INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®



Université d'Ottawa • University of Ottawa

Should population-based screening for breast cancer be extended to Canadian women aged 70 to 79?

A model-based approach

by

Leslie A. Gaudette

Thesis submitted to
the School of Graduate Studies and Research
in partial fulfilment of the requirements for the
MSc degree in Epidemiology

University of Ottawa

November 2001

© **Leslie A. Gaudette, Ottawa, Canada, 2001**

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-66041-9

Canada

ABSTRACT

Statement of Problem: This thesis applied established criteria for evaluation of a screening program to assess the population health impacts of extending population-based breast cancer screening to Canadian women aged 70 to 79. *Methods*: Experimental evidence was systematically reviewed, complemented by analyses of Canadian data for incidence, mortality, screening participation and treatment patterns for women aged 70 to 79. The MISCAN micro-simulation model was adapted to the Canadian population to assess benefits and harms. *Results*: Compared to stopping at age 69 and depending upon the level of screening participation and sojourn time, continuing to screen women aged 70 to 79 resulted in an estimated 530 to 547 false positive mammograms per 10,000 screens, 20 to 40 additional biopsies, 10 to 25 "extra" cancers, and 12.2 to13.7 prevented cancer deaths, while 30 fewer women will be diagnosed with stage T2+ cancer. About 110 life-years will be gained per 10,000 screens, with 210 to 440 life-years lived in lead-time. Quality adjustment of life-years gained reduced the benefit by up to 31% to 48% with a 5% discount factor for shorter vs longer sojourn times respectively. Between 733 and 821 screens will be needed to avert one breast cancer death, with an overall gain in life expectancy of about 4 days per screen. Overall, an estimated 459 breast cancer deaths can be prevented per year in Canada if the national target level of 70% screening participation is reached. *Conclusion:* Overall, results indicate favourable population health impacts of extending breast cancer screening to screen women aged 70 to 79. More work is needed to discern the sojourn time and the costs associated with the benefits.

ACKNOWLEDGEMENTS

I would like to thank first, my thesis advisor, Dr. Robert Spasoff, for inspiring me to contemplate a topic relevant to a policy issue, and for his always helpful, patient and constructive advice. Thanks are due to him and also to Dr. Gerry Hill, one of my committee members, for special advice and instruction regarding life tables, as well as to my third committee member, Dr. Donald T. Wigle for helpful comments at critical stages of this work.

I am very grateful to Erasmus University for agreeing to the use of the MISCAN model for this work, and want to particularly acknowledge Dr. Rob Boer for providing timely and insightful comments at all phases of this project and for patiently guiding me through the various complexities of the MISCAN model. In addition I want to acknowledge the wonderful support from the team at Erasmus, including Franka Loeve, for promptly resolving several problems that arose with the MISCAN model. Dr. Rob Boer also reviewed the results of the thesis in accordance with the agreement with Erasmus.

Dr. Francoise Bouchard deserves special mention for encouraging my interest in exploring the topic of screening women aged 70 and over and bringing me on with her team in the then Cancer Bureau at Health Canada to support the development of the project. I am also thankful for the support and comments provided by Dr. Heather Bryant and Dr. Andy Coldman of the National Committee of the Canadian Breast Cancer Screening Initiative.

Assembling the data for this project was an enormous task in which I was aided greatly by staff of the Cancer Division and other areas of Health Canada, who provided analysis files and shared data from work in progress including: Christina Bancej, Gloria Low, Asako Gomi, and Dana Paquette for the Canadian Breast Cancer Screening Database; Jay Onysko for sharing the provincial fee-for-service data and fee structures; Fan Shi and Ru-Nie Gao for their work on breast cancer treatments; Ora Kendall for assistance in obtaining the detailed population and mortality information; Robert Semenciw for aid in resolving discrepancies in mortality counts; and Dena Schanzer for sharing information from projects on ranking deaths by age group and on calculation of sojourn times for breast cancer. Their cooperation and collaboration are all very much appreciated.

My thanks are expressed to Dr. Patricia Burns, Dr. Alan Lees and Heather Jenkins of the Northern Alberta Breast Registry for inspiring my interest in breast cancer and to these many years later providing advice and agreeing to the use of the staging data from Northern Alberta in this thesis. I also thank Diane Robson and Jon Tonita of the Saskatchewan Cancer Foundation for providing breast cancer staging data from the Saskatchewan Cancer Registry.

I gratefully acknowledge Joyce Lockwood of the LCDC library at Health Canada for her professional advice assistance in developing the strategy for the literature review. Thanks are due also to Linda Mowatt-Chalu for editing and formatting the final version of the thesis. The thoughtful comments from my thesis examiners, Dr. Eva Grunfeld and Dr. Dan Krewski also aided development of the thesis.

Finally, completion of this thesis would not have been possible without the encouragement, support and understanding of my husband, Barry and my children, Darrell and Lisa. I thank Barry for his careful review of the thesis and assistance in checking data.

******

This project used data submitted by provincial/territorial organized breast screening programs to the Canadian Breast Cancer Screening Database, a collaboration of the Cancer Bureau, Centre for Chronic Disease Prevention and Control, Health Canada, and the Canadian Breast Cancer Screening Initiative Data Management Subcommittee. Fee-for-service data were provided by the provincial/territorial departments of health. The cooperation of the various departments and agencies in supplying these data is very much appreciated.

Data were provided to Health Canada from the Canadian Cancer Registry, formerly the National Cancer Incidence Reporting System, at Statistics Canada. The cooperation of the provincial and territorial cancer registries that supply the data to Statistics Canada is gratefully acknowledged.

Mortality data were provided to Health Canada by Statistics Canada. The cooperation of the registrars of vital statistics in the provinces and territories of Canada who make mortality data available to Statistics Canada under federal-provincial agreements is gratefully acknowledged.

The MISCAN breast cancer model is provided by R Boer, HJ de Koning and PJ van der Maas of the Department of Public Health of Erasmus University, Rotterdam, The Netherlands.

****

This thesis is dedicated to the many women who have died of breast cancer, including Anne Tivy, Marguerite Easingwood and Gloria Rowan, and to Josalys Scott who recently died of brain cancer but whose encouragement of my work continued until her death.

TABLE OF CONTENTS

TITLE PAGE i

ABSTRACT ii

ACKNOWLEDGEMENTS iii

TABLE OF CONTENTS v

LIST OF TABLES vii

LIST OF FIGURES ix

TABLE OF ACRONYMS AND ABBREVIATIONS x

CHAPTER 1. **Introduction** 1
- 1.1 Significance of Study 1
- 1.2 Objectives, with Criteria for Evaluation of a Screening Program 3
- 1.3 How the Approach Will Contribute to the Field 4

CHAPTER 2. **Literature Review** 6
- 2.1 Methods 6
- 2.2 Literature Overview of Benefits and Harms of Screening 8
- 2.3 Review of Quantitative Models 10
  - 2.3.1 Overview of Model Approaches 10
  - 2.3.2 Decision Analysis Models 13
  - 2.3.3 MISCAN Microsimulation Model 15
- 2.4 Review of Outcome Measures Used to Report Model Results 17
  - 2.4.1 Overview of Measures 17
  - 2.4.2 Outcome Measures Reported for Models 19
  - 2.4.3 Life Expectancy Gains 19
  - 2.4.4 Number Needed to Screen (NNS) 23
  - 2.4.5 Discounting, Utilities and Quality-Adjusted Life Expectancy 27
- 2.5 Implications 30

CHAPTER 3. **Criterion 1: Has the Effectiveness of Screening for Breast Cancer in Women Aged 70 and Over Been Demonstrated in a Randomized Control Trial?** 31
- 3.1 Background 31
- 3.2 Methods 32
- 3.3 Results 33
- 3.4 Discussion 38

3.5 Implications 43

CHAPTER 4. Burden of Cancer, Acceptability of Test, and Availability of Treatment 44
4.1 Criterion 2: What is the burden of breast cancer in women aged 70 to 79? 45
4.2 Criterion 3: Is there a valid, acceptable screening test that will reach a high proportion of women aged 70 to 79? 46
4.3 Criterion 4: Are there efficacious treatments for women aged 70 to 79? 53
4.4 Summary 55

CHAPTER 5. Criterion 5: Benefits and Harms of Screening Women Aged 70 to 79 Using the MISCAN Model 56
5.1 The MISCAN Model 56
5.2 Analytic Approach 59
5.3 Development of Input Data for MISCAN Model 61
5.3.1 Calculation of Cumulative Probability of Death Using a Life Table Approach 61
5.3.2 Calculation of the Cumulative Probability of Births before a Given Year . 62
5.3.3 Calculation of Pre-screening Incidence 64
5.3.4 Assessment of Breast Cancer Stage for Model Evaluation 66
5.3.5 Screening Participation in Canada 68
5.4 Model Validation 69
5.4.1 Population Distribution 69
5.4.2 Validation of Stage Distribution 71
5.4.3 Validation of Screening Volumes 71
5.4.4 Validation of MISCAN Output for Breast Cancer Incidence and Deaths . 74
5.5 Results: Benefits and Harms of Screening Women Aged 70 to 79 76
5.5.1 Benefits and Harms for Extra Screens, Extra Incidence, Deaths Prevented, Life-years Gained and Life-years in Lead-Time 76
5.5.2 Derived Outcome Measures 78
5.5.3 Calculation of Extra Follow-up Tests 79
5.5.4 Quality-adjusted Life-years (QALYs) 84

CHAPTER 6. Discussion 88
6.1 Criteria 1 to 4 88
6.2 Criterion 5: Review of Benefits and Harms 90
6.3 Other considerations 100

CHAPTER 7. Summary and Conclusions 104

REFERENCES 106

APPENDICES

A. Literature Review Strategy 115

B. Comparison of Models Used to Assess Breast Cancer Screening for Women aged 70+ 117

C. Data Sources, including Methods used to Develop Breast Cancer Staging Data for Analysis 124

D. Life Table Calculations for Cumulative Mortality of Death Before a Given Age, for All Causes of Death and All Causes Other than Breast Cancer Death, Canada, Females, 1995-1997 138

E. Life Table Calculations for Cumulative Probability of Birth Before a Given Year, Canada, Females, 1896-1996 140

LIST OF TABLES

2.1 Summary of outcome measures used by models reporting breast screening results for women aged 65+ or 70+ 20

2.2 Gains in life expectancy due to breast cancer screening in women aged 65+ or 70+, selected models 21

2.3 Analysis of calculation for Number Needed to Screen (NNS) 24

3.1 Design, population and characteristics of screening approach of analytic studies for women aged 70+ 35

3.2 Results of analytic studies of breast cancer screening for women aged 70+: relative risks (RR), odds ratios (OR), and 95% confidence intervals 36

4.1 Female breast cancer incidence and mortality compared to all cancers in women, by ten-year age groups in Canada, 1996 45

4.2 Participation in breast screening in Canada, 1997 and 1998 48

4.3 Results of organized breast screening in Canada, 1997 and 1998 50

4.4 Breast cancer T stage before screening implementation by age group, women aged 50-69 and 70-79, Northern Alberta and Saskatchewan 51

4.5 Number and percent distribution of T stage for breast cancers detected after screening implementation, by screening status and by age group, selected Canadian sources, women aged 50-69 and 70-79 52

4.6 Rate per 100 incident breast cancers, of hospital separations of women diagnosed with breast cancer who received breast conserving surgery (BCS) or mastectomy as in-patients by age group, Canada, 1995-1996 54

5.1 Input data for MISCAN model for cumulative mortality before a given age, Canada compared to the Dutch model used for the Netherlands 62

5.2 Input data for MISCAN model of cumulative probability of birth expressed as a ratio, by year, Canada, 1896-1996, with comparisons to the Netherlands 63

5.3 Comparison of pre-screening stage distributions from Canadian to those from other sources used in MISCAN model, age ranges from 50-69 years 66

5.4 Comparison of post-screening stage distributions for breast cancers detected at first or re-screen from Canadian screening programs, to those from sources used in MISCAN model, age ranges around 50-69 years 67

5.5 Summary of screening participation probabilities for attending biennial screening . 69

5.6 Validation of Canadian cumulative birth input data using MISCAN 70

5.7 Comparison of stage distribution in Canada of MISCAN estimates with pre- and post-screening actual data for women aged 50-79 72

5.8 Model assumptions and validation of MISCAN estimates for number of screens by screening participation level, Canada 73

5.9 Comparison of MISCAN estimates to actual breast cancer deaths in 1996 by age group and screening participation level, based on biennial screening of women aged 40-79 for 30 years (1988-2017), Canada 75

5.10 Benefits and harms of screening Canadian women aged 70-79 biennially for 30 years (1988-2017) by screening participation level, as measured by extra screens, extra incidence, deaths prevented, life-years gained and life-years in lead-time 77

5.11 Derived indicators of benefit and harm of screening women aged 70-79 biennially for 30 years (1988-2017), by screening participation level, Canada 79

5.12 Estimates of diagnostic follow-up procedures ensuing from abnormal mammograms of screening women aged 70-79 biennially for 30 years (1988-2017) by screening participation level, Canada 80

5.13 Comparison of MISCAN estimates of screen-detected cancers with cancers estimated from CBCSD detection rates for biennial screening of women aged 40-79 for 30 years (1988-2017), by screening participation level, Canada 82

5.14 Harms of screening women aged 70-79 biennially for breast cancer for 30 years (1988-2017), including extra biopsies and false positive rates, Canada 82

5.15 Summary of benefits and harms per 10,000 screens for women aged 70-79 screened biennially for 30 years (1988-2017), by screening participation level, Canada 84

5.16 Calculation of change in Quality-Adjusted Life-Years of continuing to screen women aged 70-79, assuming full (100%) biennial screening for 30 years (1988-2017), for optimistic and pessimistic sojourn times and various discount factors, Canada 86

5.17 Sensitivity analysis of Quality-Adjusted Life-Years of continuing to screen women aged 70-79 using alternative utilities for initial therapy and terminal care, other causes 87

LIST OF FIGURES

4.1 Percentage of women having a mammogram in past two years by age, Canada, 1990-99 48

5.1 Structure of MISCAN model for breast cancer 57

TABLE OF ACRONYMS AND ABBREVIATIONS

BCDDP	Breast Cancer Detection Demonstration Project
BCS	Breast Conserving Surgery
BMD	Bone Mineral Density
CBCSD	Canadian Breast Cancer Screening Database
CBE	Clinical Breast Examination
CDN	Canadian
CI	Confidence Interval
CLYS	Cost per Life-Year Saved
DCIS	Ductal Carcinoma *In Situ*
DPCP	Detectable Pre-Clinical Period
FFS	Fee-for-Service
HIP	Health Insurance Plan
m	month(s)
MBS	Model for evaluation of Breast cancer Screening
MCLYS	Marginal Cost per Life-Year Saved
MISCAN	MIcro-simulation SCreening ANalysis
MSLE	Marginal Savings in Life Expectancy
NNS	Number Needed to Screen
NPHS	National Population Health Survey
OECD	Organization for Economic Cooperation and Development
OR	Odds Ratio
POHEM	POpulation HEalth Model
PPV	Positive Predictive Value

QALY	Quality-Adjusted Life-Year
QoL	Quality of Life
RCT	Randomized Control Trial
RR	Relative Risks
TNM	Tumour, Nodes, Metastases
UK	United Kingdom
US	United States
vs	versus
w	week(s)
WHO	World Health Organization
y	year(s)

CHAPTER 1. INTRODUCTION

1.1 Significance of Study

Breast cancer is the leading cause of cancer incidence (apart from non-melanoma skin cancers) and the second leading cause of cancer deaths among Canadian women. Incidence has been rising over time, particularly since the mid-1980s, with much of the increase in recent years attributed to improved detection through screening mammography. Primary prevention of this disease in populations has proven difficult due to the lack of known risk factors amenable to change. Despite a number of treatment advances implemented from the mid-1970s onward,[1] it was not until the mid-1990s that any impact on mortality rates was observed.[2] Since the 1960s, early detection through mammography has offered the best hope to reduce mortality from breast cancer.

Mammography was initially developed as a diagnostic tool in the 1950s, and was first evaluated as a screening tool in the 1960s by the HIP (Health Insurance Plan) study carried out in New York. During the 1970s and 1980s a number of randomized controlled trials (RCTs) were conducted in Europe and North America. Meta-analyses of these RCTs found that implementation of screening mammography for women aged 50 to 69 reduced the mortality rate from breast cancer by about 30% in the screening arm.[3,4]

Debates surrounding the value of screening women 70 and over face the challenge of sparse, inconclusive data. This differs somewhat from the ongoing controversy on the value of screening women aged 40 to 49, where considerable, yet still inconclusive, evidence is available. The one clinical trial that included women aged 70 to 74 did not show a clear benefit, and genuine concerns exist about the impact of over-diagnosis, unnecessary treatment and co-morbidities, when balanced with the potential gains in life expectancy. Screening can pick up indolent cancers resulting in a temporary rise in incidence rates, yet

elderly women in particular may be less likely to die from these slow-growing cancers due to competing causes of death, and expected reductions in breast cancer mortality may not be realized. Nevertheless, screening in older women is an attractive option since life expectancy in 1990-92 for Canadian women at age 70 is 16.0 years, and at age 80 is still 9.4 years.[5] Several model-based studies have examined the population-based benefits for women aged 70 and over. Depending upon how the evidence is weighed, various professional groups have recommended a range of screening options for women over 70, including: none; only on referral from physician; biennially up to age 74, or age 79; and annually for all women over 50.[6]

Since 1988, organized population-based breast screening programs have been implemented in most provinces and territories across Canada. These target women aged 50 to 69 years, with special efforts taken to ensure their regular attendance.[7] Most programs also accept women aged 70 and over and, as of mid-1997, British Columbia specifically targets women aged 70 to 74. Demand for breast screening is increasing among elderly women, partly due to the aging cohort of women who became accustomed to regular screening in their sixties. Because of the differing policies and increasing demand, the question particularly relevant to Canada is whether sufficient benefits from organized screening can be realized to justify targeting women aged 70 and over.

Ideally, a clinical trial should be conducted to settle the question. However, because mammography screening has become relatively common among older women, RCTs may no longer be feasible.[8] And due to issues of cost, compliance and contamination, the possibility of such a trial is remote. Thus, modelling or other experimental designs based on observational data are needed.

1.2 Objectives, with Criteria for Evaluation of a Screening Program

The objective of this thesis is to use established criteria for evaluation of screening programs together with computer modeling to assess the population health impacts (i.e., benefits and harms) of extending mass breast screening programs to Canadian women aged 70 to 79. (For the purposes of this thesis, breast screening primarily refers to mammography screening, as that is the primary modality shown to be effective in randomized control trials.) The primary research question is: what are the anticipated population health impacts of continuing to screen women aged 70 to 79 compared to ending screening at age 69?

Criteria for evaluating the benefits and harms of screening programs have evolved since those first proposed for the WHO by Wilson and Jungner in the 1960s.[9] The questions below reflect a blend of those proposed by Cadman et al,[10] Muir Gray[11] and others. The approach reflects that population-based screening programs are implemented on the assumption that the benefits accruing to some members of the population will more than offset the harms experienced by others. Evidence from RCTs is first evaluated; if RCT results show a positive impact then the remaining criteria are assessed to determine the impact of implementing screening on a given population. Criteria to be assessed in this thesis for women aged 70 to 79 are:

1. Has the effectiveness of screening for breast cancer in women aged 70 and over been demonstrated in a randomized controlled trial (RCT)?

   If yes, then assess additional criteria.

   If no, do not proceed further.
2. Does the burden of suffering from breast cancer warrant screening?
3. Is there a valid screening test for breast cancer in terms of sensitivity and specificity? Is this test acceptable to women aged 70 to 79? Will the screening program reach a high proportion of women for whom it is intended?

4. Are efficacious treatments for breast cancer available for women 70 to 79? Does early detection influence outcome? Will those women with positive screen results comply with treatment?
5. What are the benefits and harms of screening among women aged 70 to 79? Overall, does screening for breast cancer do more good than harm in women aged 70 to 79?

This thesis is organized around the assessment of each criterion and thus differs somewhat from the usual presentation. The major emphasis is on Criteria 1 and 5; Criteria 2 to 4 involved relatively brief discussion. Methods and results specific to each criterion are included in the chapter assessing that criterion. Chapter 2 describes the methods for the literature review, provides a general overview of benefits and harms, and then presents an in-depth review of models used to assess benefits and harms, as well as of the outcome measures. Chapter 3 includes a systematic literature review to assess the RCT (and case-control evidence) relevant to Criterion 1, while Criteria 2 to 4 are assessed in Chapter 4. Chapter 5 presents the input data, modelling methods, and results used to assess Criteria 5. Finally, Chapters 6 and 7 discuss and summarize the results for all criteria.

1.3 How the Approach Will Contribute to the Field

A formal analysis of the health impacts of screening the population of Canadian women in this age group has not been published, although model-based results have been published for older women in the Netherlands, the United Kingdom and the United States. Results of these models are described further in Chapter 2. Benefits are generally found to ensue from increased life expectancy, as well as reduced morbidity due to, and costs for treatment of, advanced disease. Potential harms include unnecessary surgery or other

treatment for screen-detected early-stage disease that would never have become life threatening, or excessive biopsies for non-malignant breast disease. These health impacts are but one factor in the policy decision of whether a jurisdiction would choose to extend screening to women aged 70 to 79. Criteria related to resources and cost-effectiveness, that are normally also used to evaluate the feasibility of implementing a screening program, were deemed outside the scope of this thesis.

As shown later in this thesis, population-based models have been primarily applied to European and some US populations, while decision analysis models have been primarily applied in the United States. Canadian women have a breast cancer incidence rate second only to that in the United States and generally higher than among European women, and one of the longest life expectancies in the world. Further, breast screening in Canada is provided using a mix of the organized programs provided in many European countries and the more opportunistic mammography provided in the Unites States. Thus, the Canadian population differs sufficiently on several key parameters to merit a separate analysis. Model results presented in this thesis will consider the relative benefits and harms of continuing to provide Canadian women aged 70 to 79 with organized breast screening programs in terms of a number of parameters, such as follow-up of abnormal screens (including biopsies and other tests conducted), advanced cancers avoided, deaths prevented and changes in quality-adjusted life-years (QALYs).

CHAPTER 2. LITERATURE REVIEW

2.1 Methods

The following relatively simple strategy was used for the initial CancerLit search:

1. Breast neoplasms (major) (mortality, prevention and control),
2. Limit #1 to human, English and review,
3. Mass screening, and
4. 2 AND 3.

Of the 46 papers found from 1995 to 1998, the 20 most relevant were selected. As well a search was done for all papers by van Oortmarssen to locate articles using MISCAN. These were supplemented with papers already collected from my previous work on breast cancer and mammography, searching for relevant references listed in the articles found, using PubMed to find related articles for key papers, monitoring papers appearing in BMJ, JNCI, JAMA, NEJM and Lancet, and monitoring Cancer News (a clipping service) at Health Canada for articles.

Subsequently, a systematic and comprehensive search methodology was developed, using Silver Platter at LCDC to access Medline. An example of the full search strategy is shown in Appendix A. First, one set of papers was selected using the MeSH terms of mass screening (all sub-topics) AND breast neoplasms (selected topics). Then a second set of papers with a MeSH term of mammography (exploded to include all sub-topics) was identified and a third breast screening set made of those papers in either one or both of the two sets. Several search methods were used to identify a fourth set of papers on elderly women (aged 65 or over). This set was crossed with the breast screening set to identify a subset of all potentially relevant citations. Review articles were selected on the full subset. Exclusions were then made of letters, non-English citations, and citations with radio-

pharmaceuticals as a sub-topic of mammography. Further restrictions were made to include only papers where mammography OR mass screening was a major topic. This strategy was needed as European papers tended to be classified under the MeSH term of "mass screening", while US papers more commonly used "mammography". MeSH terms did not reliably identify women aged 65 and over: "aged" could be used for almost any age from 30 on up; use of "aged 80 and over" missed many articles; while "aged" as a major term eliminated most articles.

The years chosen for the systematic search were 1989 to 1999 (kept updated to the present). The earlier cutoff was chosen because there were relatively few papers published in 1989. References in later papers were also searched to pick up any relevant earlier papers. The ability of CancerLit to find additional citations was compared to Medline. We found that CancerLit contained only a few extra relevant references, all of which were conference abstracts. No additional references were found in recent years, confirming our understanding that CancerLit had not been updated with additional material beyond Medline since 1997. Therefore, we chose Medline.

A number of non-English abstracts were reviewed, but a decision was taken to exclude them as a group mainly because they: a) were very general; b) focused on how to establish a screening program in countries that did not already have programs; or c) studied populations with low to moderate rates of breast cancer, which would be less relevant than populations with rates closer to the high rates found in North America and Northern or Western Europe. References with radiopharmaceuticals as a sub-topic were omitted, thereby excluding a number of articles that focused on diagnostic mammography, with little or no relevance to screening.

Titles and abstracts of about 500 records found by the above search criteria were reviewed. Some were excluded because few, if any, women over 70 were included in the

analysis, or because the sample size was insufficient to permit a separate analysis for women 65 or 70 and over. Others were excluded because their main topics were not relevant. In general, it was difficult to narrow and focus the search without excluding at least some relevant papers. The strategy chosen was a compromise.

2.2 Literature Overview of Benefits and Harms of Screening

Benefits: Results from evaluative studies, including one RCT and one case-control study, provide only limited encouragement for screening women 70 and over. As discussed further in Chapter 3, although the point estimates for mortality reduction are less than one, the confidence intervals include 1.0. These studies demonstrate the complexity of measuring the population-health impacts due to lower participation rates, the impact of competing causes of death, and self-selection bias, which can lead to higher rates of breast cancer among those women who present for screening. Nevertheless, these studies further show that screening has the potential to benefit women by preventing deaths due to breast cancer and reducing the rate of advanced disease, thereby leading to gains in life-years, life expectancy, and quality-adjusted life expectancy.

***Harms*:** A breast screening program may cause harm through *false positive* results leading to additional screens and unnecessary biopsies, and *false negative* results leading to false reassurance. However, fewer false positives and negatives are found for women aged 70 to 79 than at younger ages due to better sensitivity and specificity. *Over-diagnosis bias* can occur where a large number of screen-detected incident cases is found, without any mortality decrease because there is no reduction in advanced cases, and many of the new cases never become clinically important.[12] Harms can also result from:

- *Over-diagnosis and over-treatment:* Ductal carcinoma *in situ* (DCIS) is more frequently diagnosed as a result of screening programs. The role of DCIS in the

development of invasive breast cancer is unclear. Harris and Leininger report that at least half of untreated women do not develop invasive cancer.[13] The 10-year mortality for DCIS is reported to be very low (about 2%-4%); despite this, women with DCIS receive similar treatment to those with early-stage invasive cancer.[14] Unnecessary treatment of this diagnosis could harm some women.[13]

- *Treatment for cancers that might never become a problem clinically (length bias):* Screening programs will differentially pick up the cancers with longer sojourn times. The sojourn time (also known as the detectable preclinical period) is the interval during which the cancer is potentially detectable by screening but cannot yet be clinically diagnosed.[15] Sojourn time differs from lead-time, which is the interval from when the cancer is actually detected by screening until when it would have been detected clinically. One effect of screening may be to diagnose and treat a cancer that would not have become clinically apparent before the woman died of some other cause of death. Harm then results from the unnecessary diagnostic and treatment procedures.
- *Longer time spent living with cancer (i.e., living in lead time):* This occurs where the breast cancer, while diagnosed earlier, either has already advanced to the point where earlier treatment is ineffective or is so indolent that it would still be treatable if found clinically at a later date. The net effect of screening is then just to extend the length of time a woman is living with breast cancer, before she dies of breast cancer or some other cause of death.
- *Psychological distress from follow-up of abnormal screens:* Steggles et al report on a systematic literature review of ten studies that assessed psychological distress related to screening. Anxiety was found to be the most prevalent consequence and was particularly evident among women requiring further investigation of abnormal

results.[16] De Haes et al examined a comprehensive range of quality of life issues of 15 states related to breast cancer screening and treatment, including 13 empirical studies of the diagnostic phase, including physical, psychological and social impacts. Utilities for each state were determined by a panel of experts, with the diagnostic state ranked as more favourable than various treatment states, but less favourable than disease-free states.[17]

2.3 Review of Quantitative Models

2.3.1 Overview of Model Approaches

This section describes a number of models identified in the literature search and assesses the feasibility of adapting them to the Canadian situation. Results from models used to evaluate screening policy options, and in some cases to make recommendations for screening policies for women 65 or older are summarized in Appendix B. Four approaches have been reported in the literature.

First, ***decision analysis*** models consider the utility of screening from the viewpoint of a decision made in consultation between a woman with selected characteristics and her physician. These models follow the consequences of the decisions through various branches representing the alternative pathways that will be taken by choice or by chance. This approach has been used to model marginal savings in life expectancy by age and health status,[18] marginal cost per life-year gained,[19] and most recently, changes in life expectancy and cost per life-year saved based on continuing to screen women aged 70 to 79.[20] These models provide interesting insights into the factors that could be considered in making a medical decision to screen, and because they form part of the evidence for screening women over 70, they will be described in detail. However, as they do not provide population-based estimates, they were not considered an appropriate method for use in this thesis.

Second, ***macro-simulation*** models use aggregate data (i.e., cell-based), rather than individual records and consider the whole population over an extended period of time. The CAN*TROL model has been used to assess breast cancer screening.[21 22 23] This model considers the full range of cancer control interventions including prevention, screening and treatment for major cancer sites. The model has some flexibility to handle age groups and population subgroups (specified in numerous ways), and to determine excess cancers, lives saved, and costs over an extended period of time. However, it does not allow for changes in the measured (detected) incidence rate caused by the introduction of a large screening program, does not consider lead-time bias, and the beta version that is currently available does not calculate QALYs. CAN*TROL has been used to assess breast cancer screening policies for women aged 50 and over; screening was found to be most cost-effective in those aged 65 to 69.[23] This model was not considered further due to its important limitations for use in screening and the difficulty in adapting it to the Canadian situation.

Third, ***micro-simulation*** considers a synthetic population of individual records over a long period of time, and employs transition probabilities between various disease and end states (i.e., death) to calculate expected numbers of cancers diagnosed, deaths prevented and other parameters. The MISCAN (MIcro-simulation SCreening ANalysis) model developed in the Netherlands considers pre-clinical disease rates and duration, screening participation patterns, breast cancer rates by stage, and mortality due to breast cancer or other causes, and several studies have applied it specifically to women aged 70 and over.[24 25 26] Jansen and Zoetelief developed a Model for evaluation of Breast cancer Screening (MBS) to evaluate lifetime risks due to radiation-induced cancers together with lifetime benefits due to early detection.[27 28] The MBS is based on a simulated Swedish population and includes estimates in each 5-year age group up to 95 to 99. Based on life-years gained from screening, the recommended screening age range is 35 to 74, but based on reduced number of fatal tumours,

this increased to ages 40 to 80 years. Finally, O'Neill et al developed a mathematical micro-simulation model that evaluates the impact of starting screening at a number of age groups in a hypothetical population of South Australian women, including those aged 70 and over.[29] They found no increase in life expectancy for women 70 and over. This model has had limited validation. A fourth micro-simulation model, the POpulation HEalth Model (POHEM) developed at Statistics Canada, has had limited development for screening interventions.[30] Of these models, MISCAN is by far the best validated for breast cancer screening and will be described in more detail in Section 2.3.3.

Fourth, ***simple life expectancy-based models:*** Law et al, and Rich and Black have separately assessed the age at which screening should stop for a number of cancers including breast.[31 32] Law calculated the number of years of life lost at each year of age as the product of the death rate and the average life expectancy at that age and then applied mortality rate reductions reported for clinical trials to determine the potential gain in life-years were screening to be implemented. Breast cancer screening for women aged 45 to 74 was reported to save on average 25 life-years per 10,000 screened; if 10 per 10,000 life-years was the minimum considered clinically significant, then the recommended age range for breast screening broadened to 35 to 79. Rich and Black calculated life expectancy as the number of days of life lost by stopping screening at various ages, again basing the benefit of screening on mortality reductions reported for clinical trials of 30% for breast cancer. Given a starting age of breast cancer screening as 50 and continuing over a life-time, they calculated a maximum potential life expectancy benefit of 42.7 days; 13.9 days of this benefit results from continuing screening after age 70, 8.5 days after age 75 and 4.6 days after age 80. They acknowledge several limitations of their simplified approach including: little consideration of the complexities of screening; the use of mortality alone as an outcome; and lack of discounting of future benefits. While these approaches may have some merit to compare

benefits of screening across various cancers or age ranges, they cannot be used to assess a broad range of benefits and harms in a population, and receive limited further consideration in this thesis.

***Summary of Models*:** The more complex models differ in approach and emphasis depending on the purpose of the analysis. Most are limited by lack of actual data documenting stage shifts due to screening in older women (stage distributions are estimated as those needed to produce mortality reductions from controlled trials). Results from the MISCAN model, the most comprehensive and best validated model, are limited by the wide range of potential sojourn times. Most, but not all, models found small benefits for women aged 70 and over. Of the five main models that provided specific recommendations for women over 65 or 70 years, four are reviewed in the next sections.

2.3.2 Decision Analysis Models

Mandelblatt et al first used decision analysis to determine whether breast screening extends life for women aged 65 and over, using three levels of health status based on presence of co-morbid conditions.[18] Marginal savings in life expectancy (MSLE) changed little as co-morbidity increased from average health to mild hypertension, but dropped by about half for severe co-morbidity as measured by congestive heart failure. MSLE also declined as age increased from 65 to 69 up to 85 years and over; the MSLE for women aged 85 and over was from one-third to one-half that of their younger counterparts. These benefits persisted when results were adjusted for long-term quality of life (QoL) factors related to stage of cancer at diagnosis. However, when MSLE was adjusted for short-term QoL factors comprising true-negative and false-positive screening results, the short-term morbidity for the oldest women appeared to outweigh any benefit. Typical values of MSLE ranged from 2.2 days for women aged 65 to 69 in good health, to 1.09 days for those in poor health, to 0.49 days for women aged 85 and over, and in poor health. Adjustment for long-term QoL tended

to improve these results slightly, possibly because of the more favourable stage distribution in screened women. However, inclusion of both short- and long-term QoL factors reduced the MSLE to 1.44 days among women in average health aged 65 to 69; MSLE was essentially 0 (zero) for women in all co-morbid groups aged 85 and over, as well as for those with the highest co-morbidity level in women aged 80 to 84. The authors note that adjustments for QoL do not reflect other, less tangible factors, and that this type of model is more useful for individual decision-making than for public health policy makers who may require more complex simulation models. Despite the small effects and, indeed, the lack of any effect in the oldest women when adjusted for QoL factors, the authors conclude that screening benefits women of all ages and that more information is needed on women's preferences for screening.

Lindfors and Rosenquist initially developed a decision analysis model to analyze the cost-effectiveness of needle-core biopsy in a breast screening program aimed at women aged 40 to 85 years.[19] This model subsequently compared cost-effectiveness of various screening frequencies (e.g., annual or biennial) for different age groups.[33 34] Women who were screened were compared to those who were observed, but not screened, to determine marginal cost per life-year saved (MCLYS). Biopsy methods were assessed in detail, but stage of disease was not considered. The initial study concluded that use of needle-core biopsy reduced the cost of breast screening.[19] The second study reported that the most cost-effective screening strategy was biennial mammography for women aged 50 to 79. A major drawback is that results are presented only for broad age ranges such as 40 to 79 or 50 to 79, with no assessment of the marginal costs that apply to ages 40 to 49 or 70 and over. However, this does not deter the authors from making recommendations in favour of screening for women aged 70 and over.

Kerlikowske et al used decision analysis to examine the impact on life expectancy and cost effectiveness of continuing screening mammography for women aged 70 to 79.[20] This study identifies a group of women at low risk for breast cancer based on low bone mineral density (BMD), and compares three scenarios: stopping screening after age 69, continuing to screen only women in the top three quartiles of BMD, or continuing to screen all elderly women up to age 79. The sensitivity analysis includes discount rates ranging from 0% to 15% (much higher than the usual rate of 3%) to explore differences in time preference among elderly women that may affect their decision about screening. The impact of increased diagnosis of DCIS was assessed by including estimates of increased incidence but lower mortality and treatment costs, depending upon the scenario chosen. Various utilities of life after treatment and of life with metastatic breast cancer were also assessed using sensitivity analyses. The benefit of continuing mammography to age 79 was small: 10.8 deaths would be prevented per 10,000 women, who would experience on average, an extra 2.4 days of life expectancy. Nearly all of the benefit (9.4 prevented deaths, and 2.1 days) would accrue to women in the top three quartiles of the BMD distribution. Notably, the small gain of about 2 days of life expectancy was very similar to that observed by Mandelblatt, but Kerlikowske is more cautious in her recommendation, emphasizing the importance of a woman's time preference (i.e., discount rate) in balancing the small gains against the potential harms, when making a decision on screening.

2.3.3 MISCAN Microsimulation Model

Boer et al used MISCAN to model the Netherlands female population to determine the best upper age limit for breast cancer screening.[24] Screening older women should prevent more breast cancer deaths due to their higher breast cancer mortality rate, although fewer life-years will be gained because life expectancy decreases with age. Early detection is further complicated in older women due to the longer sojourn (or preclinical duration) times which

result in more women dying of other causes before their breast cancer would have been diagnosed in the absence of screening. Overall, Boer et al calculated that the ratio of extra incident cases detected by screening relative to prevented breast cancer deaths will increase with increasing age, due to potentially longer sojourn times (which increase the numerator) and to competing causes of death (which reduce the denominator). However early detection has the benefit of reducing the impairment from, and costs of, advanced treatment among some women. Their conclusions are mixed: when assuming a short sojourn time and not including costs, there is no upper age limit where the balance becomes clearly negative. When costs are included, the cost-benefit ratio for screening women aged 69 to 74 appears to be at least as favourable as improving participation in women aged 50 to 69. Their results are limited by lack of accurate data, particularly on sojourn times in older women.

MISCAN was further used to evaluate the consequence of varying the screening interval for women aged 65 and over.[25 26] The basic model compared effects of biennial screening intervals in women aged 50 to 64 only, assuming 100% attendance, to the additional effects of continuing to screen these women at ages 65 to 82 at intervals of 1, 2, and 3 years(y).[26] More frequent screening increased the rate per 1,000 women of prevented breast cancer deaths from 2.9 for no screening up to 5.0 for 3y, 5.5 for 2y, and 6.2 for 1y intervals, and resulted in an increase in life-years gained per 1,000 women from 54 for no screening to 71, 76 and 83 for 3y, 2y and 1y intervals respectively. By contrast, extra incidence due to screening increased from 0.6 per 1,000 women for no screening to 5.4 for 3y, 5.8 for 2y and 6.5 for 1y intervals, and extra life-years with disease per 1,000 women rose from 41 to 114 for 3y, 125 for 2y and 140 for 1y intervals. The balance of favourable to unfavourable effects favours more frequent screening, as the negative effects increase at a slower rate than the positive effects.

2.4 Review of Outcome Measures Used to Report Model Results

2.4.1 Overview of Measures

Health policy makers need standardized outcome measures to assess the effectiveness of public health interventions and to compare the impact of alternative policies among various interventions or target population groups.[35] [36] Outcome measures usually take the form of either costs (often expressed as cost per life-year gained), or health impacts such as life-years gained. This section assesses outcome measures for the health effects to be modelled later in this thesis.

Implementation of evidence-based health policies from a population health perspective demands consideration of the effectiveness of an intervention on the health of a population, as well as the cost considerations. Health outcome measures are needed to assess both the population and the individual perspective. Both are important because while only a few individuals may benefit greatly from an intervention such as cancer screening, a population-based measure will spread this benefit over the entire population. On average, the perceived benefit to an individual prior to receiving the intervention will be quite small. At one extreme, Kopans argues forcefully that the medical and scientific analyses must be separated from the economic considerations, and that individuals should be allowed to make decisions based on medical information without consideration of the economic cost to society.[37]

What outcome measures are available to assess health policies for screening? Harris and Leininger review a number of useful measures, including the risk of developing breast cancer in the next 10 years and the risk of dying within the next 15 to 20 years, pointing out that this risk of dying per 1,000 women ranges from 7.8 at age 40 up to 25.3 at age 70.[13] They then examine the relative risk reductions from clinical trials (30%) and consider the absolute risk reduction, which ranges from 1.8 per 1,000 women at age 40 up to 7.6 at age

70. Finally, they calculate the harms in terms of numbers of procedures per cancer detected and estimate the number of lives extended through screening 10,000 women. While use of relative and absolute risk in decision-making is not straightforward, the absolute risk reduction is generally more helpful to women making a screening decision. Notably, clinicians wishing to promote screening will usually rely more heavily on the relative risk, while those opposed tend to emphasize the absolute risk.

Wright and Weinstein observe that effectiveness of preventive services is usually measured in terms of *number of cases prevented or the number of lives saved*, although this measure gives no information on premature mortality.[35] Further, a recently proposed measure–the *number of people who must be treated to prevent one death* or to produce one successful outcome–does not tell us how long the survivors will live.[38] Effectiveness of medical treatments are also often measured as the *gain in median survival time* or *gain in survival probability at five years*, or both. Instead of using these uni-dimensional measures, Wright and Weinstein propose using *changes in life expectancy*, which reflects the difference between the survival curves for the treatment and control groups. This indicator directly measures 'the shift of the survival curve caused by the intervention' and represents a two-dimensional approach. Notably, a change in life expectancy represents a probabilistic gain over the whole survival curve, and not just a gain at the end of life. In fact, most of the gain occurs very early on.

Of these measures, survival gains are generally well understood and reported in the context of clinical trials of treatment (and other) interventions that determine efficacy, but are less widely used to report effectiveness of population-based interventions such as those for screening. Life-years gained, which can often be readily converted to life expectancy if appropriate population denominators are available, is a well-established outcome measure; a detailed analysis is presented in Section 2.4.3. By contrast, the number needed to treat has

only recently been adapted to a screening context, as the number needed to screen (NNS) to avoid one death or one adverse outcome.[36] This measure is based on the absolute reduction of risk, an important criterion for policy makers. However, analysis of several papers in which it has been used (see Section 2.4.4) shows that the underlying methodology needs to be clearly defined to produce comparable results.

2.4.2 Outcome Measures Reported for Models

What measures are employed by models for breast cancer screening for older women? Table 2.1 lists various outcome measures that assess and compare benefits and harms to health as well as impacts on health resources. Of the 12 articles included in this analysis, 8 provided at least one measure of life-years gained. However, it can be difficult to compare results directly due to the variety of methods used to present data. Only a few models were used to calculate information on harms. With regard to measures of impacts on health resources, most models assessed overall effectiveness by computing one or more measures of costs, ranging from total costs to costs per life-year gained, with associated differences in methods and discounting measures used. One paper explicitly included measures of the number needed to screen to avert one death and also of the number of additional examinations needed to avert one death.[20]

2.4.3 Life Expectancy Gains

Wright and Weinstein reviewed published life expectancy gains for 83 medical interventions, and concluded that a gain in life expectancy of a month from a preventive intervention targeted at populations at average risk can be considered large.[35] Their analysis included two of the models examined in this thesis for women aged 65 and over, or 70 and over. Results related to life-years gained were extracted from published papers, converted to a common measure of increase (or decrease) in life expectancy, and summarized in Table 2.2.

Table 2.1 Summary of outcome measures used by models reporting breast screening results for women aged 65+ or 70+

Outcome Measure	Models (# of reports)	Total Reports
Benefits to Health Status		
Life-years gained (or equivalent, i.e, could be total, or expressed per 1,000 women or as life expectancy per one woman)	MISCAN (3), MBS, CAN*TROL, Mandelblatt, Lindfors, Kerlikowske	8
Quality-Adjusted Life-Years (QALY) gained	MISCAN (1), Mandelblatt (1), Kerlikowske	3
(Breast cancer) deaths averted/prevented (changes in number of deaths)	MISCAN (2), CAN*TROL (3), Kerlikowske	6
% increase in 10y survival	Australian	1
% increase in life expectancy	Australian	1
Harms to Health Status		
Extra (excess) incidence	MISCAN (2) , MBS (1)	3
Life-years in lead-time	MISCAN (2)	2
Radiation risk	MBS	1
Impacts on Health Resources		
Costs (any measure, usually cost per life-year saved, but sometimes just total costs)	MISCAN (3); CAN*TROL (3); Mandelblatt; Lindfors; Kerlikowske	9
Number needed to screen to avert 1 death	Kerlikowske	1
Number of additional examinations to avert 1 death	Kerlikowske	1

Six papers contained sufficient information to calculate life expectancy. Some were excluded because they focussed almost exclusively on costs,[23] whereas for others, population denominators were either absent [24] or possibly inappropriate,[29] making it impossible to calculate life expectancy gains per woman screened.

Comparison of results is also hampered by the variety of age ranges and study designs. Nonetheless, the models that clearly reported the increase in life expectancy ***per woman*** screened found increases of about 2 weeks to 1 month. MISCAN calculates

Table 2.2 Gains in life expectancy due to breast cancer screening in women aged 65+ or 70 +, selected models

Model	Design	Increased life expectancy in days	per woman	per screen	Comments
MISCAN (1999) [26]	compare extending screening to women aged 65-82 to stop at age 64	**Age: 65-82**			
		Screen Interval:			
		1y	30.3	2.9	Derived from data presented in paper.
		2y	28	4.4	
		3y	25.9	5.5	
		stop at age 64	20	7.7	
Decision Analysis Co-Morbidity Mandelblatt (1992) [18]	model screening for 3 co-morbid states, and 2 QoL adjustments	**Age:**			**QoL adjust** / **Ratio to average unadjusted days**
		65-69		2.2	1.4 / 0.66
		70-74		1.8	1.1 / 0.61
		75-79		1.3	0.7 / 0.52
		80-84		1.0	0.3 / 0.35
		85+		0.7	0.1 / 0.07
Lindfors & Rosenquist (1995) [33]	model screening over various ages & screen intervals	**Age: 50-79**			Results agree with Wright and Weinstein calculations, note that the misprint of 50-59 in Lindfors article was replicated by Wright & Weinstein.
		biennial	25.6		
		annual	31.5		
CAN*TROL (1989) [39]	model screening for 25% of asymptomatic US women from 1989-2000	**Age: 65-75**			Based on HIP (Health Insurance Plan) study mortality reductions of 30% @ 10y and 25% @ 20y. CBE = clinical breast examination. These results appear to be per woman.
		0% discount			
		CBE	10		
		CBE+mammo	15		
Kerlikowske et al (1999) [20]	model continuing biennial screening	**Screen Strategy:**			The main result has been extracted from a complex design; note errors in the article; results are reported as 'per woman' but appear to be 'per screen'.
		continue 70-79	2.47?		
Law et al (1999) [31]	simple life expectancy adjustments	**Age:**			Benefit may be calculated for each five-year age range only; thus might be closer to a benefit per screen.
		65-69		1.6?	
		75-79		0.5?	
Rich & Black (2000) [32]	simple life expectancy adjustments	**Age:**			This benefit is calculated to end of life from the age given.
		65	20.5		
		70	13.9		
		75	8.5		
		80	4.6		
		85	1.8		

increases of from 26 to 30 days (depending upon the screening interval) in life expectancy when screening is extended to women aged 65 to 82, after regular biennial screening at younger ages.[26] Lindfors and Rosenquist report increases for screening from age 50 to 79 on an annual or biennial basis of 31.5 and 25.6 days respectively over the whole 30 year screening period.[33] The CAN*TROL model reported increases for a combination of CBE and mammography for women aged 65 to 75 ranging from 15 to 36 days depending upon whether the mortality reduction was modelled according to results of the HIP or BCDDP studies (BCDDP data not shown in table).[39] Rich and Black find a benefit per woman age 65 of 21 days increased life expectancy and at age 70, a benefit of 14 days.[32]

Two models specifically calculated the benefit of screening in terms of ***per screen*** (or screening encounter) among women aged 65 and over or 70 and over. MISCAN found increases in life expectancy of 7.7 days to screen women aged 50 to 69, plus an additional increase of 2.9 to 5.5 days for screening intervals of 1y to 3y respectively for extending screening to women aged 65 to 82.[26] Mandelblatt reported marginal savings in life expectancy of up to 2.17 days per screening encounter for women aged 65 to 69, declining with age and greater co-morbidity to just 0.49 day for women aged 85 and over with the highest co-morbidity.[18] Data reported by Law appear to be closer to a result per screen, and show an increase of 1.6 days in life expectancy for screening women aged 65 to 69 and of 0.5 days for screening ages 75 to 79.

Results from other models did not compare well, and were also harder to interpret clearly. Kerlikowske et al reported life expectancy gains of 2.47 days for continuing to screen women from age 70 to 79, compared to stopping at age 69.[20] These results are interpreted as being the overall benefit to each woman (even though they are more comparable to results of other authors per screen). Further confusion arises when the authors compare them directly to the Mandelblatt data, which are specifically for one screening

encounter. Data from Jansen and Zoetlief data on life-years gained per 1,000,000 women convert to a value of 1.13 days per woman aged 70 to 74 (data not shown in table).[28] This may reflect an inappropriate population denominator or incorrect calculation.

2.4.4 Number Needed to Screen (NNS)

In 1998, Rembold claimed credit for being the first to formally develop a methodology to calculate the number needed to screen, "defined as the number of people that need to be screened for a given duration to prevent one death or adverse event." His results are derived from meta-analyses and associated confidence intervals for RCTs of various cancer and cardiovascular prevention and screening interventions.[36] He calculates the NNS with mammography to prevent one breast cancer death as ranging from 4,576 for women aged 40 to 49 down to 695 for women aged 60 to 69, over an 8- to 9-year period based on clinical trial results. He also reports a range for NNS of 8,054 down to 1,251 over a 5-year period (for the same age groups, respectively) in order to provide a common time period across a number of interventions. (Notably, despite Rembold's article, a search of PubMed for 'number' and 'needed' and 'screen' and 'breast' reveals that NNS appears to have been used only occasionally for assessment of screening policies. This search identified just 11 papers of which only Rembold clearly used the term.) Table 2.3 summarizes results for number needed to screen for five articles that calculated results for breast cancer screening in women.

Two papers published in 1995 used approaches similar to number needed to screen to present data, although neither actually uses the term. Wright and Mueller report that the number of women screened to achieve one less death per year ranges from 7,086 up to 63,264 to infinity, depending upon the % reduction in mortality (for all age groups) achieved in the various RCTs.[40] They conclude that screening is not beneficial to women, based on the vanishingly small number they claim would benefit. Of the several letters to the editor of

Table 2.3 Analysis of calculations for Number Needed to Screen (NNS)

	Adverse Event Prevented	Time to Accrue Benefit	Screens vs Women	Age Range	Value of NNS	Comments
Wright & Mueller (1995)	1 death	1 y	Screens (?)	40+*	7,000-63,000	Range of mortality reductions for RCTs
Harris & Leininger (1995)	(extend) 1 life	7 y?	Screens, annual	50-70*	1,700-5,000	Range reflects CIs for clinical trials
	(extend) 1 life	7 y?	Screens, biennial	50-70*	850-2,500	
Rembold (1998)	1 death	8-9y	Screens (?)	40-49	4,576	Mid-point of clinical trial results; appears to refer to number of screens, not number of women screened
				60-69	695	
	1 death	5y	Screens (?)	40-49	8,054	
				60-69	1,251	
Kerlikowske et al (1999)	(extend) life by 1 y	10y	Women (?)	70-79	150-500	Range reflects highest and lowest quartile of bone mineral density
	1 death	Life (?)	Women	70-79	1,064-7,143	
Boer et al (2000)	1 death	Life	Screens	65-79	800	Results derived from tables published in the article, and adverse events are accrued over entire life-span
	1 death	Life	Women	65-79	200	
	1 life-year	Life	Screens	65-79	58	
	1 life-year	Life	Woman	65-79	13	

* Annual screening frequency, all others are for biennial frequency

The Lancet in response to this article, de Koning was the only one to challenge these numbers, stating that fewer than 1,000 screens should be needed to prevent one death.[41] Harris and Leininger also calculate the number of women screened annually to extend one life as 1 in 1,700 to 5,000 for women aged 50 to 70; with biennial screening, these drop to 850 and 2,500 screens respectively.[13] These results are lower than those reported by Wright and Mueller, even though the mortality reductions are based on the same RCTs. The lower results occur primarily because they are "extending one life" while Wright and Mueller are

"saving one life per year", and also in part due to use of mortality reductions for just the 50-69 age group. Notably Harris and Leininger more clearly identify the outcome as a consequence of a single screening mammogram (i.e., the number of screens, and not the number of women).

Results based on NNS variations have since been published for at least one breast screening model, and can be derived from another. Kerlikowske et al calculated that the number needed to screen (biennially) with mammography for 10 years to extend a woman's life by one year ranged between 150 and 500 for women aged 70 to 79 with decreasing BMD level.[20] Expressed in other words, this seems to be the number of women needed to screen to save one life-year. In another table, the number needed to screen to avert one death from invasive (breast) cancer is reported as 1,064 for the top three quartiles of BMD and 7,143 for the bottom quartile. Results presented by Boer et al can be easily recalculated for a similar population of women aged 65 to 79 who continue to receive biennial screening, to obtain the number of screens needed to prevent one death as 800 and the number of women needed to screen as 200.[26] His figure of 800 is close to that of 1,064 reported by Kerlikowske, but the 200 figure is much lower. This further suggests that Kerlikowske's estimate might be based on number of screens and not number of women (see Section 2.4.3). Boer's data can also be translated into just 58 screens needed to save one life-year, and the number of women needed to screen to save one life-year as 13; both these estimates are much lower than Kerlikowske's results of 150 to 500 for what purports to be a similar analysis. Kerlikowske estimates that 10.8 deaths will be averted per 10,000 women (probably actually screens) while MISCAN estimates that 5.8 deaths would be prevented per 1,000 women (or 58 per 10,000–nearly five times higher).

These divergent results reflect the ease with which one can mislead (or even lie!) with statistics. Wherein lies the discrepancy? Four factors contribute. First, the *age range* chosen

to present results must be clearly defined, as the absolute mortality reductions are up to 8 times more favourable in older women, even though their relative mortality reduction may be lower. Second, NNS must consider whether it is assessing the *number of screens*, or the *number of women* following a screening regimen. Although for many public health interventions these numbers will be the same, breast cancer screening involves a number of screens over time, i.e, one woman screened biennially from age 50 to 79 will receive 15 mammograms. If results are presented per screen, the NNS will be 15 times higher than if presented per woman, a distinction not always made clear in published results. Third, the *time period* used to accrue the benefits must be specified–the very large figures reported by Wright and Meuller were based on a reduction of ***one death per year***, the intermediate figures by Rembold used a time frame of 5 or 8 years, while the lower figures for Kerlikowske and Boer used the remaining lifetime of the woman screened. Obviously, the shorter the time period in which a benefit of one death or life-year averted can be realized, the more screens would have to occur. If one assumes that deaths may be prevented over a 10-year period, this factor would produce a 10-fold difference. Fourth, the *choice of the adverse event* is important–while most analyses use the number of deaths avoided, life-years gained have also been reported. Avoiding one death will generally save more than one life-year. Norum estimates about 15 life-years for ages 50 to 69 and concludes that at least 4 to 9 life-years must be saved for mammography to be cost effective.[42] Data presented by Boer et al suggest about 8 life-years are gained per death avoided for women aged 65 to 79.[26] Thus, use of life-years will reduce the number needed to screen by a factor of as much as 15 (and perhaps less in older women). As shown by the published results above, the NNS ranges between 63,000 and 13 (a 5000-fold difference) depending on how each of these factors is defined for the analysis. This variation is well within the maximum possible limit as can be

independently calculated by multiplying together the differences estimated for each factor: 8*15*10*15 =18,000.

Despite the methodological confusion, NNS is potentially useful as an outcome measure for policy development, providing that a standardized methodology can be agreed upon and adopted. Assuming that screening is beneficial, presenting a woman at age 70 with the scenario that her participation in regular screening along with 12 other friends will save, on average, one life-year among them is a much more encouraging message than to state that tens of thousands of women must be screened to avert one death over a one-year time period.

2.4.5 Discounting, Utilities and Quality-Adjusted Life Expectancy

Quality-Adjusted Life Expectancy: Some models have been used to produce quality-adjusted life expectancy measures and thereby applied utilities for living in various diagnostic and treatment states. This approach has not been widely used to assess screening for women aged 70 and over. De Haes et al used MISCAN to estimate that QALYs were 8% lower than unadjusted life-years saved in a population of Dutch women aged 50 to 70 who were screened biennially. This small decline in effectiveness was judged too small to affect decision-making for screening in this age range.[17] Boer et al use QALYs (5% discounted) that generally agree with undiscounted life-years gained for women aged 70 and over using a model with optimistic (shorter) sojourn times. For the pessimistic model (with longer sojourn times), the QALYs declined slightly (about 5%) among women over 80, while unadjusted life-years showed a relatively flat trend.[24] Larger reductions in QALYs were reported by Mandelblatt in a population of US women aged 65 to 85 and over with the decision analysis based on one screening episode.[18] They found that life expectancy gains, when adjusted for short and long-term quality utilities, were about one-third to two-thirds lower among women of average health between the ages of 65 and 85, and almost vanished for women aged 85 and over. Kerlikowske also reported lower QALYs. Compared to an

unadjusted figure of 2.1 days, QALYs ranged from 0.1 to 2.0 days 'as the utility placed on life after treatment with breast cancer and life with metastatic breast cancer decreased to 0.8 and 0.3 respectively'. No further explanation is provided regarding the upper and lower values of the QALY range. These data are for the women in the top 3 BMD quartiles–the marginal 'increase' in QALYs becomes vanishingly small for the lowest quartile (range is -0.2 to +0.2).[20]

Utilities: Three models presented utilities used to calculate QALYs. Utilities used in the MISCAN model[24] were derived from a questionnaire evaluated by public health employees and breast cancer experts.[17] Values for some of the 15 states identified are:

- screening attendance for 1 w: 0.994;
- diagnostic phase of 5 w: 0.895;
- post-surgical treatment, disease-free for >1 y: 0.947, 0.960;
- initial treatments with duration of 2 m to 2 y: 0.717 to 0.914;
- treatment of metastatic disease for up to 20 m: 0.531 to 0.663;
- terminal illness (1 m): 0.288.

These values were subsequently reduced to three main values of 0.750 for one year of initial treatment, 0.955 for the disease-free interval and 0.624 for 1.768 years of care for metastatic disease.[43] Mandelblatt used a declining exponential approximation of life-expectancy methodology to determine utilities. In this method, values for short-term impacts are subtracted from life expectancy; these were 0.1 for 30 days for false positive mammograms, and 0.1 for 5 days for true negatives. For long-term QoL adjustments, life expectancy in each state was multiplied by utilities (range in parentheses) for stage of disease as: local, 0.9 (0.8 - 1.0); regional, 0.8 (0.7 - 0.9); and distant, 0.5 (0.4 - 0.6).[18] Finally, Kerlikowske adapted utilities from an Australian study by Hall, which had been developed empirically using a time trade-off technique as assessed by women with and without breast cancer. Hall

et al, however, had reported these utilities as an average for good and poor health states, respectively, for women aged 45 to 69. As these utilities declined with age, for women aged 65 to 69, the corresponding data for good and poor health were 0.64 and 0.18, respectively.[44] The adapted utilities used by Kerlikowske were: for life after breast cancer treatment, 0.8; and for metastatic disease, 0.3.[20]

Notably the three methods produced relatively consistent utilities of about 0.8 to 0.9 for initial treatment/ early stage disease/ good health, but varied more widely from 0.18 for poor health with breast cancer to 0.288 for terminal disease up to 0.5 for distant disease. This is due in part because definitions of states were more similar for the healthier states. Utilities for screening attendance used by Mandelblatt, however, were lower than those developed for MISCAN. The more detailed utilities developed for the MISCAN model will be used for the QALYs calculated in this thesis.

Discounting: Use of discounting varied amongst the models evaluated. Most did not use discounting for the main analysis for life expectancy gains. Several studies used discount rates of 5 or 6%, particularly where the main focus was cost-effectiveness.[24 26 39] Of the studies that reported QALYs, Boer et al used a discount rate of 5%,[24] while Mandelblatt did not discount.[18] Kerlikowske did not discount baseline results, but used a range of discount rates (0, 3, 5, 10 and 15%) to assess costs per life-year saved as well as the number needed to screen with mammography for 10 years to extend a woman's life by one year.[20]

An expert panel in the United States has recommended that societal cost-effectiveness analyses use a standard 3% discount rate.[45] This rate is based on theoretical and empirical evidence of time preference, which represents the relative value that people place on events happening now or in the future. The panel further recommended that sensitivity analyses include values of 5% (to compare to the previously accepted standard) as well as between 0 and 7%. Petitti concludes that non-monetary health benefits should be discounted at the

same rate as costs, noting that the remaining issue is the time preference depending upon whether a societal or individual perspective is taken.[46] Lipscomb et al summarize results from several studies that look at choices individuals make. They conclude that individual discount rates frequently lie outside the 0%-10% range, although the overall mean has been reported as 3.3%.[47] Kerlikowske points out that elderly women may have discount rates of as much as 10 to 20%, with women who place a higher value on the present having a higher discount rate.[20] These discount rates are normally applied to cost effectiveness and QALY analyses, and only occasionally to the basic changes in life expectancy.

2.5 Implications

Based on the review of models, MISCAN was chosen to provide a model-based approach to quantify the benefits and harms of extending screening to women aged 70 to 79 (*see* Chapter 5). Benefits to be modelled include gains in life-years, life expectancy, and quality-adjusted life expectancy, while harms may ensue from the volumes of extra diagnostic and other follow-up tests, unnecessary treatment and diagnostic procedures due to length bias, and the amount of life-years spent living in lead-time. Factors such as psychological distress and inconvenience of attending additional appointments are considered secondary to demonstrating the main population health impacts, and will not be considered further.

CHAPTER 3.

CRITERION 1: HAS THE EFFECTIVENESS OF SCREENING FOR BREAST CANCER IN WOMEN AGED 70 AND OVER BEEN DEMONSTRATED IN A RANDOMIZED CONTROL TRIAL?

3.1 Background

Kerlikowske et al and Demissie et al have published meta-analyses of results of randomized control trials (RCTs) and case-control studies.[3,4] These meta-analyses reported mortality reductions for breast cancer of 26% and 30% respectively, for women aged 50 to 74 invited for screening, where participation rates at the first screening round ranged from 60% to 80%. Mortality reductions were not found for women aged 40 to 49. Gotzsche and Olsen recently published a third meta-analysis, that concluded that screening had no effect on breast cancer mortality in women of all ages.[48] This study assessed three elements of methodologic quality – randomization methods, baseline comparability, and exclusions after randomization – and concluded that only two of the RCTs were of sufficient quality to include in a meta-analysis. Correspondents to the editor of *The Lancet* strongly criticized the inclusion criteria used in this meta-analysis, as well as the incomplete understanding by Gotzche and Olsen of basic concepts in analysis of screening studies. This meta-analysis will receive limited further consideration in this thesis.

None of these meta-analyses considered women aged 70 and over separately from the 50-69 age group. As well, the applicability of the results of these meta-analyses to women aged 70 and over has been questioned as only one of the eight RCTs (Swedish two-county) included women over 70 and only one other (Malmo) included women over 65. Demissie and Kerlikowske included six and four case-control studies respectively (none of the latter was included by Gotszche), and just two of these studies included women aged 65 to 70 (Florence)[49] and 65 to 69 (Malmo).[50] For the Nijmegen case-control study, both meta-analyses presented data only for women aged 35 to 65 years, and excluded women aged 70

and over. Although van Dijck et al briefly reviewed results of a number of studies of breast cancer screening in the elderly,[51] no systematic presentation of results of studies using meta-analysis techniques has been reported in the scientific literature. Thus a systematic search was conducted with the aim of synthesizing information for women aged 70 and over. Case-control studies were also included in this search as they are frequently analyzed together with RCT evidence.

3.2 Methods

Abstracts selected from the literature search described in Chapter 2 were further searched to identify articles mentioning case-control, case-referent or RCTs. An independent search was also conducted on Medline to look for any articles published prior to 1989. The two meta-analyses described above[3 4] were also reviewed to identify any RCTs or case-control studies that may have included women aged 70 and over. References listed in retrieved articles were further reviewed to locate relevant articles.

This search confirmed that the Swedish two-county study was the only RCT to publish results for (or include) women aged 70 and over,[52 53 54 55] and that no additional results had been published for this age group. Two case-control studies that had included women up to age 69 or 70 were also reviewed to determine if results were reported separately for women aged 65 and over. However no sub-group analyses of this age range were presented,[50 56] in part because the sample sizes were too small to permit meaningful analyses.[49] Another study, the Breast Cancer Detection Demonstration Project, provided five annual screening examinations to US women aged 35 to 74, who were recruited on a volunteer basis (i.e., non-random, non-population-based), but did not include a control group of women with no screening. The ratios of observed to expected mortality rates nine years after entry were published by age at entry for women aged 35 to 49, 50 to 59 and 60 to 74;

these ratios were 0.89, 0.76 and 0.74 respectively, with no confidence intervals provided. This study was excluded from further consideration because it was neither an RCT nor a case-control study, and also it did not provide breakdowns for women 70 and over.[57]

Two case-control studies were found that focussed on elderly women. These were not included in the original meta-analyses, although age was not stated as an exclusion criterion. First, a number of articles and letters reporting results of the Nijmegen case-control study were selected for further analyses.[51 58 59 60 61] The second case-control study was conducted by Brown and Hulka[62] who tested the hypothesis that elderly women with metastatic breast cancer were previously screened less than controls. Elderly was defined as women aged 60 and over and the final sample included just 109 cases and 211 controls, with a median age of 70. Results were not published separately for women over 70. Due to the unexpected finding that less than 6% of controls had ever had screening mammography, this study had insufficient power to draw conclusions. This study was not considered further in this analysis due to the lack of data published separately for women age 70 and over, and because the outcome variable – advanced breast disease – may not be appropriate.

3.3 Results

The study designs and screening approaches for the one RCT (Swedish two-county) and the one case-control study (Nijmegen) that reported results for women 70 and over are summarized in Table 3.1. The case-control study used a nested design as it selected breast cancer deaths in the population invited for screening and then compared the screening histories with population-based controls. Both studies were initiated in 1977 and 1978, with one-view screening mammography and no clinical breast examination. Screening attendance was lower at older than at younger ages. In the RCT, the 77% participation rate for women aged 70 to 74, while lower than the 86% found for those aged 65 to 69, was still higher than

that reported for younger women in many RCTs. Participation rates in the case-control study were much poorer: at the 7th round (the index round in one reported study) just 54% of invited women aged 65 to 69 and 43% of those aged 70 to 74 attended. However, participation rates were considerably higher for women who had already attended the sixth round.

The RCT is limited for the 70-74 age group by the relatively small sample size of 17,646, by the fact that only two screening rounds were completed, and by the relatively long screening interval of an average 33 months for women aged 50 to 74 (not reported separately for women aged 70 to 74). The screening interval found to be most cost-effective in women aged 50 and over is now considered to be 24 months;[25 26 33] longer intervals reduce screening effectiveness. Another 28,114 women aged 75 and over were initially randomized in the Swedish two-county trial, but data have never been published because the compliance rate of less than 50% was considered too poor.[52]

Both the RCT and case-control studies may be limited by the higher potential for mis-classification of deaths in older women. However, blinded assessment of the cause of death was conducted in both, with the reviewers benefitting from considerable additional information to that available on the death certificate. This step is important. Andersson[63] noted that, after the review, as many as 10% of breast cancer deaths were discordant with the vital statistics cause of death. Over all the Swedish trials, cause-specific mortality (except for breast cancer) was similar between the study and control groups and the Swedish population.[64]

Results for these two analytic studies of breast cancer screening for women aged 70 and over are presented in Table 3.2. Some results are presented for age 65 to 74, or age 65 and over, based on the age groups analyzed by the authors. The Swedish two-county trial found a relative risk (RR) for breast cancer mortality in women invited for screening of 0.98

Table 3.1 Design, population and characteristics of screening approach of analytic studies for women aged 70 +

Study (references)	Start Year	Age at Entry	Random-ization	Baseline Characteristics	Attendance	Screen Interval	No. of Views	Breast Exam	Assess Cause of Death
Randomized Control Trial: Swedish two-county [52 53 54 55 64]									
Kopparberg	Oct 1977	70-74 at entry	cluster	13% had external mammograms (all ages) [52]	77% for aged 70-74 compared to 86% for 65-69 [55]	33 months (average women aged 50-74)	1	No	Blinded assessment by End-Point Review Committee of death certificates, necroscopy reports, pathology and medical records
Östergotland	May 1978						1	No	
Case Control Study: Nijmegen			**Matching Variables**		**Attendance****				
Round 2 (first round for 65+ yrs) [59 60 61]	1977-1978	65+ (at first screen)	Same year of birth and same invitation history before case Date of Dx		*age 70+ at invitation* Round # % 2 6,180 34.5 3 6,379 25.7 4 6,571 21.0	Invitation and reminder every 2 yrs	1	No	
Round 7 [51 58]	1987-1988	65+*	Age at last invitation and number of previous invitations		Age at invitation† yrs # % *All women in round 7* 65-69 1,815 53.7 70-74 1,282 42.8 75-79 594 23.9 80+ 187 6.8 *Already in round 6* 65-69 1,569 85.2 70-74 1,057 77.9 75-79 477 68.2 80+ 130 51.8	Invitation and reminder every 2 yrs	1	No	Deaths were classified by a panel of physicians (who were unaware of screening history) based on clinical course of disease and information on serious co-morbidity

* free of breast cancer at first screening invitation at age 65 or older

** Attendance # is the number of women attending; the % represents the # attending / #invited

† attendance of women aged 70+ declined by 9th round, but was stable at about 23%-25% in 4th through 8th rounds

Table 3.2 Results of analytic studies of breast cancer screening for women aged 70+, relative risks (RR), odds ratios (OR) and 95% confidence intervals

Author (year)	Number of Women Study Control	Years of Follow-up	Ages 65-74 RR (95%CI)	Ages 65-69 RR (95% CI)	Ages 70-74 RR (95% CI)	Comments
Randomized Control Trial: Swedish two-county						
Nystrom (1993) [53]	*person-years 70-74* 91,000 64,000	up to 12			**Model used:** Follow-up 0.98 (0.63-1.53) Evaluation 0.94 (0.60-1.46)	RR of dying of breast cancer before 1990, effect of invitation to screen
Tabar (1995) [54]	*women 70-74* 10,339 7,307	up to 14			0.79 (0.51-1.22)	Results adjusted for county
Chen (1995) [55]	*women 65-74* 21,925 15,344	13	0.68 (0.51-0.89) (p < .008)	0.58 (0.39-0.86)	0.78 (0.53-1.20)	RR for 65-69 and 70-74 do not differ significantly
Case Control Study: Nijmegen			**Ages 65+ (to 92) OR (95%CI)**	**Ages 65-74 OR (95%CI)**	**Ages 75+ OR (95%CI)**	
Verbeek (1984) [61]	*Cases* *Controls* 16 80	5	0.81 (0.23-2.75)	N/A	N/A	Adjusted RR are for dying of cancer for ever screened vs never screened.
van Dijck (1994) [58]	33 165	11	0.58 (0.24-1.41)	0.34 (0.12-0.97)	2.87 (0.62-13.2)	Odds ratio is for those attending last screening vs those that did not
van Dijck (1996) [51]	82 410	16	*regular screen* 0.56 (0.28-1.13) *irregular* 0.77 (0.44-1.34)	0.45 (0.20-1.02) 0.64 (0.30-1.38)	1.05 (0.27-4.14) 0.90 (0.40-2.02)	Odds ratio is for regular or irregular screen vs reference group of no screening in index and previous four rounds

(0.63 - 1.53) after 12 years[53] and 0.79 (0.51 - 1.22) after 14 years of follow-up.[54] Only a small number of breast cancer deaths have occurred: 50 in the non-invited and 49 in the invited groups, and a high percentage (25 of 49 deaths) occurred in the 15% of women who had refused both invitations. When these data were examined specifically for women aged 65 to 74, a RR of breast cancer mortality of 0.68 (0.51 - 0.89) was found among those invited vs non-invited.[55] Much of this reduction seemed due to the low RR of 0.58 (0.39 - 0.86) for women aged 65 to 69, as a higher RR of 0.78 (0.53 - 1.20) was found for women aged 70 to 74. (That this RR of 0.78 differs very slightly from the RR of 0.79 above likely reflects slight differences in methods.) While Chen et al interpreted these results as showing no statistically significant difference in risk reduction between women aged 65 to 69 and 70 to 74, it is worth noting that the wide CI around the RR of 0.78 for women aged 70 to 74 is also consistent with *no* risk reduction in this age group.

The Nijmegen studies[59 60] used a nested case-control design, with cases being deaths among the population who had been invited to participate in screening at age 65 and over, and who had been free of breast cancer at their first invitation at or after this age.[61] These analyses were done according to age at invitation to screen (and therefore include women who were first screened before age 65). In the initial report, based on just 16 cases and 80 controls aged 65 and over at first invitation, the relative risk (based on logistic regression analysis for matched sets) of 0.81 (0.23 - 2.75) was weaker than that for women aged 50 to 64 of 0.26 (0.10 - 0.67). This smaller effect for those aged 65 and over was assumed to be due, in part, to the slow growth of breast cancer among older women.[59] This result was calculated using a logistic regression for matched sets with adjustment for residence and marital status.[59]

A subsequent analysis for women aged 65 and over, based on 33 cases and 165 controls, calculated the relative risk (term used by authors) of dying from breast cancer

among women ever screened vs women never screened as 0.58 (0.24 - 1.41).[58] The high OR of 2.87 (0.62 - 13.2) for women aged 75 and over was based on three cases in the screened group; self-selection bias was thought to be responsible.[58] The most recent analysis included 82 deaths and 410 age-matched controls.[51] The breast cancer mortality rate ratio for women regularly screened relative to those not screened was 0.56 (0.28 - 1.13); it was 0.45 (0.20 - 1.02) for women aged 65 to 74, and 1.05 (0.27 - 4.14) for women aged 75 and over (based on three deaths in the regularly screened group). Results were computed using a logistic regression model for matched sets.

In the two most recent reports,[51 58] self-selection biases were found to result in higher breast cancer incidence, and survival from all causes of death, among participants than non-participants. These biases tend to work together to decrease the effect estimates of screening because participants were relatively more likely than non-participants to have the opportunity first to develop, and then to die of, breast cancer. Further, these analyses were conducted not by the age at entry into screening, as most such studies are performed,[61 62] but by age at invitation to any screen from age 65 or older.[58] This design approach assesses the effect of *continuing* (as opposed to *initiating)* mammography screening in the elderly. Further, results are compared based only on the most recent screening rounds. Van Dijck et al[58] make the following argument for this approach: in previous case-control studies, women were classified as being never or ever screened; however, not all previously screened women will have participated in the most recent screening rounds, and it is these screening rounds, which will be more likely to take place in the pre-clinical detection period, that should contribute most to the mortality reductions.[58 62]

3.4 Discussion

Comparison of RCT and case-control results: Results for the Nijmegen case-control study were much more striking (OR = 0.45 {0.29 - 0.70}) for women aged 50 to 74 than for

the Swedish two-county RCT (RR = 0.77 {0.69 - 0.87}).[3] However, these discrepancies can be largely reconciled through an elegant approach developed by Demissie et al, who compared the results of the RCTs and case-control studies to evaluate the efficacy and effectiveness of mammography.[4] This approach is based on the fact that intention-to-treat analyses of RCTs evaluate effectiveness in the population, while case-control studies measure efficacy,[56] with the latter perhaps being of more use for decision-making by an individual woman.[4] The RCTs were compromised in many instances by low compliance rates (50% - 80%) in the invited to screen groups, and by relatively high contamination rates (20% - 30%) screening in the control groups. Demissie et al developed an algebraic expression to reconcile the differences between the RCT and case-control results based on typically reported compliance and contamination rates. When analyses are adjusted for the cross-over effects, an observed RR of 0.80 translates into a 'true' RR of 0.63, with a 70% compliance rate and 20% contamination rate, or a 'true' RR of 0.44 with a 60% compliance rate and 30% contamination rate. The RRs in these examples are similar in magnitude to those observed for women aged 70 to 74, with RR of 0.77 and OR of 0.45. (In the Swedish two-county study, Andersson[63] reports that 13% of women had undergone outside mammography, while Gullberg reports that 25% of controls in the Malmo study had mammograms.[56])

Validity of meta-analysis/systematic reviews of observational studies: For the purposes of this thesis, a meta-analysis is defined as including a quantitative summary of the outcome measure(s) of interest in addition to following all of the steps necessary to conduct a systematic review. This thesis chapter uses systematic review techniques, but is not a meta-analysis because it was judged inappropriate to combine the results from the two studies. Development of meta-analysis and systematic review techniques to synthesize known information has led to increased application of meta-analysis to observational studies. While

meta-analysis is well-established as a method to summarize results of RCTs and reduce chance variability,[65 66] its use for observational studies has been criticized[65] due to the potential to produce spuriously precise, statistically significant estimates that do not properly account for variation due to non-random factors such as confounding and selection bias, and therefore limit interpretation of results.[67] Greenland notes that confidence intervals represent only the minimum uncertainty in the data. When using them in observational studies as estimates of causal effects, four basic assumptions inherent in using confidence intervals are not fulfilled.[68] Observational studies: do not, by definition, ensure randomized exposure; seldom control for all measurement errors as data are collected from questionnaires or medical records; (for case-control studies) rarely satisfy the need for no uncontrolled selection bias; and seldom discuss the need for no mis-specification of the underlying statistical model. These problems can be partially overcome by avoiding the synthetic combination of data into a single summary estimate, and by focusing instead on use of meta-analysis to assist in critical evaluation of studies, comparing patterns among studies,[68 69] and on examining possible sources of heterogeneity.[65] Pettiti argues that meta-analysis of experimental studies must be done with great care, and along with Shapiro, questions the use of criteria to rate the quality of studies to be included or not. However, she notes that the documentation of differences among studies in dose levels, analysis categories, and so on, are important observations that could explain inconsistent results between studies. Pointing to the greater academic rigour of meta-analysis compared to narrative literature reviews, and the need to 'synthesize' information to make sense of existing information, she recommends that the method be developed further.[68 70]

Despite concerns, methodologic development continues for meta-analysis for screening studies in particular,[71] and observational studies in general.[66 71] Walter presents the first systematic review of meta-analyses of screening studies, including assessment of

methodologic quality and statistical methods. Of the 57 meta-analyses identified, 48 were based at least in part on observational study designs, 19 measured clinical outcomes generally using odds ratios, and 9 were for breast cancer screening. Walter concluded that meta-analysis, even based on weak or inadequate data, was an improvement over reliance on expert panels, but called for improvements in methods. Care must be taken to understand possible causes of inter-study heterogeneity and the potential biases and limitations of the original study data. Egger further recommended that systematic reviews should include a study protocol designed in advance, comprehensive literature searches, a list of studies selected, and data extracted in a reproducible and objective fashion.[65] Stroup et al report on the results of a workshop of US and Canadian experts who developed a detailed checklist of guidelines intended for authors, reviewers, editors, readers and decision-makers. These included background information related to study design, details of reporting of search strategy, reporting of methods, reporting of results, reporting of discussion, and reporting of conclusions.[66]

Case-control studies: Use of case-control methodology is a relatively new technique to assess screening efficacy. Clarke and Anderson published the first such study in 1979 to assess the efficacy of Pap smears to prevent invasive cervical cancer, arguing that, where an RCT was not possible, this approach could still provide convincing evidence.[72] Since that time, numerous biases and methodologic flaws have been reported; for example, Friedman and Dubin[73] describe *subject definition bias, self-selection bias, lead-time bias* and *healthy screenee bias,* while Thompson[8] also identifies *length-biased sampling,* where screening will differentially pick up slow-growing tumours.

Weiss discusses the major pitfalls in a review article.[74] Cases must be selected carefully to avoid *subject definition bias*, in consideration of the goal of the study; if the goal is to prevent mortality or other adverse consequence such as advanced disease, then cases

must be people who died or developed the adverse outcome. Second, controls must be equally carefully selected, preferably as a sample from the same population that generated the cases. Third, ascertainment of screening history should ideally be based on review of medical records to avoid *misclassification* of diagnostic tests, or recall bias for the cases and controls. This approach also helps to avoid *healthy screenee* bias, as it ensures that controls do not accumulate examinations for which cases are ineligible. Results of both negative and positive tests should be included. Fourth, the screening status should be categorized in a way that maximizes the likelihood of the screen occurring within the *pre-clinical detection period* for cases, as inclusion of screens outside this period can lead to odds ratios spuriously close to the null. Etzioni and Weiss[75] use results from a simulation model to conclude that using a longer time estimate for the pre-clinical detection period (or sojourn time) is preferable to using a shorter or mean length, as under-estimation leads to greater bias.

Fifth, various confounders must be controlled. One example is *self-selection bias* that occurs when symptomatic women, or those who are at higher risk, preferentially attend, or do not attend, screening. An example occurred in the HIP study of breast cancer screening, where of those invited, women who attended had higher breast cancer incidence than women who did not. Confounding may also be controlled by analytic methods such as stratification, matching or adjustment. Finally, results of case-control studies of screening must be interpreted carefully. An example is to ensure that positive findings at screening are followed up and treated; otherwise women with abnormal screening results may still develop and die of breast cancer.

Increasing awareness of these biases over the past 15 to 20 years has clearly influenced the design of the case-control studies of the Nijmegen trial, as the control groups have become more and more carefully designed. Even so, where screening participation rates remain low among the oldest women in the population, study authors must consider the

difficulties of fully accounting for self-selection bias of women who may be at higher or lower risk of developing breast cancer, and of length-based sampling resulting from differential detection of slow-growing tumours.

3.5 Implications

Given the above concerns, it was deemed inappropriate to pool the results of the two studies. Instead, the results presented in Section 3.3 compare and contrast the screening methodology utilized and assess reported limitations of the various studies. One criticism of the Swedish two-county RCT, the 'small' sample size of 17,000+, seems unjustified, as a sample size of about 10,000 per arm should be sufficient to identify a difference of 40% in mortality between the two groups. What is puzzling is the unexplained drop in the RR from 0.98 reported after 12 years of follow-up, to 0.79 after 14 years of follow-up. This may reflect undocumented methodological differences with respect to age-standardization noted by Gotzche. Overall, the results of this systematic review, while suggestive of a positive effect of 20%, are not conclusive. Therefore, Criterion 1 is only partially met. Other criteria will be used to assess the potential effect of screening women aged 70 to 79 for breast cancer. Criteria 2 to 4 will be considered in Chapter 4 and Criterion 5 will be evaluated through modelling the benefits and harm in Chapter 5.

CHAPTER 4.

BURDEN OF CANCER, ACCEPTABILITY OF TEST, AND AVAILABILITY OF TREATMENT

In this chapter, empirical Canadian data are analyzed to assess the three criteria related to the burden of cancer, the acceptability of the test and the availability of treatment. Data sources used in this chapter are described in Appendix C.

4.1 Criterion 2: What is the burden of breast cancer in women aged 70 to 79?

As is true for women aged 40 to 69, breast cancer remains an important health concern for elderly women. Older Canadian women may be particularly vulnerable as they experience the second highest rate for breast cancer, and have one of the highest life expectancies, of any national population. Breast cancer incidence rates increase with age. up to ages 70 to79, then decline slightly among women aged 80 and over. By contrast, mortality rates increase steadily up to the oldest age groups.[76] Among women aged 70 to 79, breast cancer accounted for 3,524 (or 24%) of all newly diagnosed cancers in 1996 (Table 4.1).

Overall, 21% of breast cancers occur in this age group, a percentage that is similar to that for women aged 50 to 59 and 60 to 69. The 1,238 breast cancer deaths occurring among women aged 70 to 79 constituted 15% of all cancer deaths, and 5% of all deaths. Women aged 70 to 79 account for the largest proportion (25%) of breast cancer deaths in any ten-year age group. Notably, women aged 70 to 79 have the highest probability (at 3.2%) of any 10-year age group of developing breast cancer in the next ten years. This probability is only slightly greater than the 2.9% for women aged 60 to 69, and moderately above the 2.3% reported for women aged 50 to 59 and the 2.2% for women aged 80 to 89.[77]

One difference between older and younger women is the increasing importance of other causes of death. Breast cancer is the leading cause of death for women aged 40 to 49

Table 4.1 Female breast cancer incidence and mortality compared to all cancers in women, by ten-year age groups in Canada, 1996

Age	0-29	30-39	40-49	50-59	60-69	70-79	80+	All Ages
New cases				***Incidence***				
all cancers (N)	1,425	2,914	6,658	9,329	12,951	**14,594**	9,558	57,429
breast cancer (N)	67	955	3,001	3,541	3,835	**3,524**	1,860	16,783
% of all cancer in age group	*4.7*	*32.8*	*45.1*	*38.0*	*29.6*	***24.1***	*19.5*	*29.2*
% of all breast cancers	*0.4*	*5.7*	*17.9*	*21.1*	*22.9*	***21.0***	*11.1*	*100*
Number of deaths				***Mortality***				
all causes (N)	2,392	1,557	3,475	5,898	12,016	**25,194**	53,137	103,669
all cancers (N)	223	520	1,747	3,314	5,646	**8,176**	7,516	27,142
breast cancers (N)	6	158	566	845	971	**1,238**	1,161	4,945
% breast cancer of all cancer deaths	*2.7*	*30.4*	*32.4*	*25.5*	*17.2*	***15.1***	*15.4*	*18.2*
% breast cancer of all deaths	*0.3*	*10.1*	*16.3*	*14.3*	*8.1*	***4.9***	*2.2*	*4.8*
Rank within age group		*2*	*1*	*1*	*3*	*5*	*10*	
% breast cancer deaths by age	*0.1*	*3.2*	*11.4*	*17.1*	*19.6*	***25.0***	*23.5*	*100.0*

Source: Canadian Cancer Registry - ORIUS extract, March 18, 2001

and 50 to 59, accounting for 16% and 14% respectively of all deaths in each age group, and ranks third, accounting for 8% of all deaths at age 60 to 69. However, by age 70 to 79, breast cancer ranks as the fifth leading cause of death, comprises 5% of all deaths, and is superceded by diseases of the heart (27%), stroke (7.8%), lung cancer (7.6%) and chronic obstructive lung disease (5.1%). By ages 80 to 89 and 90 and over, breast cancer ranks as the tenth leading cause of death, and accounts for 2.5% and 1.5%, respectively, of all deaths.

In summary, the criterion that breast cancer is a significant health problem for women aged 70 to 79 is met, based on recent statistics for Canadian women.

4.2 Criterion 3: Is there a valid, acceptable screening test that will reach a high proportion of women aged 70 to 79?

Validity of screening test: Due to the lower prevalence of benign breast disease and to less dense breast tissue among older women, screening mammography is more sensitive and specific among women aged 70 to 79, and together with their higher breast cancer rates, this leads to a higher positive predictive value (PPV). Faulk analyzed data from 32,140 mammograms performed in women 50 and older, comparing women aged 50 to 64 with those aged 65 and over. He concluded that "mammographic screening is at least as effective in detecting cancer for which there is a favourable prognosis in women aged 65 years and older as it is in women aged 50 to 64 years".[78] Hwang et al compared results of mammography over a 15-year period for women aged 40 to 69 with those 70 and over (N=1,001). They found increasing percentages in each age group of both smaller tumours (T1a and T1b) and cases diagnosed by mammography, and concluded: "the potential benefit of regular mammography to healthy women aged 70+ may equal that observed in their younger counterparts" [79]

Acceptability: Poor participation rates have been cited as a reason for not offering screening to this age group. Forrest found poor attendance among older UK women in the mid-1980s,[80] and results from the Swedish two-county trial for women aged 75 to 79 were not published due to low participation. However, these low rates accrued in trials that began entering women in the late 1970s, when population-based screening was an innovative, largely untried procedure. By contrast, in Canada, biennial mammography utilization rates of 40% to 50% for screening in both organized and outside of organized programs were reported for women aged 70 to 79 in two provinces by 1994.[81]

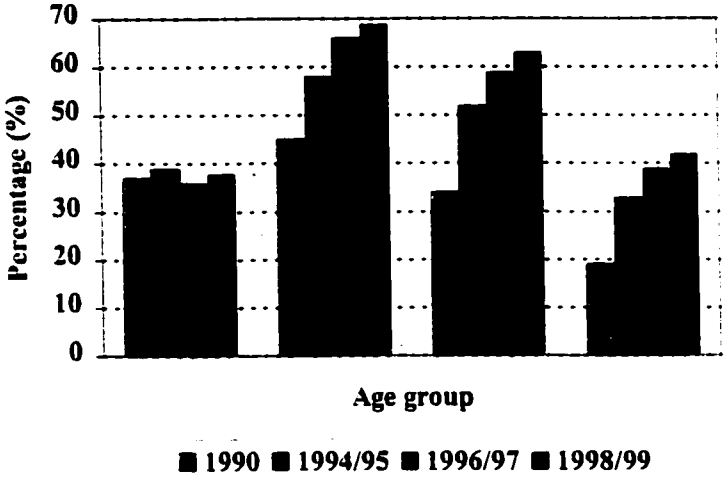
Selection bias: Self-selection into screening, and screening compliance is generally higher in women who are subsequently diagnosed with breast cancer.[51 82 83] By contrast,

among Swedish women aged 70 to 74, most deaths in the invited group occurred in the 15% who refused both invitations, suggesting that women with breast cancer were *less* likely to be screened initially.[54] Thus, breast cancer status at time of screening shows an inconsistent effect on selection bias and participation in screening among elderly women.

Influence on outcomes: For early detection to affect outcomes, the new cancers should be diagnosed at an early stage where they are still treatable, and there should be an absolute reduction in the rate of cancers diagnosed at later stages. Chu et al examined trends in breast cancer incidence, mortality and survival by age group and extent of disease (local, regional and distant metastases) in US women.[84] The observed decrease in diagnosis of regional disease in the late 1980s in women over age 40 was thought to reflect the increased use of mammography earlier in the 1980s. However, the increase in survival rates, particularly for regional disease, was interpreted as being due to a combination of improved therapy as well as early detection.

Screening participation in Canada: Significant proportions of Canadian women aged 70 to 79 participate in screening. The percentage of women in this age group who self-report being screened in the past two years increased from 36.7% to 44.2%, according to the National Population Health Surveys (NPHS) for 1994/95 and 1998/99 respectively (see Figure 4.1). In organized screening programs, 10% of women aged 70 to 79 participated in biennial screening (rates calculated over a two-year period), compared to about 20% of women in the target group aged 50 to 69 (Table 4.2). Rates for women aged 80 and over, however, plummeted to 1.6%. Not surprisingly, the proportion of first screens was lower in older women (28%) compared to women in their 60s (37%) and 50s (48%). The annual rate of fee-for-service (FFS) mammograms per 100 women more than doubled from 7.0 to 15.8 between 1988 and 1997 respectively. As the NPHS is known to over-report screening, attempts have been made to estimate screening rates by combining information

Figure 4.1 **Percentage of women having a mammogram in past two years by age, Canada, 1990-99**



Source: **Health Promotion Survey, 1990 and National Population Health Survey, 1994/95, 1996/97 and 1998/99**

Table 4.2 **Participation in breast screening in Canada, 1997-1998**

	Age Group				
	40-49	**50-59**	**60-69**	**70-79**	**80+**
Population: 1997	2,307,183	1,586,252	1,207,684	968,422	543,681
CBCSD: 1997 and 1998					
Number of screens	155,670	330,211	243,884	103,525	8,740
First screens	70,780	158,300	90,701	29,078	n.a
% first screens	*45.5*	*47.9*	*37.2*	*28.1*	n.a
Estimated number of biennial mammograms per 100 women	*6.7*	*20.8*	*20.2*	*10.3*	*1.6*
FEE-FOR-SERVICE (FFS)					
Number of mammograms per 100 women per year					
1988	13.3	15.0	11.6	7.0	n.a.
1997	16.9	30.4	26.2	15.8	n.a.
Doubled to approximate biennial screening	*33.8*	*60.8*	*52.4*	*31.6*	*n.a.*
Estimated total number of biennial mammograms per 100 women (1997-1998)	**40.5**	**81.6**	**72.6**	**41.9**	**n.a.**

Source: OECD tables, and Organized Breast Screening in Canada, 1997 and 1998 report

from the CBCSD (for screening in organized programs) and FFS billings (for mammography provided outside of organized programs). The FFS data can be doubled to approximate thebiennial screening rate, noting the limitations that they include both diagnostic and screening mammograms, they may be incomplete, and they will count a given woman more than once–for example, she may have both a screening and a diagnostic mammogram, or may be screened every year. The combined mammography rate of the Canadian Breast Cancer Screening Database (CBCSD) plus the FFS data indicate that up to 40% of Canadian women aged 70 to 79 could be receiving mammography every two years. These figures may need to be adjusted downward to account for women who receive more than one mammogram during this time period and for the 20% of diagnostic mammograms in the fee-for-service data.[31]

Interim outcome measures were generally favourable for women aged 70 and over (Table 4.3). The abnormal recall rate for first and re-screens was similar to that for other age groups. The proportion of women with abnormal screens who underwent various diagnostic procedures is very similar for women aged 70 to 79 as for women aged 50 to 69, although women aged 70-79 may be somewhat more likely to have a core biopsy. Of the cancers detected, fewer were *in situ* – just 17.4% compared to 22.1% for women aged 50-59, but similar to the rate of 17% in women aged 60-69. The cancer detection rate is substantially higher for both first and re-screens, resulting in a more favourable PPV for an abnormal screen. Finally, benign to malignant biopsy ratios are lower than those found for any other age group.

Stage distribution pre- and post-screening implementation in Canada: The distribution of breast cancer T stage before screening implementation was compared for women aged 50 to 69 and 70 to 79 using the Saskatchewan Cancer Registry (1975-1984) and the Northern Alberta breast registry (1971-1984) (Table 4.4). Stage was missing (Tx) in 9%-10% of Northern Alberta and in 6%-8% of Saskatchewan records. T1 stage (as T1x) was

Table 4.3 Results of organized breast screening in Canada, 1997 and 1998

	Age Group			
	40-49	**50-59**	**60-69**	**70-79**
Abnormal recall rate (%):				
First screens	11.1	11.5	10.4	11.3
Rescreen	5.4	6.4	6.0	5.9
Diagnostic procedure (%):		***50-69***		
diagnostic mammogram		70.3		71.2
ultrasound		42.9		38.8
fine needle aspiration		5.0		5.1
core biopsy		5.2		5.1
open biopsy (may involve fine wire localization)		13.3		17.8
Number of cancers	326	1,331	1,362	855
Invasive	222	1,037	1,131	706
In situ	104	294	231	149
% in situ	*31.9*	*22.1*	*17.0*	*17.4*
Cancer detection rate per 1000 screens:				
First screens	2.6	5.6	8.7	12.9
Rescreen	1.8	3.5	4.8	6.9
Positive Predictive Value of abnormal screen:				
	2.5	5.0	8.8	12.2
Benign to malignant ratio:				
open biopsy	4.5:1	2.0:1	1.2:1	0.8:1
core biopsy	4.6:1	2.8:1	1.5:1	0.8:1

Source: Organized Breast Cancer Screening Programs in Canada, 1997 and 1998 Report, plus special tabulations from the Canadian Breast Cancer Screening Database.

unavailable in 5%-6% of Northern Alberta records, and in 34% of Saskatchewan records, hampering comparisons between the registries. Within each registry, no difference in stage distribution was found between the two age groups based on a chi-square test, suggesting that both age groups experience similar biology of breast cancer.

The CBCSD was used to assess the distribution of T stage for breast cancers detected after screening implementation (Table 4.5). Stage distributions were compared separately for breast cancers detected at first screen and at re-screen. Using the chi-square test, statistically

Table 4.4 Breast cancer T stage before screening implementation by age group, women aged 50-69 and 70-79, Northern Alberta and Saskatchewan

	Northern Alberta, 1971-1984		Saskatchewan, 1975-1984	
Age (y)	**50-69**	**70-79**	**50-69**	**70-79**
No. of cancers	***2,441***	***691***	***1,976***	***743***
Stage:	**%**	**%**	**%**	**%**
Tis	1.9	2.3	3.0	2.7
T1a	1.1	0.9	0.4	0.3
T1b	8.6	8.1	4.0	2.6
T1c	20.5	19.8	10.8	11.0
T1x	6.5	4.5	34.2	33.0
T2	32.7	36.9	30.3	32.0
T3	5.2	4.2	4.8	4.3
T4	13.5	14.5	6.3	6.1
Tx	10.2	8.8	6.1	8.1
Total	**100.0**	**100.0**	**100.0**	**100.0**
T2+	**51.3**	**55.6**	**41.4**	**42.4**
	$\chi^2(8) = 9.84$, p=0.276		$\chi^2(8) = 7.67$, p=0.466	
	$\chi^2(6) = 7.32$, p=0.293		$\chi^2(6) = 6.92$, p=0.329	

Note: Chi-square tests compare stage distribution between 50-69 and 70-79, showing degrees of freedom in brackets. Two tests were calculated with the breakdowns including T2+[$\chi^2(6)$] and T2, T3, and T4 separately [$\chi^2(8)$].

significant differences were found between the two age groups in both screening groups. However, these differences were judged to be of relatively minor clinical importance because: 1) there was no real pattern to the categories showing differences; and 2) the largest differences for first screens were found for Tis (17.4% vs 11.0%) and T1c (33.0% vs 39.1%) and for re-screens were for T1b (24.2% vs 30.1%), Tis (17.4% vs 13.9%) and T2+ (12.2% vs 9.6%) in the younger and older age groups respectively. By contrast, Saskatchewan data, which include breast cancers diagnosed in both screened and unscreened women, show a small, but fairly consistent pattern of earlier stages occurring more often among women aged

Table 4.5 Number and percent distribution of T stage for breast cancers detected after screening implementation by screening status and by age group, selected Canadian sources, women aged 50-69 and 70-79

	CBCS Database, 1988-96, First Screen		CBCS Database, 1988-96, Re-screen		Saskatchewan, 1990-1997 (screened and unscreened)	
Age (y)	**50-69**	**70-79**	**50-69**	**70-79**	**50-69**	**70-79**
No. of cancers	***3,847***	***912***	***2,103***	***740***	***2,394***	***1,256***
Stage:	**%**	**%**	**%**	**%**	**%**	**%**
Tis	17.4	11.0	17.4	13.9	13.5	8.3
T1a	6.4	5.3	8.2	7.8	4.3	3.4
T1b	21.1	23.8	24.2	30.1	20.0	17.4
T1c	33.0	39.1	33.0	34.7	32.7	34.6
T1x	0.2	0.0	0.2	0.1	1.2	2.1
T2	15.1	14.4	11.0	7.8	20.6	24.3
T3	1.2	1.2	0.7	1.1	1.9	1.8
T4	0.6	0.9	0.6	0.7	4.0	4.8
Tx	4.9	4.4	4.9	3.6	1.9	3.3
Total	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
T2+	**16.9**	**16.4**	**12.2**	**9.6**	**26.5**	**30.9**
$\chi^2(6)$	32.95, p=0.000		17.5, p=0.008		41.0, p=0.000	
$\chi^2(8)$	33.83, p=0.000		20.0, p=0.007		41.6, p=0.000	

Note: Chi-square tests compare stage distribution between 50-69 and 70-79, showing degrees of freedom in brackets. Two tests were calculated with the breakdowns including T2+[$\chi^2(6)$] and T2, T3, and T4 separately [$\chi^2(8)$].

50 to 69 and later stages occurring among women aged 70 to 79. Data were reasonably complete; just 4% to 5% of cases in each data set were missing T stage.

Data for Saskatchewan women aged 70 to 79 show that a strong shift to lower stages occurred between 1975-1984 and 1990-1997, with 42.4% of breast cancers in the early time period and just 30.9% in the later time period being stage T2+. The proportion of *in situ* cancers rose from 2.7% to 8.3%, while the proportion of T3 and T4 cases declined. Among women whose breast cancers were detected by screening, stage was distributed even more favourably, with smaller, lower-staged cancers accounting for a higher proportion of cases.

These results provide positive support for the criterion that the mammography screen test is valid and acceptable. Even in the absence of targeted screening in most provinces, mammography screening is reaching a relatively high proportion of women aged 70 to 79. The more favourable cancer detection rates, lower *in situ* and lower benign to malignant biopsy ratios reflect the lower incidence of benign breast disease and less fatty breast tissue, as reported by others. Further, stage shifts are occurring in the expected direction (i.e., towards smaller tumour size).

4.3 Criterion 4: Are there efficacious treatments for women aged 70 to 79?

Treatment efficacy: Few clinical trials include women over age 70. The Canadian Clinical Practice Guidelines for the Care and Treatment of Breast Cancer note the paucity of evidence to support therapy choices for women over 70, and that "with increasing age or frailty, chemotherapy is less well-tolerated."[85] Otherwise, these guidelines do not recommend different treatment options for elderly women. Silliman et al point out that co-morbidity, impaired functional status, lack of social support and differences in physiology may influence treatment efficacy and effectiveness, and question the extrapolation of study findings from younger to older women.[86] Substantial variations in treatment by age have been reported. Goodwin et al found that regional variations in mortality rates for breast cancer are due primarily to variation in rates for older US women, with higher mortality found in regions with less aggressive treatment for the older age groups.[87]

Compliance with treatment guidelines: The standard initial treatment for most breast cancers diagnosed in Canada is either mastectomy, or breast conserving surgery (BCS) followed by radiation therapy. In addition, further adjuvant chemotherapy and/or hormonal therapy is provided depending upon the age of the patient, the stage of the cancer, and other factors such as tumour grade and hormone receptor status. As breast cancer is a potentially

life-threatening disease, and treatment is effective for cancers diagnosed at an early stage, most women will likely comply with the treatment recommended by their physicians. However, women with poor health status or significant co-morbidities might be more likely to choose (or to have their physician recommend) less aggressive treatment that avoids radio- or chemotherapy. For example, tamoxifen may be used to control the growth of breast tumours.

Canadian data: Treatment patterns in Canada were explored using in-patient surgical treatments, that are well documented in hospital morbidity data (Table 4.6). The

Table 4.6 Rate per 100 incident breast cancers, of hospital separations of women diagnosed with breast cancer, who received breast conserving surgery (BCS) or mastectomy as in-patients by age group, Canada, 1995-1996

Age (y)	40-49	50-59	60-64	65-69	70-79	80+
BCS	45.1	44.9	42.1	41.7	37.7	28.8
Mastectomy	38.9	38.5	39.8	39.3	42.2	39.6
Sub-total any surgery	84.0	83.4	81.9	81.0	79.9	68.4
"No" surgery*	16.0	16.6	18.1	19.0	20.1	31.6

Note: Percent is calculated by dividing hospital separations by the number of diagnosed cases.

* Calculated as the difference between incident cases and the sub-total of all surgery; includes an estimated 10% of day surgeries performed for BCS.

Source: Hospital morbidity file, Statistics Canada; and Canadian Cancer Registry, as extracted from data prepared for a special study for the OECD.

percentage of women receiving mastectomy tends to increase slightly with age, while those receiving BCS or no surgery decreases slightly with age, and most sharply for the 80+ age group. Unfortunately, this data base contains few data on radiation or systemic therapies, while day surgery treatments, such as are done for some breast conserving surgery, are not

available. Notably, only modest differences are found between the 65-to-69 age group and the 70-79 group, suggesting that most older women are able to tolerate reasonably aggressive treatment. Women in the no-surgery category may have received day surgery for BCS, or may have advanced disease where surgery is not performed. In the oldest age group some women with less advanced disease may also opt for no surgery, if the cancer is progressing so slowly as to be unlikely to cause problems in comparison to other more serious health concerns. These data provide some evidence that treatments proven effective for younger women are being used by women aged 70 to 79, given that their surgical patterns are very similar to those for younger women. However, more research is needed to provide a definitive conclusion.

4.4 Summary

Results presented relevant to the criteria regarding the burden of breast cancer in women aged 70 to 79, and the acceptability of the test for this age group show strong support for continuing screening. Evidence for the criterion of availability of an efficacious treatment shows moderate support, with an important limitation being the lack of inclusion of women over 70 in clinical trials. The next chapter will address the population-based benefits and harms using a model-based approach.

CHAPTER 5.

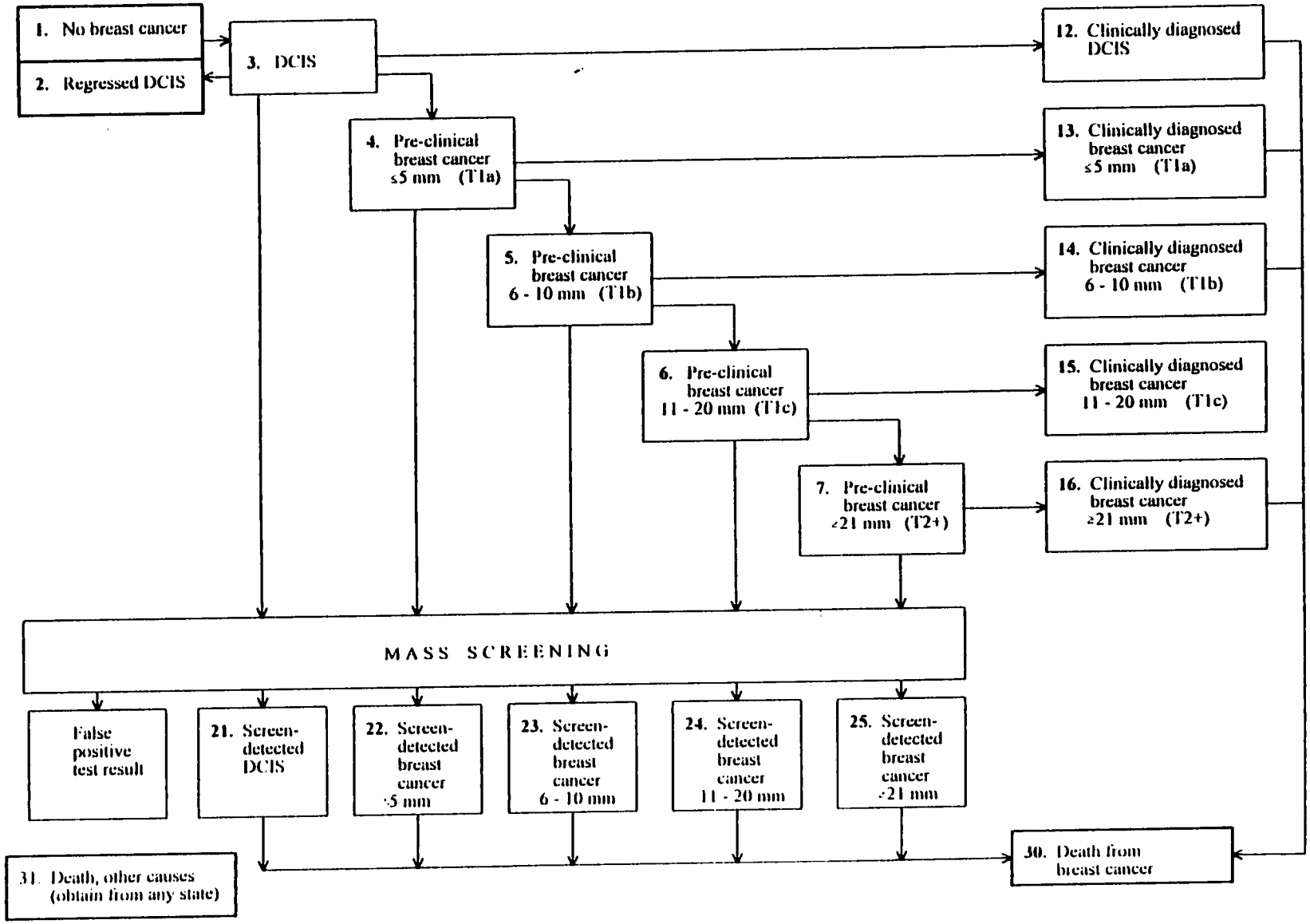
CRITERON 5: BENEFITS AND HARMS OF SCREENING WOMEN AGED 70 TO 79 USING THE MISCAN MODEL

5.1 The MISCAN Model

The MISCAN model was chosen to integrate Canadian information on demographic, mortality and cancer incidence data, screening parameters available from provincial programs, along with other data, and thereby quantify the expected impact of continuing breast cancer screening among elderly Canadian women. MISCAN is a well-validated and comprehensive model that has been used in a variety of populations and is available for third-party use. Other approaches were considered, including multi-state life tables. However, these approaches generally required making assumptions about transition probabilities or other parameters that would be less well-validated and less well understood than assumptions made in MISCAN.

MISCAN uses Monte Carlo microsimulation methods to produce outputs on the effects of screening at the individual and population level and is based on models of the natural history of breast cancer and the impact of screening on the natural history.[88] The breast cancer model has been well described.[24 89 90 91 92] It first simulates life histories for individual women in the absence of disease based on demographic characteristics of the population (births, life tables, deaths from causes other than breast cancer) and then generates incidence and mortality for breast cancer. Each simulated woman has a date of birth and an age of death from other causes.[92] In MISCAN, women first reside in a state of "No breast cancer" until a transition occurs to one of five preclinical states from which breast cancer can be detected by mammography screening (Figure 5.1). Preclinical states include DCIS, T1a, T1b, T1c, and T2+ (the latter four states generally correspond to tumour sizes of ≤5 mm, 6-10 mm, 11-20 mm, and breast cancer ≥21 mm respectively, except that T4 tumours may be

Figure 5.1 Structure of MISCAN model for breast cancer



of any size – see also Appendix C). A cancer may be detected at screening, may be diagnosed clinically in any state, or if undiagnosed, may progress to the next pre-clinical state.[25] The probability of undergoing a given transition is governed by a probability distribution for the duration (or dwelling time) in each state, which may be dependent upon age. Screened women may also enter a false-positive state when no cancer is found after a positive screening examination.

The two end states of the model are "deaths due to breast cancer" and "death from other causes".[89] Deaths from other causes are determined based on input data calculated by subtracting breast cancer deaths from all causes of death. Breast cancer survival is determined from population-based cancer registries or other sources for clinically detected cancers for a defined population (i.e., Dutch, Swedish, or North of England), and adjusted for survival of screen-detected cancers.[25 29 89 91] The duration in each state follows an exponential distribution. Age-specific assumptions for the mean duration of the preclinical states have been validated with data from the Dutch screening trials and initially ranged from 1.8 years at age 35 to 6.2 years at age 70.[91] Sojourn time is thought to continue to increase with age after 70, although few data sources exist to reliably determine sojourn times in the elderly. Thus, MISCAN uses two assumptions for modelling screening in women over 70: an *optimistic* variant that assumes no further increase after age 65 years, and a *pessimistic* variant that assumes sojourn time continues to increase up to 8.7 years at ages 80 to 84.[24]

The tumour diameter determines several key parameters of model performance, including the probability that breast cancer will be detected by screening (sensitivity), and the improvement in survival for screen-detected cancers.[89 91] Sensitivity depends on stage, and is set at 40%, 65%, 80%, 90% and 95% for DCIS, T1a, T1b, T1c, and T2+ for women aged 50 and over.[24 91] Improvement in prognosis after a screen-detected diagnosis of breast cancer is modelled by applying results from the Swedish trials to baseline survival as follows: DCIS,

1.0; T1a, 0.80; T1b, 0.73; T1c, 0.51; and T2+, 0.35. These numbers are calculated as 1 minus the ratio of the risk of dying of screen-detected cancer to the risk of dying if the cancer had been diagnosed in the absence of screening. (Thus it equals 1-RR and is equivalent to the exposed prevented fraction or the relative risk reduction.) In one study, a sensitivity analysis was performed by assuming a 15% lower 'cure' rate for all screened women than in the Swedish trials.[43]

5.2 Analytic Approach

Primary Research Question:

1. What are the differences in population health impacts of continuing to provide organized screening to women aged 70 to 79 as compared to ending screening at age 69?

Secondary Research Questions:

1. What is the ratio of additional cancers detected to deaths prevented, or to late-stage disease avoided?
2. Does screening result in improved life expectancy and quality-adjusted life-years (QALY)?

Modelling approach: The MISCAN model was first adapted to the Canadian population by developing and entering input data for Canadian births, deaths and breast cancer incidence. Five screening patterns were developed ranging from no screening to 100% participation, and including three patterns with intermediate participation rates. To reflect national recommendations for screening participation, the *target* model aimed to screen 70% of women in each age group. Two other models were set up to approximate to screening participation in 1990 and 1996. Models were then validated by comparing actual Canadian data to MISCAN estimates by age group for population counts, breast cancer

incidence and breast cancer mortality, as well as stage distributions. Analytic models were run assuming screening started in 1988 and continued biennially for 30 years for women aged 40 to 79 and 40 to 69. The difference between the two models was used to assess the impact of screening women aged 70 to 79. Sensitivity analyses included running the full model using a pessimistic sojourn time. Further details of models used are described in Section 5.4.3.

Outcome measures: MISCAN was used to produce the following outcome measures for screening volumes and population health variables, calculated by each year of screening and/or totalled over the whole simulated time period.

- number of first and repeat screens.

Selected disease-related measures include:

- number of cancers detected by screening by stage of cancer,
- number of breast cancer deaths and number of breast cancer deaths prevented,
- number of life-years lived by the population and in clinical and screen-detected stages,
- number of life-years gained and lived in lead-time,
- number of quality-adjusted life-years.

Model results were used to derive the following population health indicators:

- number of women biopsied per 10,000 women screened, and per life saved,
- number of women screened to find one cancer, and to prevent one death,
- changes in rate of advanced disease in screened vs unscreened group,
- changes in quality-adjusted life-years in screened and unscreened groups,
- ratio of extra cancers detected vs deaths prevented.

5.3 Development of Input Data for MISCAN Model

This section describes how input data were developed to adapt the MISCAN model to the Canadian situation. Data for 1996 were selected as most data series were available up to at least that date; as well 1996 was a census year. Canadian data bases used included: Vital Statistics Database (mortality); Canadian Cancer Registry (cancer incidence); Canadian population estimates; and Canadian Breast Cancer Screening Database and fee-for-service data (estimates of screening mammograms). Parameters calculated included cumulative mortality for all causes of death and for deaths other than breast cancer; cumulative probability of births by year; pre-screening incidence rates; and screening participation. Detailed descriptions of source data are provided in Appendix C.

5.3.1 Calculation of Cumulative Probability of Death Using a Life Table Approach

The MISCAN model requires a life table containing the *cumulative probability of death from other causes* for specified ages. The probability distribution of the age of death is specified by attributing to a given age *the probability of having died before this age.* The probability of death before the first age must be 0.0 and before the last age (100 yrs) must be 1.0; the maximum life span never exceeds 100. Data for the life table are entered in the following format: the number of the life table; age; and cumulative probability of death before the given age. The maximum of 40 values permitted for age allows for values by five-year age groups up to about age 75 and single year of age up to 100 and over. (This probability is used to determine the date of death on each simulated record in the absence of breast cancer. Breast cancer mortality is then superimposed on the record based on cumulative incidence probabilities and survival rates for clinical and screen-detected breast cancers.)

The MISCAN documentation does not provide a method for calculating the cumulative probability of death for all causes other than breast cancer. After consulting a

number of references on life tables,[93-95] the following method described by Selvin was selected.[96] The input data calculated over from 1995-1997 are compared to the Netherlands (Table 5.1); for detailed calculations see Appendix D.

1. Determine counts of all deaths, *dx*, and for other deaths (i.e, all deaths minus breast cancers), *dx(i)*, for a three-year period where *x* represents five-year age groups up to age 59 and then single year of age up to age 100+.
2. Calculate age specific rates for all deaths, *mx*, and for other deaths, *mx(i)*.
3. Calculate $qx = 2a*mx/(2 + a*mx)$ and $qx(i) = 2a*mx(i)/(2 + a*mx)$ where a = width of age group and mx = death rate (equivalent to Selvin, 10.32 and 10.34, p 340-1).[93]
4. Calculate the number of survivors, *l*, and the number of life-table "deaths" for all deaths, $dx = lx*qx$, and for other deaths, $dx(i) = lx*qx(i)$, starting with 100.000 survivors at age 0. (Selvin, 10.37; p 342)
5. Calculate expected deaths after age *x*, as *Wx(i)* for other "deaths" and *Wx(b)* for breast cancer "deaths". The columns for *Wx* accumulate the sum of all "deaths" due to a given cause, from age *x* upwards. These sums represent the number of individuals who reach age *x* and will ultimately die of a given cause. (Selvin, 10.38, p. 343)
6. The cumulative numbers of deaths provide the values necessary to calculate the probability of death before age *x* for each cause. Thus, for the i^{th} cause, $Fx(i) = 1 - \{Wx(i)/W0(i)\}$, is the probability of dying before age x.

5.3.2 Calculation of the Cumulative Probability of Births before a Given Year

MISCAN requires a probability distribution of simulated records by birth year to generate the observed population. Because 1996 was chosen as the model base year, the cumulative probability was calculated ranging from 0.000 in 1896 up to 1.000 by 1996. This

Table 5.1 Input data for MISCAN model for cumulative mortality before a given age, Canada compared to the Dutch model used for the Netherlands

	Canada	Dutch Model		Canada	Dutch Model
Age	Cumulative Probability		Age	Cumulative Probability	
0	0	0	85	0.5193	0.5260
5	0.0061	0	86	0.5580	
10	0.0068	0	87	0.5979	
15	0.0076	0	88	0.6384	
20	0.0094	0	89	0.6786	
25	0.0110	0.0016	90	0.7195	0.7335
30	0.0129	0.0034	91	0.7585	
35	0.0154	0.0060	92	0.7967	
40	0.0188	0.0093	93	0.8306	
45	0.0239	0.0141	94	0.8617	
50	0.0320	0.0221	95	0.8893	0.8962
55	0.0447	0.0348	96	0.9128	
60	0.0658	0.0555	97	0.9333	
65	0.0990	0.0870	98	0.9507	
70	0.1512	0.1364	99	0.9650	0.9734
75	0.2300	0.2161	99.9	0.9753	
77	0.2725		100	1.0000	1.0000
79	0.3220				

calculation reproduces the age distribution of a population in a given year (i.e., 1996), based on the current population distribution and the mortality rate. (See Appendix E for full calculations, and Table 5.2 for MISCAN input). First, the cumulative probability of death from all causes by a certain age was calculated for 1995-1997, using the life-table methods described above. This probability was then divided into the average population for 1995-1997 at that age to give the number of "births" needed at each age. Births were then summed to determine the cumulative probability of being born by each year, and then the cumulative probability of being born by each year from 1896 to 1996 was calculated to generate the observed population. These data are limited for those aged 90 and over by the smaller

Table 5.2 Input data for MISCAN model of cumulative probability of birth expressed as a ratio, by year, Canada, 1896-1996, with comparisons to the Netherlands

Ratio			Ratio			Ratio		
Birth Year	Canada	Nether-lands	Birth Year	Canada	Nether-lands	Birth Year	Canada	Nether-lands
1896	0.000	0	1914	0.096		1932	0.232	
1897	0.014		1915	0.102		1933	0.239	
1898	0.017		1916	0.109	0.110	1934	0.247	
1899	0.021		1917	0.115		1935	0.255	
1900	0.025		1918	0.122		1936	0.263	0.277
1901	0.029		1919	0.128		1941	0.306	0.323
1902	0.033		1920	0.136		1946	0.357	0.370
1903	0.038		1921	0.143	0.150	1951	0.422	0.420
1904	0.042		1922	0.151		1956	0.493	0.485
1905	0.047		1923	0.159		1961	0.572	0.552
1906	0.052	0.045	1924	0.167		1966	0.647	0.622
1907	0.056		1925	0.175		1971	0.711	0.695
1908	0.061		1926	0.183	0.190	1976	0.770	0.770
1909	0.067		1927	0.191		1981	0.828	0.839
1910	0.072		1928	0.199		1986	0.885	0.895
1911	0.078		1929	0.207		1991	0.944	0.947
1912	0.083		1930	0.215		1996	1.000	1.000
1913	0.090		1931	0.223	0.230			

population size, the poorer quality of population estimates by single year of age, and the restriction to just three decimal places in the MISCAN input, where the smallest number is 0.014.

5.3.3 Calculation of Pre-screening Incidence

Canadian rates for *in situ* and invasive breast cancer prior to screening were calculated and then used as input for a back-calculation (performed by R. Boer, personal communication) of cumulative incidence using optimistic and pessimistic assumptions of sojourn time. In the MISCAN model, cumulative incidence is defined as the probability that a person undergoes a transition from the first state (i.e., normal, no disease) to another state

(e.g., *in situ*, T1a, T1b, T1c, T2+). The input value for the optimistic model is 0.1765; and for the pessimistic model, is 0.2537.

***Calculation of rates for in situ breast cancer*:** Counts and rates of *in situ* breast cancer age-standardized to the 1991 Canadian population were computed for Canada and for each province/territory for the years 1980 to 1996. Data for *in situ* breast cancer were found to be reasonably complete for all provinces except Ontario, starting in 1982, when Manitoba and Quebec began reporting such data. (Ontario does not report data for *in situ* cancer.) From 1982 to 1985, rates for *in situ* breast cancer increased from 2.9 to 4.2 per 100,000 for Canada, excluding Ontario and the territories. *In situ* rates also varied by province; for example, for the period 1983-1985, age-standardized rates ranged from 1.5 per 100,000 in Prince Edward Island up to 6.0 per 100,000 in British Columbia. *In situ* cancers as a percentage of all breast cancers also rose from 3.1% to 4.2% between 1982 and 1985, with an average of 3.7%, a value very similar to the 4% of *in situ* cancers reported in the Netherlands prior to their screening program.

Ultimately, data for 1982-1985 for all provinces except Ontario, were chosen to represent the rates prior to screening. The variations in both percentages, and in overall rate of *in situ* cancer, reflect, in part, chance variation due to small numbers, variations in the completeness of registration of *in situ* cancers, and differences in medical practice such as early penetration of screening mammography. To minimize contamination with the National Breast Screening Study that was ongoing from 1981-1985, it is generally considered that mammography screening was not performed in Canada on a large scale prior to 1986. There is anecdotal evidence, however, that in British Columbia and Alberta some screening was started in 1985. This limited early introduction of screening may have contributed to diagnosis of some *in situ* breast cancer, but the impact should be small at the national level.

5.3.4 Assessment of Breast Cancer Stage for Model Evaluation

Canadian stage distributions were compared to those used for MISCAN model input and validation for several female populations, including Nijmegen, Utrecht, the Netherlands and the United Kingdom (Table 5.3). Canadian data did not show any important differences in stage distribution for breast cancers diagnosed pre-screening in women aged 50 to 69. Given the high rate of unknown T1x in Saskatchewan, data for both Northern Alberta and Saskatchewan were recalculated for this table by omitting cases coded as Tx from the calculation and by proportionately reassigning cancers staged as T1x to T1a, T1b and T1c. Tumours with T stage ≤ 10 mm accounted for only about 10%-15% of all tumours, while stage T2+ accounted for 45%-55%. (See Appendix C for details of staging classification.)

Table 5.3 Comparison of pre-screening stage distributions from Canadian to those from other sources used in MISCAN model, age ranges from 50-69 years

	Northern Alberta*	Saskatch-ewan*	Netherlands, Clinical [38]		Utrecht Control [29]
Year	**1971-1984**	**1975-1984**	**1989-90**	**1989 ****	**n.s**
Age Range (y)			**50-69**		
No. of cancers	***2,441***	***1,976***	***4,093***	***n.s.***	***n.s.***
Stage:	**%**	**%**	**%**	**%**	**%**
Tis	2.1	3.2	4.0	1.9	4.6
T1a	1.4	1.2	1.4	1.5	1.5
T1b	11.6	13.9	6.1	7.8	6.3
T1c	27.7	37.5	30.5	35.6	32.6
T2+	57.1	44.2	50.1	53.1	55.0
Tx	-	-	7.9	n.s	n.s
Total	100.0	100.0	100.0	100.0	100.0

Note: n.s. = not specified

* The % stage distribution for Alberta and Saskatchewan is calculated by omitting Tx from the calculation and by re-allocating cancers staged as T1x proportionately to T1a, T1b and T1c.

** Excluding Nijmegen and Utrecht because of pilot projects, email, R. Boer 2000/07/10

Source: Northern Alberta Breast Registry, Saskatchewan Cancer Registry.

Similarly, the post-screening stage distributions at first and re-screen were close to those observed in other screening programs (Table 5.4). Exact comparisons are hampered due to the small numbers for some of the comparison populations, the slightly different age ranges used (i.e., 50-64 vs 50-59), and the changes in the MISCAN model over time from use

Table 5.4 Comparison of post-screening stage distributions for breast cancers detected at first or re-screen from Canadian screening programs, to those from sources used in MISCAN model, age ranges around 50-69 years

	Canada	Utrecht [89]	Netherlands			UK [29 97]
	1988-96		1990-91	1992	Boer[97]	
Age (y)	50-69	50-64	---------- 50-69 ----------			50-64
Interval	2y	1-4y	---------- 2y ----------			3y
	---------- First Screen ----------					
No. of cancers	*3,847*	*108*	*1,068*	*1,447*	*1,754*	*392*
Stage:†	%	%	%	%	%	%
Tis	17.4	14.0	13.0	13.9	14.7	15.1
T1a*	6.4	25†	3.6	4.3	6.3	9.2*
T1b	21.1		22.0	19.1	21.1	30.6
T1c	33.0	40.0	38.9	34.5	37.9	30.4
T1x	0.2	n.s	n.s	n.s	n.s	n.s
T2+	16.9	21.0	16.9	19.8	20.0	14.8
Tx	4.9	n.s	5.6	8.4	n.s	n.s
	---------- Re-screen ----------					
No. of cancers	*2,103*	*81*			*449*	*92*
Tis	17.4	6.0			14.4	20.7
T1a*	8.2	32†			5.6	11.9*
T1b	24.2				23.3	27.2
T1c	33.0	46.0			37.8	28.3
T1x	0.2	n.s.			n.s	n.s
T2+	12.2	16.0			18.9	12.0
Tx	4.9	n.s.			n.s.	n.s.

Note: n.s. = not specified

† Stage represents 0-9, 10-19, and 20+mm

* Includes microinvasive for UK data

Source: Canadian Breast Cancer Screening Database; published results for MISCAN model.

of four stage groups (Tis, <10, 11-19, and 20+ mm) to five stage groups with the latter based on the TNM classification, where the upper cutoffs are 5, 10, and 20 mm. As commonly occurs due to digit preference, many tumours are sized at exactly 5, 10, 15, 20 mm and so on. Thus, changing the cut-off can have considerable impact on the percent distribution.

5.3.5 Screening Participation in Canada

Screening participation is modelled in MISCAN using three probabilities: the probability that a woman will attend the first screen; the probability that she will return for the next screen having attended the previous screen; and the probability that she will attend the next screen having not attended the previous screen. Canadian data from administrative files were used to develop plausible screening scenarios as shown in Table 5.5.[31 98] In early models, screening was assumed to occur every 2 years, at ages 51.0 to 79.0 exactly at the beginning of each new year of age – the lower age was modified to 41.0 as models evolved. The starting age was selected to reflect the reality that not all women will start screening at exactly 50 years of age. Screening was continued over a 30-year period.

In the full scenario, 100% participation is assumed at all biennial screens. The target scenario assumes 70% of women in each age group will be screened and that 80% will return for the next screen, having attended the previous screen. Finally, 50% of women who do not attend the previous screen are assumed to attend the next biennial screen. The 80% return rate is based approximately on that observed in the CBCSD for women aged 50 to 79, albeit up to 30 months post-screen, while the 50% uptake rate for lapsed or never screened women is an estimate. Participation at initial screens for the next two scenarios use observed mammography utilization rates for 1990 and 1996 derived from a combination of CBCSD and fee-for-service data. Return rates are adjusted downward for the 40-49 age group in line with results for that age group reported for the CBCSD. Each scenario is varied by ending screening at age 69.

Table 5.5 Summary of screening participation probabilities for attending biennial screening

Screening Scenario	Age Group				
	40-49	50-59	60-69	70-79	
				Screened	No Screen
Full (100% participation)					
1st screen	1.00	1.00	1.00	1.00	0.00
Next screen, previous screen attended	1.00	1.00	1.00	1.00	0.00
Next screen, previous not attended	1.00	1.00	1.00	1.00	0.00
Target					
1st screen	0.70	0.70	0.70	0.70	0.00
Next screen, previous screen attended	0.80	0.80	0.80	0.80	0.00
Next screen, previous not attended	0.50	0.50	0.50	0.50	0.00
1996 levels					
1st screen	0.39	0.72	0.60	0.39	0.00
Next screen, previous screen attended	0.65	0.80	0.80	0.82	0.00
Next screen, previous not attended	0.26	0.46	0.35	0.24	0.00
1990 levels					
1st screen	0.35	0.43	0.34	0.20	0.00
Next screen, previous screen attended	0.65	0.80	0.80	0.82	0.00
Next screen, previous not attended	0.24	0.33	0.25	0.14	0.00
No screening					
1st screen	0.00	0.00	0.00	0.00	0.00
Next screen, previous screen attended	0.00	0.00	0.00	0.00	0.00
Next screen, previous not attended	0.00	0.00	0.00	0.00	0.00

5.4 Model Validation

5.4.1 Population Distribution

The number of women simulated in MISCAN was varied until both the overall population counts and the distribution of the Canadian population by age could be reasonably well reproduced. Model runs were based on over 300 million simulations to improve precision of results and then rescaled to approximate the actual population. The number of life-years produced in each five-year age group for 1996 was tabulated and compared to the average Canadian population from 1995-1997 (Table 5.6).

Table 5.6 Validation of Canadian cumulative birth input data used in MISCAN

Age Range	Canadian Population 1995-1997 average	MISCAN life-years 1996 estimate	Ratio of Life-years to Canadian pop'n*10
00-04	955,973	8,621,076	0.90
05-09	984,900	9,973,471	1.01
10-14	977,609	9,697,202	0.99
15-19	976,011	9,793,660	1.00
20-24	989,348	9,992,600	1.01
25-29	1,071,655	10,730,221	1.00
30-34	1,274,427	12,428,390	0.98
35-39	1,306,067	13,213,250	1.01
40-44	1,192,783	11,990,477	1.01
45-49	1,058,883	10,886,990	1.03
50-54	838,412	8,611,067	1.03
55-59	673,703	7,280,031	1.08
60-64	614,485	6,148,550	1.00
65-69	590,574	5,967,172	1.01
70-74	539,632	5,386,566	1.00
75-79	408,452	4,150,046	1.02
80-84	287,145	2,955,340	1.03
85-89	157,995	1,660,778	1.05
90-94	64,806	700,058	1.08
95+	19,570	178,217	0.91
Total (0-100+)	**14,982,430**	**150,365,161**	**1.00**
80+	**529,516**	**5,494,393**	**1.04**
100+	**2,672**		

Note: Scenario modelled was with full participation from 50-79 with screening from 1996 to 2023

Source: Based on final run December 17, 2000, using file all_rev80_33.in

Model estimates for the total population and those for most age groups agree well with actual data. The main deviations occurred among the oldest and the very youngest age groups.

The discrepancy for the youngest age group is not surprising as no special adjustments were made to the life-table as would normally be done to account for most deaths occurring in the first year of life. However, as this age group is too young to affect model results, further adjustment was deemed unnecessary. MISCAN tends to over-estimate

the 45-59 age group by about 3%-8%. The reasons for this are not known, but may reflect rapid changes in birth rates between 1937 and 1951. The deviation in the oldest age groups may reflect limitations in the population and death data, particularly for women aged 90 and over. The small proportion of women in these single years of age limits the precision of the MISCAN input, for both birth and life-table data. Attempts to smooth out the estimates for the 80+ age groups, however, did not achieve a closer fit.

5.4.2 Validation of Stage Distribution

A key test of the ability of MISCAN to predict breast cancer in a population is the comparability of the stage distributions produced by the model to those actually recorded in a population. Stage distributions from Alberta and Saskatchewan prior to 1984 were compared to modelled output for clinically detected breast cancer for a population of women aged 50 to 79 undergoing biennial screening (Table 5.7). Model estimates by stage agreed reasonably well with the stage distribution adjusted for unknown stage. The exception was stage T1b, where the model output was lower, although this may reflect the difficulties of adjusting for unknown stage. The stage distribution of screen-detected breast cancer was also compared between the MISCAN model and the Canadian Breast Cancer Screening Database. Again the modelled output was within the range of stage distributions for the CBCSD, with the exception of T1b, where the modelled output was slightly higher. It was concluded that MISCAN model results reasonably reflected the Canadian population and that further adjustments to model parameters were not needed.

5.4.3 Validation of Screening Volumes

Model assumptions shown in Table 5.8 were used in subsequent analyses unless otherwise stated. Sojourn times may be optimistic (shorter) or pessimistic (longer) (see Section 5.1), and four screening participation levels are modelled for the optimistic assumption of sojourn time. In all scenarios, the age range screened is 40 to 79 and the start

Table 5.7 Comparison of stage distribution in Canada of MISCAN estimates with pre- and post-screening actual data for women aged 50-79

	MISCAN estimates				50-69 pre-1984, adjusted	
	Screened		No screening		AB	SK
Clinically Detected Breast Cancer						
	#	%	#	%	%	%
DCIS	40,400	3.4	49,161	3.4	2.1	3.2
T1a	16,449	1.4	16,901	1.2	1.4	1.2
T1b	73,575	6.2	81,314	5.6	11.6	13.9
T1c	375,994	31.8	457,409	31.4	27.7	37.5
T2+	674,921	57.1	851,972	58.5	57.1	44.2
Total	1,181,339	100.0	1,456,757	100.0		
					CBCSD, Rescreen	
Screen-detected Breast Cancer					**50-69**	**70-79**
DCIS	38,108	12.6			17.4	13.9
T1a	19,170	6.4			8.2	7.8
T1b	99,741	33.0			24.2	30.1
T1c	114,323	37.9			33.0	34.7
T2+	30,535	10.1			12.2	9.6
Total	301,877	100.0				

Note: Estimates were obtained from a model assuming full (100%) participation and optimistic sojourn time for ages 50-79. Biennial screening starts in 1996 and continues to 2026.

Source: Output file: CdnIMB_opt_all_initial.pp.xls. Model results were divided 10.

year and end year for screening are 1988 and 2017. (These parameters were revised from the models described in Sections 5.4.1 and 5.4.2 to permit screening histories to be built up from 1988, and to improve the ability to validate counts of incident cancers and deaths – see Section 5.4.4.) The model assumes that all women are invited, thereby giving a population count in the first two years, 1988 and 1989. A larger simulated population must be modelled for the pessimistic model and the counts produced vary slightly from optimistic models. The larger population is needed because records in which cancer is simulated are weighted 50 to 1 to records without cancer, and more cancer records are simulated in the pessimistic model due to longer sojourn times.

Table 5.8 Model assumptions and validation of MISCAN estimates for number of screens by screening participation level, Canada

Model assumptions:						
Canadian data for:		Incidence, Cumulative Births and Mortality				
Population simulated:		33 million	33 million	33 million	33 million	45.69 million
Sojourn time:		optimistic	optimistic	optimistic	optimistic	pessimistic
Screening Policy:	Age Range	40-79	40-79	40-79	40-79	40-79
	Start year	1988	1988	1988	1988	1988
	End year	2017	2017	2017	2017	2017
	Participation	1990 level	1996 level	Target ~70%	Full - 100%	Full - 100%
Screening volumes						
Screened women at start				***MISCAN Estimates***		
First screens 1988		857,781	1,302,668	1,727,718	2,468,996	2,466,156
1989		842,266	1,249,126	1,681,103	2,374,385	2,376,540
Number of women screened		1,700,046	2,551,794	3,408,821	4,843,380	4,842,697
Number of women invited		4,843,380	4,843,380	4,843,380	4,843,380	4,842,697
% screened		*35.1*	*52.7*	*70.4*	*100.0*	*100.0*
Total screens	**Actual**					
1994	1,400,000	1,331,204	1,567,203	2,025,139	2,839,656	2,835,008
1996	1,489,355	1,412,473	1,631,484	2,105,053	2,948,323	2,942,330
				Ratio of MISCAN estimates to actual screens		
1994		0.95	1.12	1.45	2.03	2.03
1996		0.95	1.10	1.41	1.98	1.98
Total screening activity from 1988 to 2017				***MISCAN estimates***		
First screens		9,465,304	10,084,267	11,002,900	11,385,574	11,386,807
Repeat screens		37,767,068	44,796,687	58,596,316	86,278,548	86,261,522
% receiving first screen		*83.1*	*88.6*	*96.6*	*100.0*	*100.0*
% receiving repeat screens		*43.8*	*51.9*	*67.9*	*100.0*	*100.0*

The number of women receiving first screens increases from 1.7 million in the 1990 level model to 4.8 million in the full model, with the percentage of women initially screened being 35% for the 1990 level, 52% for the 1996 level and 70% for the target. MISCAN estimates were compared to total 'actual' screens in 1994 and 1996: the model estimate for

the number of screens was 5% lower than actual, based on 1990 levels and 10-12 % higher than actual based on 1996 levels. The proportion of women receiving a first screen at some point over the 30-year screening period, ranged from 83% in the 1990 level model to 100% in the full model, while the proportion of repeat screens was 44% in the 1990 level scenario, 53% in the 1996 level and 68% in the target scenario. These results indicate that the models of screening participation cover a reasonably realistic range based on Canadian patterns.

5.4.4 Validation of MISCAN Output for Breast Cancer Incidence and Deaths

MISCAN estimates for breast cancer deaths and incidence were compared to actual data for deaths for 1997, and incidence data for 1996 (the most recent years for which data were available), for the full range of screening activities (Table 5.9). The estimated number of deaths varied by age and screening behaviour in a predictable manner, with fewer deaths occurring at ages 40 and over, as screening participation increased.

The model based on the 1990 level of screening most accurately predicted the overall 1997 mortality from breast cancer, and gave the closest estimates for the 70+ age groups. This might be expected as mortality reduction generally appears about seven years after screening is initiated. MISCAN predicts more deaths in women under 40 years of age than are actually observed, which could indicate that factors other than screening (e.g., treatment) are affecting the current mortality levels for this age group, or that further work may be needed to ensure that mortality rates used accurately reflect reality. MISCAN also predicts progressively fewer deaths in the 80+ age group as screening participation increases, reflecting the delayed impact of screening women younger than 80. Interestingly, the pessimistic model showed a higher percentage reduction in deaths among women aged 70 and over. This likely reflects the lessened lethality of the slower-growing tumours simulated with the longer sojourn times in this model. Screening is not necessarily responsible for all of the mortality reduction, as treatment advances may also have contributed.

Table 5.9 Comparison of MISCAN estimates to actual breast cancer deaths in 1996 by age group and screening participation level, based on biennial screening of women aged 40-79 for 30 years (1988 to 2017), Canada

Screening participation level:		1990 level	1996 level	Target	Full -opt	Full-pess
Age at Death:	**Actual deaths in 1997**	*MISCAN estimates, breast cancer deaths, 1997*				
<40	164	217	217	217	217	220
40-49	566	614	610	576	545	549
50-69	1,816	1,976	1,880	1,809	1,624	1,610
70-79	1,238	1,179	1,090	1,031	932	910
80-99+	1,161	1,053	1,010	950	899	819
All	4,945	5,039	4,806	4,582	4,217	4,108
% reduction in deaths						
all ages		***1.9***	***2.8***	***7.3***	***14.7***	***16.9***
ages 70+		***6.9***	***12.5***	***17.4***	***23.7***	***27.9***
		Ratio of MISCAN estimates to actual deaths, 1997				
<40		1.32	1.32	1.32	1.32	1.34
40-49		1.08	1.08	1.02	0.96	0.97
50-69		1.09	1.04	1.00	0.89	0.89
70-79		0.95	0.88	0.83	0.75	0.74
80-99+		0.91	0.87	0.82	0.77	0.71
All		1.02	0.97	0.93	0.85	0.83
Age at Diagnosis	**Actual new cancers, 1996**	*MISCAN estimates, new invasive cancers, 1996*				
<40	1,022	1,069	1,069	1,069	1,069	1,066
40-49	3,001	3,069	3,122	3,188	3,184	3,268
50-69	7,376	6,790	6,761	6,723	6,798	6,931
70-79	3,524	3,330	3,318	3,341	3,329	4,232
80-99+	1,860	1,936	1,836	1,719	1,624	1,342
All	16,783	16,193	16,105	16,040	16,004	16,837
		Ratio MISCAN estimates to actual new cancers, 1996				
<40		1.05	1.05	1.05	1.05	1.04
40-49		1.02	1.04	1.06	1.06	1.09
50-69		0.92	0.92	0.91	0.92	0.94
70-79		0.94	0.94	0.95	0.94	1.20
80-99+		1.04	0.99	0.92	0.87	0.72
All		0.96	0.96	0.96	0.95	1.00

Estimated total incidence in 1996 among the optimistic models varied only slightly and was 4%-5% lower than the actual new cases. These results likely occurred because most of the prevalent cancers are detected by MISCAN in the first few years of screening, and by 1996, eight years later, any variability of excess incidence due to differing screening participation levels had already occurred. Ratios of model estimates to actual counts by age group were fairly similar across optimistic models, except for the 80+ age group, where the model estimates increased as screening participation decreased.

5.5 Results: Benefits and Harms of Screening Women Aged 70 to 79

5.5.1 Benefits and Harms for Extra Screens, Extra Incidence, Deaths Prevented, Life-years Gained and Life-years in Lead-Time

Model assumptions were implemented as described in Table 5.8, with screening participation varied in separate model runs for women. The impact of screening women aged 70 to 79 was then calculated as the difference between screening women aged 40 to 79 and 40 to 69 (Table 5.10). As a result of screening over a 30-year period, an estimated 36,000 and 84,000 fewer cancers will be detected clinically, while an additional 42,000 to 118,000 cancers will be detected by screening. The extra incidence will vary from 6,300 for the 1990 level of screening up to 12,600 for full participation, using optimistic assumptions of sojourn time, and up to 35,000 for pessimistic assumptions. These represent 210 to 1,168 additional cancers per year of screening. These extra incident cancers occur over the entire screening period and represent cancers that would never have been detected in the absence of screening. The initial increase in incidence due to detection of prevalent cases is largely balanced by a decline in incidence when screening stops. To provide some perspective, the net additional cancers represent about 4% to 22% of the 5,384 cases diagnosed in women aged 70 and over in 1996. As shown in Table 5.10, the net effect is that most cancers will be

Table 5.10 Benefits and harms of screening Canadian women aged 70-79 biennially for 30 years (1988-2017) by screening participation level, as measured by extra screens, extra incidence, deaths prevented, life-years gained and life-years in lead-time

Screening Participation	1990 level	1996 level	Target	Full	Full
Sojourn Time	---------- Optimistic ----------				Pessimistic
	Difference in counts for screening 40-79 and 40-69				
Extra screens	6,260,170	8,077,705	10,097,655	14,141,948	14,082,738
Extra incidence					
Incident breast cancers					
clinically detected	-35,850	-46,374	-58,674	-71,858	-83,594
screen-detected	42,151	54,548	68,999	84,447	118,624
Total extra incidence	**6,301**	**8,174**	**10,325**	**12,590**	**35,030**
per year of screening	*210*	*272*	*344*	*420*	*1,168*
Extra incidence by stage					
DCIS	2,856	3,661	4,591	5,720	6,994
T1a	2,620	3,399	4,258	5,909	12,670
T1b	13,239	17181	21,600	29,357	47,859
T1c	6,751	8776	11,297	11,421	11,607
T2+	-19,166	-24,842	-31,421	-39,816	-44,100
Breast cancer deaths prevented					
from screening over 30 years	8,432	10,915	13,774	17,234	19,327
per year of screening	*281*	*364*	*459*	*574*	*644*
Life-years gained	69,571	89,620	112,903	140,954	151,362
Life-years in lead-time	150,027	193,779	244,489	298,535	625,290

Note: 1990 and 1996 levels of screening participation are estimated based on both CBCSD and fee-for-service data (*see* Table 5.5)

detected at an early stage, with far fewer breast cancers occurring at stage T2+, and substantially more cancers diagnosed at stage T1b, when they are 10 mm or less in size.

As a result of screening women aged 70 to 79 for 30 years, fewer breast cancer deaths are predicted to occur, ranging from 8,400 if screening participation is at 1990 levels, 17,200 if full participation is reached and up to 19,300 if pessimistic sojourn times apply. These range from 281 to 644 breast cancer deaths prevented per year of screening, which represent

a considerable proportion of the 2,400 deaths that actually occurred in the 70+ age group in 1997. More deaths are prevented if pessimistic sojourn times are assumed because the slower cancer growth rates mean it would take longer to die from the cancer, and also that cancers are likely to be detected at an earlier stage, when they are less lethal. (Note the considerably increased number of cancers diagnosed at stages DCIS, T1a and T1b for the pessimistic model.)

Estimated life-years gained as a result of screening women aged 70 to 79 over 30 years ranged from 69,600 at the 1990 level up to 141,000 for full participation, with a small rise to 151,400 for full participation with a pessimistic sojourn time. Life-years in lead-time increases from an estimated 150,000 up to 298,500 for optimistic assumptions, but then rose dramatically to 625,300 for full participation with pessimistic assumptions for sojourn time.

5.5.2 Derived Outcome Measures

Data presented in Table 5.10 were manipulated to derive the outcome measures presented in Table 5.11. As examples, the *number of screens needed to prevent one breast cancer death* is calculated by dividing the number of extra screens (i.e., 6,260,170 for the 1990 level) by the corresponding number of deaths prevented (8,432), to obtain 742. Similarly for the 1990 level of screening, *life expectancy gains in days per screen* is calculated by first multiplying the number of life-years gained by 365 to get the number of life-days gained (i.e., 69,571*365) and then dividing by the number of extra screens (6,260,170) to obtain 4.06 days. Most indicators remained rather constant across various screening scenarios using optimistic assumptions, while some varied for pessimistic sojourn times. The ratio of life-years in lead-time to life-years gained was about 2.1 for optimistic, but increased to 4.1 for pessimistic, sojourn times for women aged 70 to 79 who participated in screening over a 30-year time period. The number of screens needed to prevent one breast cancer death generally varied between 730 to 740. The higher number of 821 found for full

Table 5.11 Derived indicators of benefit and harm of screening women aged 70-79 biennially for 30 years (1988-2017), by screening participation level, Canada

Screening Participation:	1990 level	1996 level	Target	Full	Full
Sojourn Time:	----------optimistic----------				pessimistic
	Based on comparison of age 40-69 and 40-79				
Number of screens to prevent one breast cancer death	742	740	733	821	729
Number of screens to gain one life-year	90	90	89	100	93
Life expectancy gains in days per screen	4.06	4.05	4.08	3.64	3.92
Number of life-years gained per breast cancer death prevented	8.3	8.2	8.2	8.2	7.8
Ratio of life-years in lead-time to life-years gained	2.16	2.16	2.17	2.12	4.13
Number of life-years in lead-time per breast cancer death prevented	17.8	17.8	17.7	17.3	32.4
Ratio of breast cancer deaths prevented per extra breast cancer diagnosed	1.34	1.34	1.33	1.37	0.55

Source: These data are derived from results presented in Table 5.10

participation may reflect diminishing returns. Similarly, about 90 screens were needed to gain one life-year, and life-expectancy is estimated to increase by about 4 days for each screen. However, for each breast cancer death prevented, 8.2 life-years are gained under optimistic assumptions, and 7.8 life-years under pessimistic assumptions. Finally, about 17 to 18 years were lived in lead-time under optimistic assumptions, and 32 years under pessimistic assumptions, for each breast cancer death prevented.

5.5.3 Calculation of Extra Follow-up Tests

Rates calculated from screening reported to the CBCSD are applied to screening volumes produced by MISCAN to assess the volume of diagnostic follow-up procedures that would ensue if biennial screening is extended to women aged 70 to 79 over a 30-year period

(Table 5.12). In 1988 and 1989, 735,600 additional first screens would have occurred if women had fully participated in screening; this number would drop to 541,700 if the 1990

Table 5.12 Estimates of diagnostic follow-up procedures ensuing from abnormal mammograms of screening women aged 70-79 biennially for 30 years from 1988 to 2017 by screening participation level, Canada

Screening Participation:		1990 level	1996 level	Target	Full	Full
Sojourn Time:		---------- optimistic ----------				pessimistic
Estimated additional screens (MISCAN)	**CBCSD Rate**	**Difference in MISCAN estimates for screening ages 40-79 vs 40-69**				
1st screen		541,669	620,764	777,435	735,554	735,773
repeat screen		5,718,501	7,456,941	9,320,220	13,406,394	13,346,965
All extra screens		6,260,170	8,077,705	10,097,655	14,141,948	14,082,738
Abnormal recall		**Estimated abnormal mammograms, CBCSD recall rates**				
first screens	11.3%	61,209	70,146	87,850	83,118	83,142
repeat screens	5.9%	337,392	439,960	549,893	790,977	787,471
All screens		398,600	510,106	637,743	874,095	870,613
Diagnostic procedures		**Estimated diagnostic procedures due to abnormal mammograms, CBCSD procedure rates**				
diagnostic mammogram	71.2%	283,803	363,195	454,073	622,356	619,877
ultrasound	38.8%	154,657	197,921	247,444	339,149	337,798
fine needle aspiration	5.1%	20,329	26,015	32,525	44,579	44,401
core biopsy	5.1%	20,329	26,015	32,525	44,579	44,401
open biopsy	17.8%	70,951	90,799	113,518	155,589	154,969
total biopsies	22.9%	91,279	116,814	146,043	200,168	199,370
		Estimated procedures per year from abnormal mammograms				
diagnostic mammogram		9,460	12,107	15,136	20,745	20,663
ultrasound		5,155	6,597	8,248	11,305	11,260
fine needle aspiration		678	867	1,084	1,486	1,480
core biopsy		678	867	1,084	1,486	1,480
open biopsy		2,365	3,027	3,784	5,186	5,166
total biopsies		3,043	3,894	4,868	6,672	6,646

Note: Estimated diagnostic procedures are calculated by multiplying the number of total extra screens by the CBCSD percentage of each procedure

level of screening had been in effect. Overall, 14 million more screens would take place if all women aged 70 to 79 had participated fully over the 30-year period (1988-2017). The 6 to 14 million extra mammograms performed, depending upon the screening participation rate, will result in an estimated 400,000 to 875,000 abnormal mammograms based on 1997 and 1998 data from the CBCSD.[98] The ensuing follow-up diagnostic procedures to determine if cancer is present will amount to about 15,000 diagnostic mammograms, 8,000 ultrasounds, and 5,000 biopsies each year, if 70% of women are screened. Ideally, these figures should be adjusted to exclude those procedures that would be necessary to assess the cancers that would be clinically detected in the absence of screening. However, this adjustment cannot be done due to the lack of information on the rates both of women undergoing diagnostic mammography in the general population, and of the resultant follow-up procedures and biopsies. (It is impossible to accurately count the number of diagnostic mammograms, first, because many are done outside the fee-for-service system as part of block funding to hospitals, and second because provincial billing codes often do not effectively distinguish between diagnostic and screening mammography services. Further, no easy or reliable method exists to tabulate the number of follow-up diagnostic procedures.)

At the target screening participation level, these tests will results in an estimate of 74,300 screen-detected cancers, based on applying CBCSD rates to MISCAN estimates of the number of screens, a figure within 8% of the 69,000 screen-detected cancers estimated directly by MISCAN (Table 5.13) However, as estimated by MISCAN, there will be fewer clinically detected cancers, so the actual percentage of extra cancers will be about 15%. (Table 5.14). These figures can be used to make a very approximate adjustment for biopsies by multiplying the number of biopsies by the ratio of the total extra cancers vs extra screen-detected cancers, both as estimated by MISCAN. This gives an estimate of 728 extra

Table 5.13 MISCAN estimates of screen-detected cancers compared to cancers estimated from CBCSD detection rates, biennial screening of women aged 40-79 for 30 years (1988-2017), by screening participation level, Canada

Screening Participation:		1990 level	1996 level	Target	Full	Full
Sojourn Time:		---------- Optimistic ----------				pessimistic
CBCSD cancer detection rate per 1,000 screens		Estimated screen-detected cancers due to abnormal mammograms				
first screen	12.9	6,988	8,008	10,029	9,489	9,491
repeat screen	6.9	39,458	51,453	64,310	92,504	92,094
all CBCSD screens		46,445	59,461	74,338	101,993	101,586
screen-detected cancers per year		*1,548*	*1,982*	*2,478*	*3,400*	*3,386*
MISCAN Estimates:						
total screen-detected cancers		42,151	54,548	68,999	84,447	118,624
screen-detected cancers per year		*1,405*	*1,818*	*2,300*	*2,815*	*3,954*
Ratio of CBCSD to MISCAN		**1.10**	**1.09**	**1.08**	**1.21**	**0.86**

Note: CBCSD rates include cancers detected by CBE, and thus are slightly higher than if mammography only screens were used.

Table 5.14 Harms of screening women aged 70-79 biennially for breast cancer for 30 years (1988-2017), including extra biopsies and false positive rates, Canada

Screening Participation:	1990 level	1996 level	Target	Full	Full
Sojourn Time:	---------- Optimistic ----------				pessimistic
Extra biopsies					
MISCAN estimates					
total screen-detected cancers	42,151	54,548	68,999	84,447	118,624
total extra cancers (MISCAN)	6,301	8,174	10,325	12,590	35,030
Ratio of extra to screen-detected cancers	**0.15**	**0.15**	**0.15**	**0.15**	**0.30**
After adjusting for the ratio of extra to screen-detected cancers					
Total "extra" biopsies per year	*455*	*584*	*728*	*995*	*1962*
False positives					
abnormal mammograms	398,600	510,106	637,743	874,095	870,613
screen-detected cancers	42,151	54,548	68,999	84,447	118,624
False positives	356,449	455,558	568,744	789,648	751,989
per year	*11,882*	*15,185*	*18,958*	*26,322*	*25,066*

biopsies per year for the target level of screening, increasing to 1,962 for full screening with pessimistic sojourn times.

Potential benefits and harms are summarized in Table 5.15. Per 10,000 screens there will be between 530 and 560 women with false positive results. The 21 to 42 extra biopsies will result in about 10 to 25 extra breast cancers per 10,000 screens, while averting about 13 to 14 deaths. Overall, just over 100 life-years will be saved, while 211 to 444 life-years will be lived in lead-time, among those women diagnosed with screen-detected cancers. For the target screening participation, the extra cancers reflect that 68 cancers will be detected per 10,000 screens, while 58 fewer cancers will be detected clinically. More cancers will be found at stage T1b (21 per 10,000) and T1c (11 per 10,000) and many fewer (31 per 10,000) will be found at stage T2+, compared to no screening.

5.5.4 Quality-adjusted Life-years (QALYs)

Quality-adjusted life-years were calculated based on the full participation in biennial mammography to age 79, using both optimistic and pessimistic assumptions of sojourn time. MISCAN calculates, by initial stage of breast cancer at diagnosis, the life-years in initial therapy (estimated at one year for all stages of breast cancer). Life-years in terminal care for breast cancer for all stages combined are produced, assuming that 1.768 years will be spent in terminal care prior to each breast cancer death. Life-years in terminal care for other causes are also calculated, assuming on average one year. Disease-free years in continuing care are calculated as total life-years with breast cancer minus the total of life-years in initial therapy and life-years in terminal care due to breast cancer or other causes. States are assigned hierarchically, with the terminal care being assigned first, then initial therapy, and finally continuing care. MISCAN estimates of the number of screens were used directly and also adjusted (Table 5.12) using the proportion of abnormal mammograms to determine the number of women requiring diagnostic follow-up. The number of screens was multiplied by

Table 5.15 Summary of benefits and harms per 10,000 screens for women aged 70-79 screened biennially for 30 years (1988-2017), by screening participation level, Canada

Screening Participation:	1990 level	1996 level	Target	Full	Full
Sojourn Time:	---------- Optimistic ----------				pessimistic
Estimated procedures:	*per 10,000 screens*				
abnormal mammogram	637	631	632	618	618
diagnostic mammogram	453	450	450	440	440
ultrasound	247	245	245	240	240
fine needle aspiration	32	32	32	32	32
core biopsy	32	32	32	32	32
open biopsy	113	112	112	110	110
total biopsies	146	145	145	142	142
'Extra' biopsies after adjustment	21.8	21.7	21.6	21.1	41.8
False positive tests	569	564	563	558	534
Screen-detected breast cancers					
CBCSD	74.2	73.6	73.6	72.1	72.1
MISCAN	67.3	67.5	68.3	59.7	84.2
extra breast cancers	10.1	10.1	10.2	8.9	24.9
breast cancer deaths prevented	13.5	13.5	13.6	12.2	13.7
life-years saved	111	111	112	100	107
life-years in lead-time	240	240	242	211	444
Incident breast cancers					
clinically detected	-57	-57	-58	-51	-59
screen detected	67	67	68	60	84
by stage:					
DCIS	4.6	4.5	4.5	4	5
T1a	4.2	4.2	4.2	4.2	9
T1b	21.1	21.3	21.4	20.8	34
T1c	10.8	10.9	11.2	8.1	8.2
T2+	-30.6	-30.8	-31.1	-28.2	-31.3

Source: Derived from data presented in Tables 5.11 and 5.13

7/365 to derive the number of life-years in the screening phase (based on one week in this phase), while the number of women requiring diagnostic follow-up was multiplied by 26/365 to obtain the number of life-years in the diagnostic phase (based on the median of 3.7 weeks for diagnostic follow-up recently reported for Canadian women).[99] Utilities (U) reported in the literature were applied to breast cancer-related life-years (LY) to adjust the life-years for each state (x), as (1-U(x))*LY(x) (Table 5.16). As no utility was available for terminal care for other causes, an estimate of 0.8 was used in the basic model. Results were discounted by 0, 3 and 5% to account for time preference, as women may prefer to have time in good health now rather than additional years of life in the future.

Life-years in initial therapy are only slightly less than those for total cancers detected as some patients will die of breast cancer (or other causes) in the first year or two after breast cancer is detected, and the time spent in terminal care is applied first. Assuming full screening participation and optimistic sojourn times, an estimated increase of 20,400 life-years in initial therapy and 414,000 life-years in continuing care was calculated if screening is continued to age 79, vs. stopping screening at aged 69, while 25,400 fewer life-years will be spent in terminal care for breast cancer, and 17,200 more years in terminal care for other causes of death. With pessimistic sojourn times, the life-years spent in initial therapy and continuing care approximately doubled due to detection of additional slower-growing cancers with good prognosis, while life-years spent in terminal care for other causes increased slightly, and for breast cancer decreased slightly (just over 10% in both cases).

Overall, when quality adjustment is used and assuming 0% discounting, the screening program is predicted to be about 18% less effective under optimistic assumption and 29% less effective using pessimistic assumptions. When time preference is considered, the effectiveness decreases by 25% for optimistic assumptions with a 3% discount factor and 31% with a 5% discount factor, and using pessimistic assumptions, by 39% and 48%

Table 5.16 Calculation of change in Quality-Adjusted Life-Years of continuing to screen women aged 70-79, assuming full (100%) biennial screening for 30 years (1988-2017), for optimistic and pessimistic sojourn times and various discount factors, Canada

Discount Factor:		0%		3%		5%	
	Utility	Unadjusted Life-Years	QALYs Lost	Unadjusted Life-Years	QALYs Lost	Unadjusted Life-Years	QALYs Lost
Optimistic Sojourn:							
Screens	0.994	271,215	1,627	177,420	1,065	139,328	836
Diagnostic follow-up	0.895	62,264	6,538	41,668	4,375	33,294	3,496
Initial therapy	0.750	20,363	5,091	16,863	4,216	14,937	3,732
Continuing care	0.955	414,496	18,652	230,137	10,356	164,922	7,422
Terminal care, other causes	0.800	17,234	3,447	7,407	1,481	4,506	901
Terminal care, breast cancer	0.624	-25,420	-9,558	-13,635	-5,127	-9,504	-3,573
Total QALYs lost			25,797		16,863		12,816
Life-years gained		140,954		65,580		41,902	
Net QALYs gained			115,157		49,213		29,087
Ratio: QALY : life-years gained			**0.82**		**0.75**		**0.69**
Pessimistic Sojourn:							
Screens	0.994	270,080	1,620	176,671	1,060	138,742	832
Diagnostic follow-up	0.895	62,016	6,512	41,505	4,358	33,166	3,482
Initial therapy	0.750	43,672	10,918	33,837	8,459	29,074	7,269
Continuing care	0.955	706,864	31,809	391,806	17,631	280,252	12,611
Terminal care, other causes	0.800	19,327	3,865	8,235	1,647	4,981	996
Terminal care, breast cancer	0.624	-28,437	-10,692	-14,987	-5,635	-10,340	-3,888
Total QALYs lost			43,032		27,520		21,303
Life-years gained		151,362		69,919		44,486	
Net QALYs gained			107,329		42,399		23,182
Ratio: QALY : life-years gained			**0.71**		**0.61**		**0.52**

Note: Unadjusted life-years are the life-years spent in each state. Life-years for screening and diagnostic follow-up were calculated by multiplying the number of events by 7/365 (for one week for screening) and 26/365 (for the median 3.7 weeks spent in diagnostic follow-up). QALYs lost were calculated by applying 1-utility to the life-years. The QALYs lost in each state are summed, and the total is subtracted from the life-years gained to obtain the quality-adjusted life-years.

Table 5.17 Sensitivity analysis of Quality-Adjusted Life-Years of continuing to screen women aged 70-79 using alternative utilities for initial therapy and terminal care, other causes

Sojourn Time:		Optimistic			Pessimistic		
Discount Factor:		0%	3%	5%	0%	3%	5%
Utilities		Ratio of QALY : life-years gained					
Base model	***Table 5.16***	***0.82***	***0.75***	***0.69***	***0.71***	***0.61***	***0.52***
Sensitivity analysis							
Terminal care	0.600	0.79	0.73	0.67	0.68	0.58	0.50
Terminal care	0.900	0.83	0.76	0.70	0.72	0.62	0.53
Initial therapy	0.900	0.84	0.79	0.75	0.75	0.68	0.62
Initial therapy & terminal care	0.900	0.85	0.80	0.76	0.77	0.69	0.63

respectively. These calculations are largely driven by the time spent in initial therapy and in the diagnostic follow-up, as terminal care due to future deaths has a much lower impact as discount rates increase. In both optimistic and pessimistic models, the gain in life-years for terminal care due to fewer breast cancer deaths considerably outweighed the life-years lost due to other causes of death, and the gap widened as the discount rate increased.

Sensitivity analyses were conducted by varying the utilities for the year in terminal care in the range of 0.6 to 0.9 (Table 5.17). This had little effect on the model results, as the ratio of QALY to life-years gained differed by at most 2-3%. An alternative utility of 0.9 for initial therapy was also tested to take into consideration that 90% of breast cancers diagnosed in women aged 70 and over will be estrogen receptor positive and thus most women will receive the relatively benign tamoxifen in place of more debilitating chemotherapy. The ratios improved by 2 to 6% for optimistic, and by 6 to 10% for pessimistic, sojourn times.

CHAPTER 6. DISCUSSION

6.1 Criteria 1 to 4

Criterion 1: Evidence from randomized control trials

A systematic review was conducted to examine the evidence for screening women age 70 and over for breast cancer. One RCT (Swedish two-county) and one case-control study (Nijmegen) were identified that included women aged 70 and over and that also published analyses for the older age groups. The results of this systematic review, while suggestive of a positive effect of 20%, are not conclusive, due to broad confidence intervals that included 1.0, but also included point estimates for mortality reductions for women aged 50 to 69. Both studies were limited by small sample sizes, lower participation rates and contamination of the control group. The RCT was further limited as only two screening rounds were completed and the screening interval was reported for all ages to be 33 months. All of these factors interfere with the ability of the RCT to find an effect.

The Nijmegen case-control study design evolved over time as subsequent analyses were published, in order to eliminate the impact of various biases that can occur, especially those related to the pre-clinical duration period (or sojourn time). Where screening participation rates remain low among the oldest women in the population, it is difficult to account fully for self-selection bias of women who may be at higher or lower risk of developing breast cancer, and for length-based sampling resulting from differential detection of slow-growing tumours. The contradictory results of higher breast cancer mortality among the screened group in the Nijmegen study for those aged 75 and over were interpreted as being due to self-selection bias, in that women most likely to develop breast cancer may be those most likely to participate in screening, and thus more likely to be able to die from the disease.

Given these limitations, on balance the results appear to be as or even more consistent with a net benefit from screening, than with no difference. Further, the efficacy for an individual woman who is screened can be higher than the effectiveness found over an entire population as determined by intention-to-treat analysis.[100] Thus, a thorough review of the evidence using other methods was undertaken to understand the impact of extending breast cancer screening to Canadian women aged 70 to 79.

Criterion 2: Burden of breast cancer in women aged 70 to 79

The analysis of Canadian incidence and mortality data showed that the burden of breast cancer in women 70 to 79 is generally closer to that for women aged 60 to 69 than for women aged 80 and over. Age-specific incidence is highest in the 70-79 age group. Women aged 70 to 79 account for the highest proportion of breast cancer deaths of any age group, although breast cancer is less important in comparison to all causes of death. Reducing mortality in this age group could have a significant impact on overall breast cancer mortality. These data clearly show breast cancer to be a significant health problem for women in this age group.

Criterion 3: Screening acceptability

Lower participation rates found in early screening studies in the 1970s along with initial findings in the UK have been used in the past to justify not including women over 69 years of age in organized screening programs.[80] By contrast, over the past 15 years in Canada, screening participation among women aged 70 to 79 continues to increase, and more and more organized programs in Canada are accepting and/or recalling women in this age group, while one program targets women aged 70 to 74. Interim measures of screening effectiveness show that abnormal recall rates and cancer detection rates are higher than for younger women, leading to a higher positive predictive value for women aged 70 to 79. At the same time, the percent of DCIS and the benign to malignant ratio are lower for women

aged 70 to 79 than for younger age groups, while a similar strong shift to lower-staged disease in screen-detected breast cancers is seen. The favourable results for these indicators reflect the lower incidence of benign breast disease and less dense breast tissue, making it easier for mammography to detect breast cancer. These results provide strong positive support for the criterion that mammography is a valid and acceptable screen test.

Criterion 4: Treatment efficacy and availability

Despite limited clinical trial evidence, the Canadian Clinical Practice Guidelines do not recommend different treatment options for healthy women, although with increasing age, frailty and co-morbidity, less aggressive treatment may be considered. Still, an analysis of surgical treatment for breast cancer for Canadian women aged 70 to 79 shows only modest differences in rates compared to those for younger women, indicating that most women in their 70s are likely opting for the standard treatment regimens. More research is needed to assess treatment patterns including radiation and systemic therapies. Clinical trials are needed to assess optimal treatment regimens suitable for older women, especially those with early-staged cancers, those in frail health, and those aged 80 and over.

6.2 Criterion 5: Review of Benefits and Harms

This thesis uses the MISCAN micro-simulation model to explore the benefits and harms of screening. A major advantage of MISCAN over most other models is that it incorporates knowledge of the biology of breast cancer. Thus, MISCAN can consider the effects of slow-growing cancers through alterations to the sojourn time (or pre-clinical detection period). The *optimistic* model assumes only a limited increase in sojourn time after age 70, while the *pessimistic* model assumes continuing large increases as women age.

Source data: Canadian data for population distribution, cancer incidence, overall mortality, and screening patterns were assembled for input into MISCAN in order to adapt it

to the Canadian population. While Canada has a wealth of information, some limitations did occur. Pre-screening breast cancer incidence rates, including the percentage of *in situ* cancers could only be calculated in a narrow window of the 1982-1985 time period. Earlier years could not be used due to poor cancer registration in one province, while by 1985 some screening was likely occurring in some provinces. Mortality data were generally reliable, and were calculated over a three-year period (1995-1997) to increase stability of rates. Population data by single year of age from 90 to 100 and older were obtained from special calculations conducted by the Demography Division at Statistics Canada, again for 1995-1997. However the small size of the 90+ population and the fact that only population counts were used in developing estimates, with no consideration of mortality rates at these ages, may have limited the accuracy of the population data and consequently the ability of the MISCAN model to simulate the eldest populations. Developing realistic screening histories was also challenging. Screening is conducted in both organized and opportunistic settings, with limited relevant information available from the latter setting. While this impeded the ability to construct screening scenarios that fully represented the Canadian reality, several plausible scenarios were developed for this analysis.

Model validation: Adaptation of the MISCAN model to Canada was validated in four ways. First, MISCAN was generally able to reproduce the overall Canadian female population, including the age distribution for most age groups. While a few age groups were more problematic, all MISCAN estimates were within 10% of the actual Canadian population, and values for just 5 of 20 age groups were more than 3% discrepant.

Second, MISCAN estimates for the distribution of breast cancers by stage for both clinically-detected and screen-detected cancers were similar to data available from Canadian sources. However, this validation is limited by the high level of missing data in at least one of the registries and the rather wide range in the clinically detected stage distribution between

the two registries. For the screen-detected cancers, very little missing information occurred in the CBCSD, and the distributions were generally quite close. The strengths of the Canadian data sources for staging were the relatively large numbers of cancers, and the availability of data from population-based cancer registries.

Third, screening volumes were generated from MISCAN using a range of screening participation rates, based on biennial screening of women aged 40 to 79, starting in 1988. Notably, those scenarios established to approximate 1990 and 1996 levels of screening showed the closest correspondence to MISCAN predictions for 1994 and 1996, two recent years for which actual data were available.

Fourth, four screening scenarios, which ranged from relatively low participation in 1990, to moderate in 1996, to a target of about 70% participation, and finally to full 100% participation, were used to explore the ability of MISCAN to predict actual counts by age group for breast cancer mortality in 1997 and breast cancer incidence in 1996, the most recent years for which actual data were available. Again, screening was started in 1988. Remarkably, MISCAN estimates of total 1997 deaths were very close to actual total counts when the 1990 and 1996 screening participation levels were modelled, although more variation occurred within age groups. This is what might be expected as screening generally appears to reduce deaths starting about seven years after inception. Again as expected, the higher the level of screening participation, the greater the reduction in the estimated number of deaths. MISCAN also showed good ability to reproduce the cancer incidence counts for 1996, with estimates being about 5% lower than the actual. Incidence estimates showed less variation by level of screening, reflecting that most prevalent cancers would have been picked up well before 1996. It was interesting to note that, as screening participation increased to 100%, there were *fewer* breast cancers predicted for the 80+ age group, suggesting that screening was picking them up at younger ages. Also, the pessimistic model

at full participation predicted the number of cancers in 1996 almost exactly, although there was more fluctuation by age group.

Overall, these results indicate that the MISCAN model produces reasonable estimates of the volume of screening and its impact on the incidence and mortality in the Canadian population. At the same time, MISCAN is not able to reproduce reality exactly, due in part to the considerable variability in screening participation through a mix of organized and opportunistic approaches in Canada, unlike the apparently more orderly approach taken in many European countries. While further fine-tuning might improve the model slightly, results seldom deviated more than 5 to 10% from what might be expected, and were often much closer. This version of the model was judged adequate to model the impact of screening in Canada.

Global results: The results presented highlight the importance of regular screening at high participation rates to fully experience benefits. Biennial screening participation at a target of 70% will prevent 459 deaths annually, or about 180 more deaths than the 281 prevented by a lower 1990 level of screening. Measurable benefits were found to occur at lower screening participation levels, although in practice, selection bias could affect these one way or the other, if women who are either relatively more or less likely to develop breast cancer participate at different rates. Screening one age group of women has a strong impact on deaths in older age groups. Thus, screening women aged 60 to 69 will influence deaths in women aged 70 and over, while screening women aged 70 to 79 will reduce both incidence and mortality for the 80+ age group.

A key concern in the literature is that screening will detect numerous "extra" cancers in older women that will have limited impact on overall mortality, and will lead to needless worry due to the phenomena of living in lead-time. MISCAN results reported here for extending screening to women aged 70 to 79 reveal a more balanced effect. The "extra"

incidence is calculated as the difference between the higher number of screen-detected cancers, and the considerably reduced number of clinically detected cancers that will occur in the presence of screening. Using optimistic assumptions of sojourn time, the difference amounts to between 210 and 420 cancers per year on average over the 30-year period for screening, or roughly 10% of the 3,500 cancers diagnosed in 1996. These extra cancers are more than offset in optimistic scenarios by the reduction in deaths of from 281 to 574 per year. The situation is somewhat less favourable if the pessimistic assumption is used for sojourn time, as an estimated 1,168 cancers will be detected per year. However, the slower-growing nature of cancer in this scenario also means that women are less likely to die from breast cancer, and 644 deaths will be prevented if all women were screened.

The impact of extra cancers is also offset by the strong favourable stage shifts, with large reductions in women diagnosed with cancers larger than 20mm (Stage T2+), along with large increases in the number diagnosed with cancers sized between 6mm and 20mm (Stages T1b and T1c). Only modest increases were found for *in situ* cancers. These results clearly show that screening can be expected to pick up some extra cancers, at a rate of about 10 per 10,000 screens. Clearly, mammography screening is not just picking up smaller tumours, but is also substantially reducing the number of larger tumours that is essential for subsequent mortality reductions.

From the perspective of the women who are diagnosed, the extra cancers will be detected throughout the screening time period. However from a program implementation perspective, there will be a large increase in prevalent cancers in the first few years of screening as a result of screening women aged 70 to 79 for the first time (or prevalent screen). The impact of prevalent cancers would be reduced if women aged 70 to 79 who were not previously screened were not invited to attend for the first time. At the same time, women could be invited to continue with regular screening initiated in their 50s and 60s,

which better reflects the current Canadian situation. Further development of the MISCAN model would be needed to simulate the situation of continued screening, and exclude the prevalent screen.

The number of life-years lived in lead-time vs the number of life-years gained is an issue for older women. Results for the full model with optimistic assumptions find 298,000 years in lead-time balanced against 141,000 life-years gained. This balance worsens for the pessimistic model, where the number of years lived in lead-time at 625,300 is almost double that for the optimistic model, with only a modest increase in life-years gained to 151,400. The large difference in life-years in lead-time between optimistic and pessimistic assumptions reflects the biologic reality that the earlier-staged but slower-growing cancers detected in the pessimistic model are not very likely to result in death (hence only a small increase in life-years gained). However, these slow-growing cancers are more likely to be detected earlier, with the result that more women will be aware that they have cancer over a longer period of time (i.e, living in lead-time).

The combination of extra cancers diagnosed and treatment, and the consequent increase in life-years in continuing care, have a considerable effect on the quality-adjusted life-years (QALYs), particularly when the gain in life-years is discounted for time preference. For the optimistic model the QALYs were from 18% to 25% to 30% lower than the actual life-years as the discount factor increased from 0% to 3% to 5%, while for the pessimistic model, QALYs were from 29% to 39% to 48% lower. This reduction in QALYs is much greater than the 8% reduction reported by de Haes et al for biennial screening of Dutch women aged 50 to 70. [17] Further, the currently recommended value to use for time preference is 3% could be much higher for elderly women.[20] Thus, the value of screening to a woman aged 70 to 79 will be affected by the extent she wishes to trade off time in good health now (by avoiding breast cancer treatment) with additional years of life in the future.

Derived indicators: Results from the MISCAN model for Canadian women were generally comparable to international results published in the literature. The number of screens needed to prevent one breast cancer death ranged from 733 to 821 for optimistic assumptions, and was actually least for the pessimistic model, at 729 compared to 821 for the optimistic model. This converted to about 90 to 100 screens needed to save one life-year. The overall life expectancy gain was about 4 days per screen, again comparable to the published data from various models showing a range of from about one day up to about 10 days. Since these results are calculated per screen, the value of screening to a woman aged 70 who decides to continue screening for another 10 years would be about 5 times greater, i.e., one would need to screen about 150 to 200 women five times to avert one death. Unfortunately, it was not feasible for this analysis to set up MISCAN to calculate results for an individual woman who participated regularly in screening. Finally, women who do avert a breast cancer death win big as, on average, about eight life-years are gained per death prevented.

Impact of extra follow-up tests: Screening is generally thought to generate considerable extra volumes of diagnostic tests to follow-up on the abnormal mammograms. Per 10,000 women aged 70 to 79, 146 of the 637 receiving an abnormal mammogram (or 25%) will go on to have a biopsy, and of these biopsies about half will involve a cancer. However, as most of the cancers would have eventually been diagnosed, the number of additional biopsies might be as few as 21 per 10,000 screens.

Overall benefits and harms per 10,000 screens: Calculation of benefits and harms per 10,000 screens was useful to quantify and compare the population impact. About 530 to 560 women will experience a false positive result from screening. The 20 to 40 extra biopsies done as a result of screening will find about 10 to 25 extra cancers, and avert from 12.2 to 13.7 deaths. There is a saving of about 100 life-years, but an increase of from 211 to

444 life-years lived in lead-time. (The higher values are for the pessimistic model.) There will be 51 to 59 fewer breast cancers detected clinically and from 60 to 84 cancers detected through screening. Finally, about 30 fewer women will be diagnosed with stage T2+ cancer, about 30 more will be diagnosed with stage T1b or T1c cancers and another 8 with either DCIS or stage T1a cancers. Overall, given the mortality benefits and potentially less invasive therapy for lower-staged cancers, the balance seems quite favourable to continuing to screen women aged 70 to 79. And the much lower rate of clinically detected cancers could lead to less worry for those women who might have noticed clinical symptoms but would not seek medical attention right away.

The balance of benefits and harms worsens somewhat if pessimistic sojourn times are assumed as there will be 25 extra cancers per 10,000 women screened and these will mostly occur at stages T1a and T1b, with only very modest additional reductions in deaths. However it is not clear from this analysis whether the pessimistic or the optimistic model best reflects the Canadian situation. Data presented in Table 5.13 show that the ratio of cancers predicted by applying CBCSD parameters to MISCAN screening volumes compared to the cancers directly estimated by MISCAN are about 8% to 21% higher than those for the optimistic model and about 14% lower than the pessimistic model. The higher results based on the CBCSD rates may be due in part because the CBCSD estimates include a small percentage (about 5%) of cancers detected only through clinical breast examination. If these were eliminated from this calculation the CBCSD estimates would be even closer to the MISCAN estimates for the optimistic model and further apart from those for the pessimistic model. It is also possible that the CBCSD mammography as practiced in the 1990s may have slightly greater sensitivity than that used in the MISCAN model. In general, these results indicate that the sojourn times used in the pessimistic model may be too long, and that the true sojourn times may be closer to those used in the optimistic model.

Sojourn time: MISCAN provides two estimates of the sojourn time (or the pre-clinical detection period) based on the ratio of detection rates at first screening and incidence rates prior to screening implementation. Ratios were calculated from the Swedish two-county trial, the Nijmegan trial, and the Breast Cancer Detection Demonstration Project (BCDDP) and showed that the sojourn time increased from 1.0 to 3.0 years at age 40 up to 2.5 to 4.7 years at age 60. There is some evidence that sojourn time continues to increase up to age 70, when the Swedish two-county detection rates were compared to the control group. However, the values stabilize if compared to the Swedish cancer registry, while sojourn times show, if anything, a slow increase after age 70 in the Nijmegen and BCDDP data. The optimistic sojourn times used in the MISCAN model range from 2.0 years at age 50 to 54, up to 3.4 at age 65 to 69, and then increase very slowly to 4.0 at age 80 to 84. The pessimistic times continue to increase rapidly from 3.4 at age 65 to 69, to 4.6 at 70 to 74, to 6.0 at age 75 to 79, and finally to 8.4 at age 80 to 84.[24]

These sojourn times have been calculated in Northern European and American populations that may or may not be generalizable to Canada. Estimates of sojourn time based on the Screening Mammography Program in British Columbia range from 1.6 years for ages 40 to 49, 1.88 for ages 50 to 59, 2.42 for ages 60 to 69, up to 2.8 for ages 70 to 74.[101] An estimate from the CBCSD using similar methodology to that used to produce the MISCAN estimates ranges from 2.7 for ages 50 to 54, to 3.7 for 60 to 64, to 4.5 for 70 to 74, and 5.3 for 80 and over. (D. Schanzer, personal communication). Although it is difficult to directly compare the absolute values of sojourn times given differences in methodologies, the general impression is that while sojourn time continues to increase with age, it does not increase in Canadian women as rapidly as in the pessimistic model.

Limitations: While MISCAN produces more relevant output than nearly any other model, some limitations arose with the models implemented for this thesis. First, the model

was not developed to simulate continuing screening and thereby exclude the first prevalent screen for women aged 70 to 79. Second, without considerable further work, results could not be calculated per woman screened. Third the model assumes a 30% improvement in prognosis (or mortality reduction based on the Swedish trial results for 50 to 69), which is higher than the point estimate from the RCT. Modelling these scenarios would require specifying, in detail, appropriate cohorts and their associated mortality rates as part of the model input.

On balance, favourable population impacts were found for continuing to screen women aged 70 to 79 due to the mortality reductions and lower-staged disease, which generally more than offset the increase in extra cancers. Quality-adjustment of the results, however, shows that time preference is important, as the quality-adjusted life-years with 5% discounting were 31% less than unadjusted life-years with no discounting. Results are less favourable if longer sojourn times are assumed, but are moderately improved if a utility for initial therapy is used that reflects the more benign and appropriate tamoxifen therapy generally provided to the vast majority of women in this age group. Comparisons of screen-detected cancers in Table 5.13 based on direct MISCAN estimates and indirect estimates calculated by applying CDCSD rates to MISCAN estimates of the number of screens indicate that the true sojourn times may be closer to the shorter optimistic times. Further, the direct calculation of sojourn times appear to be closer to the optimistic than pessimistic assumptions. While alternate mortality reductions were not modelled, it seems likely that the 30% reduction may represent a balance between the 20% point reduction observed in the Swedish two-county study and the higher reduction found in the Nijmegen case-control study. In other words, for women who actually participate in screening, the 30% may represent a lower limit to the expected mortality reduction.[100] Overall, these results are consistent with previously published work that showed a generally favourable balance in

favour of screening up until at least age 80, and that the optimal screening interval for ages 65 and older is two years.[25 26]

6.3 Other considerations

Further work: Some sensitivity analysis has already been done by comparing the various screening participation levels and by comparing optimistic and pessimistic assumptions of sojourn times. Setting up MISCAN as a cohort model would enable results to be calculated for continuing to screen women aged 70 to 79 and to calculate results per woman, as well as to permit the mortality reduction to be varied from about 30% to 20% for women aged 70 and over. Another possibility would be to vary the sensitivity of the test. This is a particularly intriguing adaptation of the model, due to possible variations in sensitivity of mammography units across Canada and the difficulty some units are having to meet accreditation standards due to the relatively old age of the equipment.

Other criteria in making a decision regarding screening policy: This thesis examines five criteria identified by various sources as important to assess prior to taking the decision to implement a screening program. Some information has also been presented on additional criteria not assessed. The question of the presence of a significant detectable pre-clinical period (DPCP) was not addressed directly, as this has been well-established for breast cancer in women of all ages. However, the length of the DPCP does vary, being somewhat shorter at younger ages, and longer at older ages. The question of sojourn times also deserves more attention based on empirical Canadian data, given the importance of this parameter on overall model results.

Health system resource implications: Other criteria include ensuring that there are sufficient resources in the health care system to provide all the necessary screening, follow-up diagnostic and treatment. Results from this model provide some insight. The volume of

additional tests seems modest, particularly with optimistic assumptions, and would not be that large even with pessimistic assumptions, were screening restricted to those women who had already been screened at least once prior to age 70. A point in favour of screening is that streamlined follow-up processes may lead to more efficient use of these resources.

Treatment: While MISCAN is limited in that it only considers size of the tumour and not the nodal status, it is nevertheless true that smaller tumours are less likely to have involved lymph nodes. The earlier stage of disease found through screening should, for at least some cancers, reduce the amount and extent of treatment needed. Less aggressive therapy may be an option for some of these early-stage node-negative cancers, and particularly for DCIS. A consideration is the post-operative mortality rate, which does increase with age. This thesis assumes that all women diagnosed with breast cancer will receive one year of initial treatment having a utility of 0.750. A sensitivity analysis was also conducted assuming a utility of 0.900 to reflect that the vast majority of women aged 70+ will receive the relatively benign tamoxifen. The improved results with this higher utility indicate that to the extent that less arduous treatment is available resulting in a shorter time frame or a higher utility or both, the unfavourable aspects of screening could be further reduced. More research in this area would help to reduce the negative impact of breast cancer in elderly women.

Cost-effectiveness: Most analyses of screening effectiveness include an estimate of cost per quality-adjusted life-years saved. As this thesis aimed to concentrate on the population health impacts, this calculation was not done for Canada. However, a number of models have calculated the cost per life-year saved (CLYS) for screening women; most results are in the range of US $20,000 to US $40,000 (Appendix B).[18 23 26 33 34 39] The CLYS rises with increasing age and screening frequency and is most sensitive to the cost of the screening mammogram. (One study found a CLYS of US $30,000 for mammograms costing

US $55, and US $39,000 for mammograms costing US $115.[23]) These results are all within the borderline effectiveness range of CDN $20,000 to CDN $100,000 proposed by Laupacis.[102]

One recent study has calculated a higher marginal CLYS of US $66,800, which is close to the upper end of this limit, for continuing to screen women aged 70 to 79 as compared to screening women up to age 69.[20] The higher costs are attributed by the authors to higher mammography costs (US $116 instead of US $55 to US $115), costs adjusted to 1998 US$ (instead of 1984 and 1990 costs) and the inclusion of costs of detecting and treating DCIS (assessed as about 17% of all cancers). The cost of initial treatment in the US was estimated as $27,000 for DCIS, $37,000 for invasive breast cancer in the screened group and $39,000 for the unscreened group, in 1998 US$.

The CLYS as calculated for US populations may be considerably higher than those that would apply to Canada. Rates for mammography in Canada are somewhat lower, depending upon whether a diagnostic or screening fee-for-service is applied, and generally do not exceed CDN $100, depending upon the province. Initial treatment costs in the US are much higher than the initial costs in Canada, calculated as $8,286 (1995 CDN $) for Stage I and II breast cancer.[103] However, lifetime treatment costs in Canada are higher, ranging from CDN $23,275 for Stage I up to CDN $36,340 for Stage IV disease.[104]

Health status: This thesis does not directly address the range in health status that occurs in women aged 70 to 79 and which may affect the policy decision to implement screening for older women. Women who have poor health status and perceived lower life expectancy may benefit less from screening. Mandelblatt et al showed that severe co-morbidity reduced the benefit of screening by about half.[18] Kerlikowske et al reported that low bone mineral density was an indicator of low risk of breast cancer and concluded that mammography was not cost-effective in women aged 70 to 79 with low bone mineral

density.[20] More work could be done in Canada to assess the impact of health status on screening effectiveness.

Overall life expectancy: Canadian women experience the second highest rate in the world for breast cancer and have one of the highest life expectancies of any national population. Thus older Canadian women have the potential to benefit more from screening than any other population. Competing causes of death are an important concern, although their impact among women aged 70 to 79 may be less than expected. Life table analysis (see Appendix D) indicates that a woman aged 60 has a 6.5% chance of dying by age 70, while a 70 year-old has a 20% chance of dying by age 80, and an 80 year-old a 35% chance of dying by age 90. A woman diagnosed with a clinically detected breast cancer has a 40% chance of dying of breast cancer in the next ten years.[105] Results presented in this thesis indicate that on average eight life-years are saved for every breast cancer death prevented. Thus, saving a women from dying of breast cancer can extend life for some years until death from another cause.

CHAPTER 7. SUMMARY AND CONCLUSIONS

Results presented in this thesis show strong positive support for continuing to screen women aged 70 to 79 on two criteria, and provide moderate support on three others. No criterion contra-indicated screening. The evidence from the randomized control trial and the case control study is inconclusive. However, given the low participation rates and other study limitations, most of which bias the results toward the null, the evidence can be interpreted as leaning towards a benefit. Strong support for screening is provided by the analysis for two criteria regarding the burden of disease and the availability of an acceptable and valid screening test. With regards to the fourth criteria, effective treatment is available and appears to be being used, but more research is needed in this area.

The relative benefits and harms due to screening were assessed (criterion 5) using a model-based approach. Results highlight the importance of regular screening at high participation rates to fully experience benefits; however significant benefits occur even at relatively modest screening participation levels. Overall, the balance of benefits and harms was moderately favourable. The impact of "extra" cancers was less than expected in terms of their overall number and the harms ensuing from these "extra" cancers was balanced by the much lower stage distribution of all cancers. Quality adjustment of life-years gained, together with discounting for time preference, reduced the benefits by up to 31%, although this was improved to a 25% reduction if utilities for time in initial therapy are adjusted upwards to account for greater use of tamoxifen among women over age 70. The benefit is reduced still further if pessimistic assumptions regarding sojourn times prevail, or if mortality reductions are less than the 30% reduction modelled. Further sensitivity analysis could provide additional insights into these issues. On the other hand, benefits could be greater than anticipated if the mortality reduction for women who are screened were greater than that

modelled, while harms could be further reduced by optimizing treatment of early-stage cancer for the elderly.

These results describe the impact of extending screening to women aged 70 to 79 on a population basis. There may be trade-offs for individual women who are particularly anxious or who want to avoid bothersome appointments. Women in very poor health may not live the five to seven years needed to benefit from screening, or may be too frail to withstand treatment. The preference of a woman for healthy life-years now vs additional future life-years is also important.

Taken together, these results support continuing to screen healthy women aged 70 to 79 for breast cancer. Further work is recommended to refine the model and to document the benefits and harms when smaller mortality reductions are assumed. This additional effort could also assess the net benefits to be gained by continuing to screen women aged 70 to 79 as opposed to screening all women, evaluate alternative sojourn times, and determine the cost-effectiveness of continued screening. A case-control study of the value of breast screening in women aged 70 to 79 would also be enlightening, given the considerable levels of screening now occurring in elderly women.

REFERENCES

1. Olivotto IA, Bajdik CD, Plenderleith IH *et al.* Adjuvant systemic therapy and survival after breast cancer. N Engl J Med 1994; 330(12):805-10.

2. Gaudette LA, Wysocki M, Gao R-N, Nault F. Update on breast cancer mortality, 1995. Health Rep 1997; 9(1):31-4.

3. Kerlikowske K, Grady D, Rubin SM, Sandrock C, Ernster VL. Efficacy of screening mammography. A meta-analysis. JAMA 1995; 273(2):149-54.

4. Demissie K, Mills OF, Rhoads GG. Empirical comparison of the results of randomized controlled trials and case-control studies in evaluating the effectiveness of screening mammography. J Clin Epidemiol 1998; 51(2):81-91.

5. Millar WJ, David P. Life Tables, Canada and Provinces, 1990-1992. Statistics Canada cat no. 84-537. Ottawa: Minister of Industry, 1995.

6. Lipskie TL. A summary of cancer screening guidelines. Chronic Dis Can 1998; 19(3):112-30.

7. Health Canada. Organized Breast Cancer Screening Programs in Canada - 1996 Report. cat. no. H1-9/13-1999. Ottawa: Minister of Public Works and Government Services, 1999.

8. Thompson RS, Barlow WE, Taplin SH *et al.* A population-based case-cohort evaluation of the efficacy of mammographic screening for breast cancer [see comments]. Am J Epidemiol 1994; 140(10):889-901.

9. Wilson JMG, Jungner G. Principles and Practice of Screening for Disease. Geneva: World Health Organization, 1968. (Public Health Papers; 34).

10. Cadman D, Chambers L, Feldman W, Sackett D. Assessing the effectiveness of community screening programs. JAMA 1984; 251(12):1580-5.

11. Muir Gray JA. Evidence-Based Health Care. New York: Churchill Livingstone, 1997.

12. Miller AB. Is routine mammography screening appropriate for women 40-49 years of age? Am J Prev Med 1991; 7(1):55-62.

13. Harris R, Leininger L. Clinical strategies for breast cancer screening: weighing and using the evidence. Ann Intern Med 1995; 122(7):539-47.

14. Ernster VL, Barclay J, Kerlikowske K, Wilkie H, Ballard-Barbash R. Mortality among women with ductal carcinoma in situ of the breast in the population-based

surveillance, epidemiology and end results program. Arch Intern Med 2000; 160(7):953-8.

15. Last JM, Ed. A Dictionary of Epidemiology. 4th edition. New York: Oxford University Press, 2000.

16. Steggles S, Lightfoot N, Sellick SM. Psychological distress associated with organized breast cancer screening. Cancer Prev Control 1998; 2(5):213-20.

17. de Haes JC, de Koning HJ, van Oortmarssen GJ, van Agt HM, de Bruyn AE, van Der Maas PJ. The impact of a breast cancer screening programme on quality-adjusted life-years. Int J Cancer 1991; 49(4):538-44.

18. Mandelblatt J, Wheat ME, Monane M, Moshief RD, Hollenberg JP, Tang J. Breast cancer screening for elderly women with and without comorbid conditions. A decision analysis model. Ann Intern Med 1992; 116(9):722-30.

19. Lindfors KK, Rosenquist CJ. Needle core biopsy guided with mammography: a study of cost-effectiveness. Radiology 1994; 190(1):217-22.

20. Kerlikowske K, Salzmann P, Phillips KA, Cauley JA, Cummings SR. Continuing screening mammography in women aged 70 to 79 years: impact on life expectancy and cost-effectiveness. JAMA 1999; 282(22):2156-63.

21. Eddy DM. A computer-based model for designing cancer control strategies. In: Greenwald P, Sondik EJ, editors. NCI Monographs: Cancer Control Objectives for the Nation 1985-2000. Vol. 2. Bethesda: National Cancer Institute, 1986: 75-82.

22. Levin DL, Gail MH, Kessler LG, Eddy DM. A model for projecting cancer incidence and mortality in the presence of prevention, screening, and treatment programs. In: Greenwald P, Sondik EJ, editors. NCI Monographs: Cancer Control Objectives for the Nation: 1985-2000. Vol. 2. Bethesda: National Cancer Institute, 1986: 83-93.

23. Brown ML. Economic considerations in breast cancer screening of older women. J Gerontol 1992; 47 Spec No:51-8.

24. Boer R, de Koning HJ, van Oortmarssen GJ, van der Maas PJ. In search of the best upper age limit for breast cancer screening. Eur J Cancer 1995; 31A(12):2040-3.

25. Boer R, de Koning H, Threlfall A *et al.* Cost effectiveness of shortening screening interval or extending age range of NHS breast screening programme: computer simulation study. BMJ 1998; 317(7155):376-9.

26. Boer R, de Koning HJ, van der Maas PJ. A longer breast carcinoma screening interval for women age older than 65 years? Cancer 1999; 86(8):1506-10.

27. Jansen JT, Zoetelief J. MBS: a model for risk benefit analysis of breast cancer screening. Br J Radiol 1995; 68(806):141-9.

28. Jansen JT, Zoetelief J. Assessment of lifetime gained as a result of mammographic breast cancer screening using a computer model. Br J Radiol 1997; 70(834):619-28.

29. O'Neill TJ, Tallis GM, Leppard P. A review of the technical features of breast cancer screening illustrated by a specific model using South Australian cancer registry data. Stat Methods Med Res 1995; 4(1):55-72.

30. Will BP, Le Petit C, Berthelot JM, Tomiak EM, Verma S, Evans WK. Diagnostic and therapeutic approaches for nonmetastatic breast cancer in Canada, and their associated costs. Br J Cancer 1999; 79(9-10):1428-36.

31. Law MR, Morris JK, Wald NJ. The importance of age in screening for cancer. J Med Screen 1999; 6(1):16-20.

32. Rich JS, Black WC. When should we stop screening? Eff Clin Pract 2000; 3(2):78-84.

33. Lindfors KK, Rosenquist CJ. The cost-effectiveness of mammographic screening strategies [see comments] [published erratum appears in JAMA 1996 Jan 10;275(2):112]. JAMA 1995; 274(11):881-4.

34. Rosenquist CJ, Lindfors KK. Screening mammography beginning at age 40 years: a reappraisal of cost-effectiveness. Cancer 1998; 82(11):2235-40.

35. Wright JC, Weinstein MC. Gains in life expectancy from medical interventions--standardizing data on outcomes. N Engl J Med 1998; 339(6):380-6.

36. Rembold CM. Number needed to screen: development of a statistic for disease screening. BMJ 1998; 317(7154):307-12.

37. Kopans DB. An overview of the breast cancer screening controversy. J Natl Cancer Inst Monogr 1997; (22):1-3.

38. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. N Engl J Med 1988; 318(26):1728-33.

39. Eddy DM. Screening for breast cancer. Ann Intern Med 1989; 111(5):389-99.

40. Wright CJ, Mueller CB. Screening mammography and public health policy: the need for perspective [see comments]. Lancet 1995; 346(8966):29-32.

41. de Koning HJ. Screening for breast cancer, time to think--and stop? National Evaluation Team for Breast Cancer Screening (NETB). Lancet 1995; 346(8972):438-9.

42. Norum J. Breast cancer screening by mammography in Norway. Is it cost-effective? Ann Oncol 1999; 10(2):197-203.

43. Boer R. From Evidence to Decision Support in Cancer Screening: applications of MISCAN models. Chapter 8: In search of the best upper age limit for breast cancer screening. 2000.

44. Hall J, Gerard K, Salkeld G, Richardson J. A cost utility analysis of mammography screening in Australia. Soc Sci Med 1992; 34(9):993-1004.

45. Weinstein MC, Siegel JE, Gold MR, Kamlet MS, Russell LB. Recommendations of the Panel on Cost-effectiveness in Health and Medicine. JAMA 1996; 276(15):1253-8.

46. Petitti DB. Advanced Cost-Effectiveness Analysis. In: Petitti DB, editor. Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis. New York: Oxford University Press, 2000: 182-201.

47. Lipscomb J, Weinstein MC, Torrance GW. Time Preference. In: Gold M, Sigegel JE, Russell LB, Weinstein MC, editors. Cost-Effectiveness in Health and Medicine. New York: Oxford University Press, 1996: 214-35.

48. Gotzsche PC, Olsen O. Is screening for breast cancer with mammography justifiable? Lancet 2000; 355(9198):129-34.

49. Palli D, Del Turco MR, Buiatti E *et al.* A case-control study of the efficacy of a non-randomized breast cancer screening program in Florence (Italy). Int J Cancer 1986; 38(4):501-4.

50. Janzon L, Andersson I. The Malmo mammographic screening trial. In: Miller AB, Chamberlain J, Day NE, Hakama M, Prorok PC, editors. Cancer Screening . Cambridge: Cambridge University Press (on behalf of UICC, International Union Against Cancer), 1991: 37-44.

51. Van Dijck JA, Verbeek AL, Beex LV *et al.* Mammographic screening after the age of 65 years: evidence for a reduction in breast cancer mortality. Int J Cancer 1996; 66(6):727-31.

52. Tabar L, Fagerberg CJ, Gad A *et al.* Reduction in mortality from breast cancer after mass screening with mammography. Randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. Lancet 1985; 1(8433):829-32.

53. Nystrom L, Rutqvist LE, Wall S *et al.* Breast cancer screening with mammography: overview of Swedish randomised trials [see comments] [published erratum appears in Lancet 1993 Nov 27;342(8883):1372]. Lancet 1993; 341(8851):973-8.

54. Tabar L, Fagerberg G, Chen H-H *et al.* Efficacy of breast cancer screening by age. New results from the Swedish Two-County Trial. Cancer 1995; 75(10):2507-17.

55. Chen H-H, Tabar L, Fagerberg G, Duffy SW. Effect of breast cancer screening after age 65 [published erratum appears in J Med Screen 1995;2(2):118]. J Med Screen 1995; 2(1):10-4.

56. Gullberg B, Andersson I, Janzon L, Ranstam J. Screening mammography. Lancet 1991; 337(8735):244.

57. Morrison AS, Brisson J, Khalid N. Breast cancer incidence and mortality in the Breast Cancer Detection Demonstration Project. J Natl Cancer Inst 1988; 80(19):1540-7.

58. van Dijck JA, Holland R, Verbeek AL, Hendriks JH, Mravunac M. Efficacy of mammographic screening of the elderly: a case-referent study in the Nijmegen program in The Netherlands. J Natl Cancer Inst 1994; 86(12):934-8.

59. Verbeek AL, Hendriks JH, Holland R, Mravunac M, Sturmans F. Mammographic screening and breast cancer mortality: age-specific effects in Nijmegen Project, 1975-82. Lancet 1985; 1(8433):865-6.

60. Verbeek AL, Hendriks JH, Holland R *et al.* Screening and breast cancer. Lancet 1984; 2(8404):690.

61. Verbeek AL, Hendriks JH, Holland R, Mravunac M, Sturmans F, Day NE. Reduction of breast cancer mortality through mass screening with modern mammography. First results of the Nijmegen project, 1975-1981. Lancet 1984; 1(8388):1222-4.

62. Brown JT, Hulka BS. Screening mammography in the elderly: a case-control study. J Gen Intern Med 1988; 3(2):126-31.

63. Andersson I, Aspegren K, Janzon L *et al.* Mammographic screening and mortality from breast cancer: the Malmo mammographic screening trial. BMJ 1988; 297(6654):943-8.

64. Nystrom L, Larsson LG, Wall S *et al.* An overview of the Swedish randomised mammography trials: total mortality pattern and the representivity of the study cohorts. J Med Screen 1996; 3(2):85-7.

65. Egger M, Schneider M, Davey Smith G. Spurious precision? Meta-analysis of observational studies. BMJ 1998; 316(7125):140-4.

66. Stroup DF, Berlin JA, Morton SC *et al.* Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA 2000; 283(15):2008-12.

67. Shapiro S. Is there is or is there ain't no baby? Dr. Shapiro replies to Drs. Petitti and Greenland. Am J Epidemiol 1994; 140(9):788-91.

68. Greenland S. Can meta-analysis be salvaged? Am J Epidemiol 1994; 140(9):783-7.

69. Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. Am J Epidemiol 1994; 140(3):290-6.

70. Petitti DB. Of babies and bathwater. Am J Epidemiol 1994; 140(9):779-82.

71. Walter SD, Jadad AR. Meta-analysis of screening data: a survey of the literature. Stat Med 1999; 18(24):3409-24.

72. Clarke EA, Anderson TW. Does screening by "Pap" smears help prevent cervical cancer? A case-control study. Lancet 1979; 2(8132):1-4.

73. Friedman DR, Dubin N. Case-control evaluation of breast cancer screening efficacy. Am J Epidemiol 1991; 133(10):974-84.

74. Weiss NS. Application of the case-control method in the evaluation of screening. Epidemiol Rev 1994; 16(1):102-8.

75. Etzioni RD, Weiss NS. Analysis of case-control studies of screening: impact of misspecifying the duration of detectable preclinical pathologic changes. Am J Epidemiol 1998; 148(3):292-7.

76. Gaudette LA, Silberberger C, Altmayer CA, Gao RN. Trends in breast cancer incidence and mortality. Health Rep 1996; 8(2):29-37(Eng); 31-40(Fre).

77. National Cancer Institute of Canada. Canadian Cancer Statistics 2000. Toronto: The Institute, 2000.

78. Faulk RM, Sickles EA, Sollitto RA, Ominsky SH, Galvin HB, Frankel SD. Clinical efficacy of mammographic screening in the elderly. Radiology 1995; 194(1):193-7.

79. Hwang E-S, Cody HS. Does the proven benefit of mammography extend to breast cancer patients over age 70? South Med J 1998; 91(6):522-6.

80. Forrest AP, Aitken RJ. Mammography screening for breast cancer. Annu Rev Med 1990; 41:117-32.

81. Gaudette LA, Altmayer CA, Nobrega KM, Lee J. Trends in mammography utilization, 1981 to 1994. Health Rep 1996; 8(3):17-27(Eng); 19-30(Fre).

82. Miller AB. Screening for cancer of the breast. In: Miller AB, editor. Screening for Cancer. Toronto: Academic Press, 1985: 325-45.

83. Threlfall AG, Woodman CB. Risk of breast cancer in women who attend the NHS breast screening programme: cohort study. BMJ 2001; 323(7305):140.

84. Chu KC, Tarone RE, Kessler LG *et al.* Recent trends in U.S. breast cancer incidence, survival, and mortality rates [see comments]. J Natl Cancer Inst 1996; 88(21):1571-9.

85. The Steering Committee on Clinical Practice Guidelines for the Care and Treatment of Breast Cancer. Clinical practice guidelines for the care and treatment of breast cancer: a Canadian consensus document. CMAJ 1998; 158 Suppl 3:S1-2.

86. Silliman RA, Balducci L, Goodwin JS, Holmes FF, Leventhal EA. Breast cancer care in old age: what we know, don't know, and do. J Natl Cancer Inst 1993; 85(3):190-9.

87. Goodwin JS, Freeman JL, Freeman D, Nattinger AB. Geographic variations in breast cancer mortality: do higher rates imply elevated incidence or poorer survival? Am J Public Health 1998; 88(3):458-60.

88. De Koning HJ, Fracheboud J, Boer R *et al.* Nation-wide breast cancer screening in The Netherlands: support for breast-cancer mortality reduction. National Evaluation Team for Breast Cancer Screening (NETB). Int J Cancer 1995; 60(6):777-80.

89. van Oortmarssen GJ, Habbema JD, van der Maas PJ *et al.* A model for breast cancer screening. Cancer 1990; 66(7):1601-12.

90. van den Akker-van Marle E, de Koning H, Boer R, van der Maas P. Reduction in breast cancer mortality due to the introduction of mass screening in The

Netherlands: comparison with the United Kingdom. J Med Screen 1999; 6(1):30-4.

91. de Koning HJ, Boer R, Warmerdam PG, Beemsterboer PM, van der Maas PJ. Quantitative interpretation of age-specific mortality reductions from the Swedish breast cancer-screening trials [see comments]. J Natl Cancer Inst 1995; 87(16):1217-23.

92. Loeve F, Boer R, van Oortmarssen GJ, van Ballegooijen M, Habbema JD. The MISCAN-COLON simulation model for the evaluation of colorectal cancer screening. Comput Biomed Res 1999; 32(1):13-33.

93. Health and Welfare Canada. User's Guide to 40 Community Health Indicators Ottawa: Minister of Supply and Services Canada (Cat. No. H39-238/1992E), 1992.

94. Péron Y, Strohmenger C. Demographic and Health Indicators. Presentation and Interpretation. Ottawa: Statistics Canada, 1985.

95. Spasoff RA. Epidemiologic Methods for Health Policy. New York: Oxford University Press, 1999.

96. Selvin S. Statistical Analysis of Epidemiologic Data. New York: Oxford University Press, 1996.

97. Boer R. From Evidence to Decision Support in Cancer Screening - applications of MISCAN models. Chapter 12. Stage distribution at first and repeat examinations in breast cancer screening. Rotterdam: Erasmus University, 2000.

98. Health Canada. Organized Breast Cancer Screening Programs in Canada- 1997 and 1998 Report. cat. no. H1-9/13-1998. Ottawa: Minister of Public Works and Government Services Canada, 2001.

99. Olivotto I, Bancej C, Goel V. *et al.* Waiting times from abnormal breast screen to diagnosis in 7 Canadian provinces. CMAJ 2001; 165(3):277-83.

100. Tabar L, Vitak B, Tony HH, Yen MF, Duffy SW, Smith RA. Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality. Cancer 2001; 91(9):1724-31.

101. Chan, KJ. A Semi Markov Model for Mammographic Detection of Breast Cancer. Vancouver: University of British Columbia, 1999.

102. Laupacis A, Feeny D, Detsky AS, Tugwell PX. How attractive does a new technology have to be to warrant adoption and utilization? Tentative guidelines for using clinical and economic evaluations. CMAJ 1992; 146(4):473-81.

103. Evans WK, Will BP, Berthelot JM, Logan DM, Mirsky DJ, Kelly N. Breast cancer: better care for less cost. Is it possible? Int J Technol Assess Health Care 2000; 16(4):1168-78.

104. Will BP, Berthelot JM, Le Petit C, Tomiak EM, Verma S, Evans WK. Estimates of the lifetime costs of breast cancer treatment in Canada. Eur J Cancer 2000; 36(6):724-35.

105. National Cancer Institute of Canada. Canadian Cancer Statistics 1993. Toronto: The Institute, 1993.

APPENDIX A: LITERATURE REVIEW STRATEGY

The strategy below was used to search MEDLINE from 1989 onwards to obtain papers relating to breast cancer screening in women aged 70 and over. Search results shown below for 1991 to 1995 illustrate the logic and indicate the volume at each step.

No.	Records	Request	Comments
1	11,766	"Breast-Neoplasms"/ diagnosis , economics , epidemiology , mortality , pathology , prevention-and-control , radiography	*Selects for selected breast mesh headings*
2	7,706	"Mass-Screening"/ all subheadings	*Selects for screening*
3	74	"Multiphasic-Screening"/ all subheadings	*Selects for screening*
4	7,776	#2 or #3	*Combined screening set*
5	2,903	explode "Mammography"/ all subheadings	*Selects all mammography*
6	226,468	"Aged"/ all subheadings	*Selects for aged*
7	1,569	#6 in mjme	*Selects aged as a major MeSH heading*
8	64,864	"Aged-80-and-over"/ all subheadings	*Another approach to elderly*
9	65,986	#7 or #8	*Combined 80+ and aged major*
10	19,685	elderly or (older adj woman) or (older adj women)	*Another approach to select older women*
11	69,601	(age* or over) near1 (65 or 66 or 69 or 70 or 74 or 75 or 80)	*Another approach to select older women*
12	81,598	#9 or #10 or #11	*Combined set of older women*
13	1,189	#1 and #4	*Combined set for breast and screening*
14	3,276	#13 or #5	*Combined set of breast and screening and mammography*
15	423	#14 and #12	*Combined set of breast screening, mammography and older women*
16	376	#15 and (la=english)	*Restricts set to English*
17	116,796	pt=letter	

18	364	#16 not #17	*Removes letters from set*
19	5,998	explode "Radiopharmaceuticals"/ all subheadings	*Selects radiopharmaceuticals*
		Note: Steps 19 and 20 removes papers on technical aspects of screening, which were quite numerous in some years	
20	362	#18 not #19	*Removes papers with a mesh heading of radiopharmaceuticals*
21	218,551	pt=review	*Selects review articles*
22	330	#20 not #21	*Removes review articles from main set*
23	1,704	#5 in mjme	*Selects mammography as a major mesh heading*
24	4,266	#4 in mjme	*Selects mass screening as a major set*
25	168	#22 and #23	*Combined set of major mammography and non- review articles*
26	168	#22 and #24	*Combined set of major mass screening and non- review articles*
27	12	#22 and #7	*Combined set of major aged and non-review articles*
* 28	176	#25 or #26 or #27	*Combined set of articles in any one of the above 3 sets.*
29	32	#20 and #21	*Review articles*

APPENDIX B: Comparison of Models Used to Assess Breas

Model >>	MISCAN	Model for Breast Screening:MBS	Mathematical Model	C/
Authors	Boer, de Koning, van Oortmarssen et al (1990-1999)[1 2 3 4 5]	Jansen and Zoetelief, (1995, 1997)[6 7 8]	O'Neill, Tallis, Leppard, (1995)[9]	Eddy (198 (1986)[11]; E Brown (19
Type	Micro-simulation	Micro-simulation (Monte Carlo)	Micro-simulation	Macro-sim chains
Basic Steps	For each individual, calculates presence of pre-clinical disease (PCD); then duration of PCD at each stage; applies screening sensitivity; calculates transition to breast cancer (5 stages); then transition to death from breast cancer or other causes; output can be used for QALYs and costs.	Randomly simulates individual breast tumours, age at tumour onset, tumour growth rate, life-years expectancy, tumour size threshold for clinical and screen-detected cancer. Only model to assess *risk of radiation-induced cancer*; also assesses benefit from improved survival in smaller tumours.	Starts cohort at age 0, applies breast cancer initiation function, assigns tumour growth rates to calculate size, and determines rate of first report of breast cancer; when screening introduced, models with detection rate dependent on size; then applies survival function based on age.	Starts pop no cancer and sex-s rates and distributio Deaths fr can occur Assigns c to each st Calculate costs per (C/LS). preventio treatment
Population	Can generate results to compare with observed rates in population, e.g., for Swedish trials or UK women.	Simulated 1,000,000 Swedish women.	• Validated on South Australia population. • Modelled on hypothetical 100,000 women.	Any synth populatio groups, e aged 65+ 50-84.
Age Range	Most use 40-69; 3 articles focus on 70+	Starts at age 20, results presented to age 95-99	Screening starting at age 30, 40, 50, 60 and 70	ages 65-1 and 50-6
Study Design Features	Two variants: • optimistic (no increase in sojourn time); and • pessimistic (increase in sojourn time). Also explores screening interval on effect of continuing to screen.	Includes risks of radiation-induced cancers; can vary parameters such as: • tumour growth rate; • detection rates; • screening interval; • age-dependent survival.	• Modelled effect of screening, starting at the beginning of each 10-year age group. • Focus on mathematical derivation of various parameters and functions, with clear explanations.	Eddy (19 impact of acceptan 30% and Levin (19 preventio Eddy (19 detailed a aged 65- Brown (1 cost-effe mammo frequenc

ss Breast Cancer Screening for Women Aged 70 and Over

CAN*TROL	Decision Analysis: ages 65+, Co-morbidity	Decision Analysis: ages 40-79, Strategies	Decision Analysis: ages 65-79, BMD
Eddy (1986)[10] ; Levin (1986)[11]; Eddy (1989)[12]; Brown (1992)[13]	Mandelblatt, Wheat, et al, (1992)[14]	Lindfors & Rosenquist, (1994) (1995) (1997)[15 16 17]	Kerlikowske et.al., (1999)[18]
Macro-simulation, Markov chains	Decision Analysis, (SmlTree)	Decision Analysis, using (SmlTree) Markov model	Decision Analysis, using Markov model
Starts population in state 0, no cancer; then applies age- and sex-specific incidence rates and a stage distribution (5 stages). Deaths from other causes can occur each year. Assigns cancer death rates to each stage /state. Calculates life years, and costs per life year saved (CYLS). Can model prevention, screening and treatment interventions.	Starts with no screen (=> late Dx) vs screen. then abnormal => biopsy vs normal screen =>interval cancer. Then new tree: cancer vs. no cancer; if cancer, then local, regional, distant; if local or regional, then surgery, then die or survive surgery; if no cancer, then estimates mortality rates for level of co-morbidity.	Starts with mammo, then +ve or -ve outcome; +ve outcome => FU mammo or biopsy (by type), then M or B biopsy, then treatment and then determine, death (br ca, oth) or next mammo. For -ve mammo, then death (oth), interval cancer, or next scrn mammo; then determine prob and cost.	Starts with healthy women age 65, screened regularly in previous 5 years. Screen every 2 y. At end of each 1y cycle, women will be: healthy; develop DCIS & alive; develop breast cancer & alive; die of breast cancer after Dx of breast cancer OR DCIS; died of other causes.
Any synthetic or real population and/or sub-groups, e.g., US women aged 65+, or women aged 50-84.	Synthetic (US): compare 20 groups of women by age group and co-morbidity.	Compare 2 hypothetical populations (US): • screened • observed, not screened.	Compare US female populations ages 65-79.
ages 65-100, 65-80, 65-75; and 50-64, 65-69, ... 80-84	age 65 and over, by 5y age group, up to 85+	Various age groups, primarily : 40-79, 50-79	Age groups: 65-69, 65-79
Eddy (1986) examines impact of mammography acceptance rates of 100%, 30% and 10% for age 65+; Levin (1986) looks at prevention, screening, Tx; Eddy (1989) conducts detailed analysis for women aged 65-75; Brown (1992) examines cost-effectiveness by cost of mammo & screen frequency.	Three levels of co-morbidity: • normal, • medium (ie, hypertension), and • high (ie, congestive heart failure) Two QoL adjustments: • long term = stage distn • short term = screening impacts	Measures cost effectiveness as marginal cost per life-year saved (MYCLYS) for: • various screening strategies; • various age ranges; and • biennial vs annual screens. Oldest age group modelled separately is 50-79.	Biennial screening for age 65-69, then: • stop: or • continue to 79 for: - all women; or - only for top 3 quartiles of bone mass density (BMD)

Model >>	MISCAN	Model for Breast Screening:MBS	Mathematical Model	CAN
Frequency of and Participation in Screening	Can be varied; attendance and return rates for first and repeat screens may be based on empirical results (e.g., from Nijmegen, Utrecht, Two-County, BCCDP)	Varied; analysis of semi-annual, annual, biennial and triennial. Appears to have assumed 100% in screened group.	Annual, biennial Apparently assumed 100% in screened group.	Annual and t Participation as proportio receiving sc intervention period of 10
Preclinical Duration/ Lead Time	Preclinical duration as: prevalence /incidence from above studies; lead time is time from screen- to clinically-detected.	Indirectly modelled by including and varying parameters for tumour growth rate (assume geometric mean of .690) and tumour diameter thresholds. Do not correct for lead time.	Modelled using rate of initiation function h(y), and rate of tumour growth x (t;β).	Not consider variation of t distribution.
Screen Findings (+ve, -ve)	Provides for false +ve result, and clinically detected cancers (i.e., false -ve).	• Spontaneous and interval cancers. • Does not include mammo findings	??	Specify prop +ve screens Can specify and mammo
Abnormals, Biopsy	Utilities for QALYS, and costs, are determined for diagnostic phase, which includes assessment and biopsy.	Not considered.	??	Abnormals r considered, of Dx work-u
Breast Cancer Incidence	Uses rates in population prior to screening, and for screen-detected cancers. Provides for interval cancers.	Validated on the Swedish population.	Incidence rate of 1st report = i(a), which is varied to produce actual range of size of tumour (x) based on South Australian women.	Defined by cancer incid population c be SEER or
Cancer Stage	Uses T stage: • Tis, in situ; • T1a, 0-5mm; • T1b, 6-10mm; • T1c, 11-20mm; and • T2+, 21+ mm.	Parameters based on tumour diameter in mm: <10mm, 11-14mm, 15-19mm, 20-29mm, 30-49mm; and 50+mm.		Stage distn data, adjust mortality re screening t use empiric shifts from

CAN*TROL	Decision Analysis: ages 65+, Co-morbidity	Decision Analysis: ages 40-79, Strategies	Decision Analysis: ages 65-79, BMD
Annual and biennial. Participation may be defined as proportion of population receiving screening intervention over a phase-in period of 10 yrs.	Assumed one screen at t = 0. Compares 100% screened with not screened.	Annual and biennial. Compares 100% screened and unscreened.	Biennial. Assumes 100% screened.
Not considered, except in variation of the stage distribution.	Not considered, except in variation of stage distribution (below).	Not considered.	Not considered.
Specify proportion of false +ve screens. Can specify results for CBE and mammography	Uses 75% sensitivity and 90% specificity for mammo. False -ve have different stage dist; considers some false +ve.	Considers true +ve, false +ve, true -ve and false -ve (interval cancers).	Not considered.
Abnormals not specifically considered, except as costs of Dx work-up.	Abnormal screens result in biopsy leading to cancer Dx or not. Does not consider impact of unnecessary (ie, -ve) biopsies.	Includes follow-up for some +ve not immediately biopsied, does not consider adverse impacts of excess biopsies.	Assumes annual rate of 2.0% for abnormal screening mammography for women aged 60-79.
Defined by user, based on cancer incidence rates for population of interest, could be SEER or other rates.	SEER data for unscreened (late) Dx; BCDDP and Swedish trial for screen-detected; and Swedish trials for interval cancer rates.	Uses SEER data for baseline incidence; interval cancers based on data from Tabar (Swedish 2-County)	SEER rates
Stage distn based on SEER data, adjusted to reproduce mortality reductions from screening trials (does not use empirical data on stage shifts from screening trials).	Stage distn varied for screen-detected, interval and late Dx cancers using SEER for unscreened (late Dx), BCDDP and Swedish trials for screen-detected, and BCDDP for interval.	Not considered.	DCIS is only 'stage'; screen-detected rates from National Breast and Cervical Cancer Early Detection Program; pre-screen rates from SEER.

Model >>	MISCAN	Model for Breast Screening:MBS	Mathematical Model	CAN
Survival: Breast Cancer	Fraction cured in each clinical stage (ie, size) based on Swedish Cancer Registry.	Survival calculated as a function of tumour diameter and constants adjusted to Tabar's data for screened group; also calculates age-dependent survival.	Modelled mathematically as a hazard function: k (a,x) varies by age and size based on 1347 breast cancer cases.	Age- and st survival rate SEER data, screening ef of 32% at 5y 10y.
Other Causes	Life tables for all causes except breast cancer.	Lifetime expectancy.	Uses historical life tables.	Us s cause- mc tality rat causes.
Follow-up	To end of life (100 yrs) for all women in cohort.	To end of life (age 99).	Seems to follow to end of life.	To end of m up o 40 yea 19 2 screer foll wed to 2
Health and Cost Outcomes	• Deaths prevented. • Life-yrs gained. • Life-years in lead time. • Extra incidence. • QALYs gained.	• Life-years lost for spontaneous cancer; gained from screen-detected; & lost from radiation induced. • # of tumours by spontaneous, screen-detected, extra, radiation-induced, interval, etc.	• Lead time. • % detected at screening. • % increase in 10 yr survival. • % increase in life expectancy.	• D eaths av • C hanges i s reening c hange fo • F erson-ye • C osts per • C uality of • N mber o prevalenc

CAN*TROL	Decision Analysis: ages 65+, Co-morbidity	Decision Analysis: ages 40-79, Strategies	Decision Analysis: ages 65-79, BMD
Age- and stage- specific survival rates based on SEER data, adjusted for screening effect from RCTs of 32% at 5y and 28% at 10y.	Age-, race- and stage-specific survival due to breast cancer based on SEER rates; assumes same survival for screened and unscreened.	Based on mortality reductions expected from screening trials, adjusted by screening interval; interval cancers considered separately.	Uses SEER 10y mortality rates for 1984-89 for DCIS and 10y from 1978-83 for breast cancers for no screening; adjusts mortality rates by 27% reduction found in screening trials
Uses cause-deleted mortality rates for other causes.	Uses US life expectancy for women with no co-morbidity; and Framingham study mortality rates for women with hypertension and congestive heart disease.	Calculates age-specific mortality rates by subtracting the probability of breast cancer deaths from that for all deaths.	Uses life tables for overall mortality.
To end of modelling period, up to 40 years; ie, Brown 1992 screened 1990-2010; followed to 2020.	Apparently to end of life (as main outcome measure is life expectancy).	Only follow-up to end of screening (ignores benefits after screening period).	Follow-up for 5 years after end of screening (ie, to age 85).
• Deaths avoided. • Changes in cost of screening and Tx, and net change for each year. • Person-years of life saved. • Costs per death saved. • Quality of life. • Number of new cases and prevalence.	• Marginal savings in life expectancy (MSLE). • Determines marginal savings in QALY for: - for long term effects. and - both long and short term effects. • Determines cost per year of life saved CLYS).	• Life-years gained (as difference between screened and observed groups). • Main outcome is marginal cost per year of life saved (MCLYS) for screened vs not screened.	Calculates health states at age 85 based on cohort of 10,000 women, as: alive, with no in situ or invasive cancer; or dead from other causes (no br ca) or from other causes and DCIS or invasive br ca. Measures costs for screening, follow-up, Tx, and life-years saved.

Model >>	MISCAN	Model for Breast Screening:MBS	Mathematical Model	CA
Methods to Estimate Costs, including utilities and discount rate	Cost estimates based on actual costs of some procedures, tariffs for medical procedures, and a chart review of 68 women with advanced disease. Utilities: • screening attendance: 0.994; • post Tx, disease-free: >0.9; • various Tx (2mo to 2yr): 0.8; • terminal illness (1mo): 0.3 Discount rate: 3%	Does not consider costs.	Does not consider costs.	Eddy (1989 for: • screening mammo (• abnormal Dx worku $500); • Tx by sta • terminal c • dying of c Vary cost c $55 to $11 Utilities: nc Discount ra life-years:
Strengths and Limitations	Model used and validated in many populations. Analysis limited by: • assumptions re sojourn time, attendance rates; • limited data on improvement of prognosis (ie, mortality reductions, etc) *Note:* The increase, if any, in sojourn time for women aged 70+ is a critical factor. Cost analysis limited by: • simplified data.	• Does not consider costs. • Most work done on Swedish population. • Good presentation of results by 5-yr age groups, showing lots of numeric values. • Very systematic approach. • Unique in calculating radiation risk.	• Gives very detailed description of all parameters and functions used, and validates incidence results vs the SA population. • Model is provisional and results are illustrative only. • Limited by small population. • Age group 70+ is not main focus of model. • Looks at *starting* screening at each age, thus at 70+ so the lead time becomes an issue. • Precision of life expectancy changes, and population (at 100,000) may be too small to be measured for age 70+.	CAN*TRO algorithm c for: • changes caused b screenin unscreer • estimate Model sen assessed incidence distributio specific su Users can input data. Note: Reli good estir shift and s specific su

CAN*TROL	Decision Analysis: ages 65+, Co-morbidity	Decision Analysis: ages 40-79, Strategies	Decision Analysis: ages 65-79, BMD
Eddy (1989) Includes costs for: • screening program and mammo ($68); • abnormal follow-up, and Dx workups (biopsy, $500); • Tx by stage ($8-12K); • terminal care ($24K); • dying of other causes. Vary cost of mammo from $55 to $115. Utilities: not mentioned Discount rate for costs and life-years: 5% Brown (1992).	Estimates marginal costs for: • screening ($65 + $7); • abnormal work-up, including biopsy ($687) using Medicare charges. Utilities: *short-term, subtract from life expectancy (l.e).* • false +ve: 0.1 for 30 days • false -ve: 0.1 for 5 days; *long-term, muliply by l.e.* • local: 0.9 (0.8-1.0) • regional: 0.8 (0.7-0.9) • distant: 0.5 (0.4-0.6) Discount rate: 0%	Estimates costs for: • mammo ($84) • 4 types of biopsy procedure ($150-$2800) • definitive Tx ($6.1K) (One cost only, no mention of stage.) Utilities: no mention. NB: Tx data are not recent. Discount rate: • primary analysis: 0% • secondary analysis: 5%	Estimates in 1998 $US for: • mammo ($116); • abnormal results ($379) • Tx: screened, $37K; in unscreened, $38K; & in DCIS, $26K Utilities (from Australia): • main model, assume 1; • for sensitivity analysis, after breast cancer Tx, 0.8, and after • metastatic disease, 0.3 Discount rate: • baseline: 3%; • sensitivity: 0-15%
CAN*TROL screening algorithm does not provide for: • changes in incidence rate caused by introduction of screening program in unscreened population; • estimates of lead time. Model sensitivity is assessed by adjusting the incidence rates, the stage distribution, and the stage-specific survival rate. Users can manipulate the input data. <u>Note</u>: Relies heavily on good estimates of stage shift and stage- and age-specific survival rates.	• Authors note that as more primary data become available on the natural history of breast cancer and screening effects in older women, then more precise estimates of screening benefit can be made. • Only allows for one screening session; does not consider first or repeat screens. • Reports limited cost information. Use of one-point-in-time view - seen as useful for individual decision-making.	Criticisms in several letters to editor include lack of: • consideration of unnecessary surgery, • sensitivity analysis of mortality reductions, • consideration of impacts after end of screening period, and • validation vs actual data. Also, does not seem to make any adjustment for stage, or to differentiate screened vs non-screened by cost of advanced Rx.	Strengths: • considers costs of DCIS; • sensitivity checked for utilities of disease states, discount rate, mortality reduction, 10-yr breast cancer mortality, BMD costs and % breast cancer reduction with low BMD. Limitations: • very sensitive to utilities; • uses $ ranges from Laupacis in 1990 $Cdn to apply to 1998 $US; • states cut-off as $US50K, instead of $Cdn20K (but Weinstein uses $50K cutoff).

Model >>	MISCAN	Model for Breast Screening:MBS	Mathematical Model	CA
Conclusions	• Balance between +ve and -ve effects of screening remains up to at least 80 yrs with long lead time and continues at older ages with shorter lead time. • Cumulative life-yrs gained show little increase after age 80; life-yrs in lead time increase at all ages; and extra incidence increases more rapidly when assume longer lead time. • QALYs (at 5% discount) increase to about age 80, then level off (assuming short lead time) and drop (assuming long lead time). With $US100 mammograms, CLYS for continuing to screen from age 65-82: • no screen: $8.3K; • 3-yr interval: $9.6K; • 2-yr interval: $19.7K; • 1-yr interval: $38.6K.	For 70+ find: • life-yrs gained by screening is 3052 for 70-74 age group, and 1476 for 75-79, then drops rapidly. For 80+ find: • # of excess screen-detected tumours is at least 40% => much Tx of little benefit to women; For age 76-80 only marginal improvement in mortality from increased screening frequency. Recommend age 35-74 as screening range, based on life-time gained, but 40-80 based on reduced # of fatal tumours.	For 70+ find: • longer lead times; • good % detected at screening; • 5-6% increase in 10-yr survival after reporting disease (due to lead time?). • no increase in life expectancy (i.e. 0.1% at age 70+ compared to 1-2% increase if screening started at age 60 or 50).	Eddy (1989) • If 30% of screen ea would hav deaths in 719 fewer of 22,000 • Cost of sa screen ag Brown (199 Screening i effective in group; with screening,a mammogra saved (CLY • for 65-69, • for 75-79, • for 80-84, if mammo i $115, CLY • for 65-69 • for 80-84

CAN*TROL	Decision Analysis: ages 65+, Co-morbidity	Decision Analysis: ages 40-79, Strategies	Decision Analysis: ages 65-79, BMD
Eddy (1989): • If 30% of women accept screen each year then would have 229 fewer deaths in first year, and 719 fewer after 15 yrs, out of 22,000 deaths for 65+. • Cost of saving one life if screen age 65-75 is $26K. Brown (1992): Screening is most cost-effective in 65-69 age group; with biennial screening,and $55 mammograms, cost per life saved (CLYS) is: • for 65-69, $26K; • for 75-79, $30K; • for 80-84, $38K. if mammo increases to $115, CLYS rises : • for 65-69 to $39K; and • for 80-84 to $62K	• Marginal savings in life expectancy (MSLE) range from 2.2 days (age 65-69) to 0.8 days (85+) for long-term QoL in average health to 1.2 to 0.6 days for worst health. • For long-term + short term QoL, MSLE for normal health dropped from 1.44 to 0.05 days, and worst health from 0.43 to -0.15 days between ages 65-69 and 85+. Savings (as CYLS) for women in average health: • age 65-69: $23K; • age 70-74: $28K; • age 85+: $73K.	• Most cost-effective approach to mammography is biennial screening for age 50 to 79, with a MCLYS of $16,000; • Annual screen of 40-49 is not unreasonable, cost-wise, with MCLYS of $27,100 for ages 40-79, based on 23% mortality reduction (for 40-49), rising to $33,600, if assuming 4% reduction; • Analysis for 70-79 is peripheral to study, and based on 19% mortality reduction; results are not reported separately Cost summary; MYCLYS: • 50-79, biennial; $16K; • 40-79, 23% mortality drop in 40s; $27K; • 40-79, 4% mortality drop in 40s; $33.6K.	Continuing screening mammography after age 69 up to age 79 would: • prevent 10.8 deaths; • add 2.4 days to life expectancy (most benefits accrue to women in the top 3 quartiles of BMD); • be moderately cost-effective for those with high BMD, with cost per life saved as $66,773; and • be more costly ($117,689) for those with low BMD (lowest quartile). Women's preferences are important

REFERENCES

1. van Oortmarssen GJ, Habbema JD, van der Maas PJ *et al.* A model for breast cancer screening. Cancer 1990; 66(7):1601-12.

2. de Haes JC, de Koning HJ, van Oortmarssen GJ, van Agt HM, de Bruyn AE, van Der Maas PJ. The impact of a breast cancer screening programme on quality-adjusted life-years. Int J Cancer 1991; 49(4):538-44.

3. Boer R, de Koning HJ, van Oortmarssen GJ, van der Maas PJ. In search of the best upper age limit for breast cancer screening. Eur J Cancer 1995; 31A(12):2040-3.

4. Boer R, de Koning H, Threlfall A *et al.* Cost effectiveness of shortening screening interval or extending age range of NHS breast screening programme: computer simulation study. BMJ 1998; 317(7155):376-9.

5. Boer R, de Koning HJ, van der Maas PJ. A longer breast carcinoma screening interval for women age older than 65 years? Cancer 1999; 86(8):1506-10.

6. Jansen JT, Zoetelief J. MBS: a model for risk benefit analysis of breast cancer screening. Br J Radiol 1995; 68(806):141-9.

7. Jansen JT, Zoetelief J. Optimisation of mammographic breast cancer screening using a computer simulation model. Eur J Radiol 1997; 24(2):137-44.

8. Jansen JT, Zoetelief J. Assessment of lifetime gained as a result of mammographic breast cancer screening using a computer model. Br J Radiol 1997; 70(834):619-28.

9. O'Neill TJ, Tallis GM, Leppard P. A review of the technical features of breast cancer screening illustrated by a specific model using South Australian cancer registry data. Stat Methods Med Res 1995; 4(1):55-72.

10. Eddy DM. A computer-based model for designing cancer control strategies. In: Greenwald P, Sondik EJ, editors. NCI Monographs: Cancer Control Objectives for the Nation 1985-2000. Vol. 2. Bethesda: National Cancer Institute, 1986: 75-82.

11. Levin DL, Gail MH, Kessler LG, Eddy DM. A model for projecting cancer incidence and mortality in the presence of prevention, screening, and treatment programs. In: Greenwald P, Sondik EJ, editors. NCI Monographs: Cancer Control Objectives for the Nation: 1985-2000. Vol. 2. Bethesda: National Cancer Institute, 1986: 83-93.

12. Eddy DM. Screening for breast cancer. Ann Intern Med 1989; 111(5):389-99.

13. Brown ML. Economic considerations in breast cancer screening of older women. J Gerontol 1992; 47 Spec No:51-8.

14. Mandelblatt J, Wheat ME, Monane M, Moshief RD, Hollenberg JP, Tang J. Breast cancer screening for elderly women with and without comorbid conditions. A decision analysis model. Ann Intern Med 1992; 116(9):722-30.

15. Lindfors KK, Rosenquist CJ. Needle core biopsy guided with mammography: a study of cost-effectiveness. Radiology 1994; 190(1):217-22.

16. Lindfors KK, Rosenquist CJ. The cost-effectiveness of mammographic screening strategies [see comments] [published erratum appears in JAMA 1996 Jan 10;275(2):112]. JAMA 1995; 274(11):881-4.

17. Rosenquist CJ, Lindfors KK. Screening mammography beginning at age 40 years: a reappraisal of cost-effectiveness. Cancer 1998; 82(11):2235-40.

18. Kerlikowske K, Salzmann P, Phillips KA, Cauley JA, Cummings SR. Continuing screening mammography in women aged 70 to 79 years: impact on life expectancy and cost-effectiveness. JAMA 1999; 282(22):2156-63.

APPENDIX C:

DATA SOURCES, INCLUDING METHODS USED TO DEVELOP BREAST CANCER STAGING DATA FOR ANALYSIS

C.1 Data Sources: Overview

Data from sources described in this Appendix were analyzed in Chapter 4 to provide empirical data to assess Criteria 2 to 4. These data sources were further used in Chapter 5 to develop input data for, and to validate results from, the MISCAN model used to produce information to assess Criterion 5. Section C2 describes the sources used to obtain breast cancer stage data, while Section C3 compares the 3rd and 4th editions of the TNM staging classification system, and details the methods used to develop T stage categories for analysis.

C.1.1 Canadian Cancer Registries

National and provincial cancer registries provide excellent coverage of the Canadian population.[1] The patient-oriented Canadian Cancer Registry (CCR) (1992-) and its event-oriented predecessor, the National Cancer Incidence Reporting System (NCIRS) (1969-1991) contain cancer incidence data as reported annually by provincial and territorial cancer registries to the Health Statistics Division at Statistics Canada.[2,3] Statistics Canada annually provides an extract of this data file to Health Canada, where it is loaded into the ORIUS system to facilitate analysis. The NCIRS and CCR contain demographic, diagnostic and residence data for each invasive tumour diagnosed in residents of each province and territory; all provinces and territories except Ontario also provide information on *in situ* cancers. Completeness of registration for invasive cancers is considered to be about 95% or more from 1983 onwards.

Data were originally coded in ICDA8 until 1977, then in ICD9 until 1991 and subsequently in ICDO-2.[3] Subsequently, data for all years were converted to ICD9 at Statistics Canada. For this project, data on female breast cancers were extracted from the

ORIUS system at Health Canada, using ICD9 code 174._ , while *in situ* breast cancers were selected using ICD9 code 233.0 and Sex = Female. (174._ means any 4th digit suffix for code 174). Stage of cancer is not available at the national level. Several provincial registries, however, have coded TNM stage of breast cancers diagnosed in part or all of their population as follows, including Northern Alberta (defined as north of the 52nd parallel of latitude) from 1974 to1988, and Saskatchewan from 1971 to 1997; data from these registries were provided to Health Canada for use in this project. *(See* Section C.2*)*

C.1.2 Vital Statistics Mortality Data

Statistics Canada maintains mortality data for Canada as compiled from the vital statistics offices in each province and territory. Underlying cause of death is coded using the International Classification of Diseases, 9th revision. Data for breast cancer and all causes of death for this study were selected using ICD9 code 174._ and 000.0 through 999.9, respectively, from a version of this data set supplied annually to Health Canada.

C.1.3 Population Data

Statistics Canada conducts a census of the Canadian population in Canada every five years. Post-censal population estimates for Canada by single year of age up to age 90+ are produced annually based on census counts updated for births, deaths, immigration and emigration. Inter-censal population estimates have been revised back to 1971 to account for census undercount and non-residents.[4] For this project, special tabulations were prepared at Statistics Canada to provide estimates by single year of age from 90 to 100+ yrs. The percentage distribution by single year of age was determined using unadjusted census counts, and then applied to the adjusted total 90+ population estimate in each census year. Populations for single year of age from 90 to 100+ for inter-censal years were then calculated by interpolation.

C.1.4 Canadian Breast Cancer Screening Database

The Canadian Breast Cancer Screening Database (CBCSD) contains individual micro-data records on every screening visit for each woman who has participated in an organized breast screening program in Canada.[5] British Columbia was the first province to institute a provincial breast screening program in 1988, followed closely by Alberta, Saskatchewan, Yukon, Ontario and Nova Scotia. All 10 provinces and the Yukon territory have implemented organized breast screening programs. Breast cancer screening at all organized programs includes a two-view screening mammogram. Five provinces, Manitoba, Ontario, Nova Scotia, Prince Edward Island and Newfoundland also provide a clinical breast examination (CBE).[5] The target population is defined as asymptomatic women between the ages of 50 and 69, with no prior diagnosis of breast cancer. All programs now accept women aged 70 and over, although policies vary regarding the extent to which women are actively recruited or recalled (see Table C.1).

Established in 1993, the CBCSD is maintained by the Centre for Chronic Disease Prevention and Control (CCDPC) at Health Canada, based on data provided by provincial programs. As of the 1996 reporting year, the CBCSD contains data on over 1,300,000 screening visits, of which 310,000 (including 45,000 in women aged 70+) were conducted in 1996. For each woman screened, organized programs collect information on risk factors for breast cancer, screening history, screening results, referral status (referred for diagnostic follow-up or not), diagnostic tests, and the final diagnosis for all women referred for diagnostic follow-up. Cancer diagnoses are usually confirmed in consultation with the provincial cancer registry.

Table C.1 Breast cancer screening programs in Canada: practices regarding women aged 70+ in 1999

Program	Start Year	Program Practices For Women Aged 70+		Clinical Breast Exam	Screening Frequency
British Columbia*	1988	70-74 75-79 80+	Actively recruit (since 1997) Recall, do not actively recruit Accept with physician referral	No	Biennial (Annual until mid-1997)
Yukon	1990	70+	Accept, do not recall	No	Biennial
Alberta	1990	70-74 75+	Recall, do not actively recruit Accept, do not recall	No	Biennial
Saskatchewan	1990	70+	Recall, do not actively recruit	No	Biennial
Manitoba	1990	70+	Accept, do not recall	Nurse	N/A
Ontario	1991	70+	Accept, do not recall	Technologist	Biennial
Northwest Territories	1994	70+	Accept, do not actively recruit	No	Biennial
New Brunswick	1995	70+	Accept with physician referral	No	Biennial
Nova Scotia	1995	70+	Recall, do not actively recruit	Nurse or technologist	Biennial
Newfoundland	1996	70-74	Recall, do not actively recruit	Nurse	Biennial
Quebec	1998	70+	Accept with physician referral	No	Biennial
Prince Edward Island	1998	70-74	Recall, do not actively recruit	Technologist	Biennial

Source: adapted from *Organized Breast Cancer Screening Programs in Canada*[6]

C.1.5 Mammography Billings/Fee-for-Service Data

Not all screening occurs in organized programs. Non-program mammography can be tracked through the fee-for-service billings system to provide a comprehensive picture of mammography utilization in Canada.[7] Data on the annual number of mammograms performed on women by five-year age group are collected from the provincial/territorial departments of health responsible for operating the physician billings data bases that track the fees paid for various health services. This data source includes both diagnostic and screening

mammograms, of which about 80% are for screening purposes. Between 1981 and 1994 the annual number of mammograms performed in Canada increased from less than 200,000 to more than 1.4 million, with almost 15 % of all mammograms in 1994 performed in screening programs.[7] Due to the high percentage of screening mammograms that occur outside of screening programs, these billings data are needed to develop a comprehensive picture of mammography utilization in Canada.

C.1.6 National Population Health Surveys

Longitudinal and cross-sectional national health surveys complement the administrative data by providing an independent assessment of mammography utilization together with information on socio-demographic characteristics of women. Information on mammography utilization was first collected by the 1990 Health Promotion Survey.[3] In 1994/95, the biennial National Population Health Survey (NPHS) was established at Statistics Canada to collect information about the health of the Canadian population from both household and institutional residents in all provinces and territories. For the first three cycles, the NPHS had both a cross-sectional and longitudinal component. The 1994/95 NPHS collected detailed health information from 17,626 household respondents, including questions on frequency of, and reasons for mammography utilization posed to women aged 35+. The larger 1996/97 survey collected information on about 66,000 respondents of which 22,000 were women aged 35 and over.

C.2 Sources of Breast Cancer Staging Data

C.2.1 Overview

Breast cancer stage distributions can be measured over time to assess whether the expected shift to lower stage disease has occurred, which is a necessary first step for mortality reductions to ensue. Comparing stage distributions by age group also provides

insight into differences in biology of breast cancer that could affect mortality reductions for older women. Finally, T stage distributions are needed to validate input and outputs from the MISCAN model. Data for breast cancer T stage were developed for the situation *before screening* using breast cancers recorded on the population-based Northern Alberta and Saskatchewan cancer registries. Breast cancer T stage *in the presence of screening* was determined from the Canadian Breast Cancer Screening Data base.

Comparisons in this thesis were based on T stage, which was derived and tabulated for each Canadian data sets for time periods before and after screening implementation, to compare the stage distribution for women aged 50-69 to those aged 70-79. This was needed to determine if the biological characteristics of breast cancer were similar for the two age groups and to validate model output. Finally, T stage distributions at first screen and rescreen from the Canadian Breast Cancer Screening Database were tabulated to compare with published results from other geographic areas world-wide, where data were modeled using MISCAN. Detailed comparisons between the 3rd and 4th editions of the TNM staging systems and the method used to derive T stage are given in Section C.3. Details of the source data are described below.

C.2.2 Northern Alberta Breast Registry (1971-1988)

The Northern Alberta Breast Registry, created in 1971, initially used the 1968 version (or 2nd edition) of the TNM.[9] However, data were subsequently retrospectively restaged from the 1968 to the 1973 version (or 3rd edition) of the TNM.[10] Staging information was coded by nurses from patient charts based on information recorded at the initial clinical assessment of the patient by oncologists at the breast assessment clinic conducted at least weekly at the Cross Cancer Institute. Pathologic tumour size was recorded to the nearest millimetre. From 1971-1973, 643 pathologically confirmed cases were registered, of which clinical (ie, pre-operative) TNM stage was available for 622 cases; the trend was for younger

women to have less advanced disease. When data were transformed into the 1973 version, the stage distribution was greatly altered, because tumour size was now emphasized, and Stage II now included disease without palpable lymph nodes. Thus, in the 1973 version, Stage II consisted of clinically less-advanced regional disease than that for 1968. A total of 3,132 records of invasive and in situ breast cancers diagnosed between 1971 and 1984 in women aged 50-79 years was used in this analysis. The extent of missing data varied across the time period.

C.2.3 Saskatchewan Cancer Registry (1975-1997)

Breast cancer stage was coded to the 4th revision of the TNM by registry staff based on retrospective and prospective chart review, with the latter starting in 1990. Analyses in this report were based on 6,369 records of invasive and *in situ* breast cancer diagnosed between 1975 and 1997 in women aged 50-79. About 40% of records have pathologic size recorded. The staging recorded is primarily pathological as most of the women are seen in the clinics subsequent to surgery. The staging information was coded by Health Record Technicians (HRTs) from the information in the cancer centre chart including the stage as reported by the oncologist. Any discrepancies between the oncologist's assigned staging and that determined by the HRTs were returned to the oncologist for final decision. The staging determined within the cancer registry is passed to the Screening Program for Breast Cancer for incorporation in their records.

C.2.4 Canadian Breast Cancer Screening Database (1990-1996)

Staging data are recorded in each screening program using the TNM, 4th revision. Coding is done in some provinces by the screening program staff based on information in the pathology reports supplemented by other sources. In others, stage is provided from the cancer registry. Analyses in this report were restricted to 7,602 screen-detected (as defined by each screening program) invasive and *in situ* breast cancers diagnosed in women aged 50-

79, and excluded interval and non-compliant cancers. The combination of reported T stage with pathologic size may be problematic for Alberta and Nova Scotia, where core biopsies are used. As this procedure removes part of the tumour, the pathologic size of the tumour may be smaller than at the initial (clinical) assessment.

C.3 Staging Classifications and Development of T Stage

C.3.1 Analysis of Systems to Record Cancer Stage

Staging information is recorded in these databases using the UICC TNM system (Union Internationale Contre le Cancer, Tumour, Nodes, Metastases). The TNM 3rd edition was used in Northern Alberta, and the TNM 4th edition, in Saskatchewan (all years), and in the Canadian Breast Cancer Screening Database. A comparative table for the summary stage for the two systems is presented in Table C.2.

Table C.2 Comparison of Summary Stage from TNM 3rd and 4th edition

Summary Stage	TNM, 3rd edition (Alberta)	4th edition (same as AJCC) (SK, CBCSD, SEER)
0 (in situ)	not stated, presume same as 4th	Tis N0 M0
I	T1 (a or b) N0, N1a* M0	T1 N0 M0
II	T0, T1 N1b M0 T2 N0, N1a M0 T2 N1b M0	**IIA** T0,T1 N1 M0 T2 N0 M0 **IIB** T2 N1 M0 T3 N0 M0
IIIA	T3 N0, N1 ?M0 T1, T2, T3 N2, N3** M0	T3 N1, N2 M0 T0, T1, T2 N2 M0
IIIB	T1, T2, T3 N3 M0 T4 N1, N2, N3* M0	T0, T1, T2, T3 N3** M0 T4 N1,N2,N3 M0
IV	Any T Any N M1	Any T Any N M1

Notes: * shading indicates differences in assignment of detailed T and N codes to Summary Stage
** different definitions used for N3 in the two systems.

The TNM system provides for a series of codes based on Tumour size, Nodal involvement and presence of Metastases. The elements of the TNM for the 3rd and 4th revision are shown in Table C.3. The TNM Summary stage (0, I, II, III, IV) is derived from known values of T, N and M; cases with the unknown values of Tx, Nx, and Mx can not be assigned a Summary Stage using this system in its pure form. The TNM is a dual system with a (pre-treatment) Clinical classification (cTNM) and a (post-surgical histopathological) Pathological classification (pTNM). Both classifications are retained unaltered in the patient's record.* The former is used for the choice of treatment; the latter is used for the estimation of prognosis and the possible selection of adjuvant therapy.

The main classification for Stage T1, T2, T3 and T4 is the same when comparing the 3rd and 4th editions; however, the subcategories differ. In the 3rd edition, the 'a' or 'b' suffix refers to (a) no extension, or (b) extension to chest wall. By contrast, in the 4th edition, T1a,b,c refer to the size of the tumour, T2 and T3 have no sub-categories, and T4 has a further sub-category for inflammatory carcinoma. Thus T1a and T1b in the 3rd revision have a different meaning than in the 4th revision.

Differences also occur for N stage. The N1a category appears in the 3rd and 4th editions as a P (pathology) stage to indicate micro-metastases[11] but only in the 3rd edition does it appear as a C (clinical) stage. For the 4th edition, Summary Stage IIA includes all N1 (both a and b) with T1, yet in the 3rd edition, N1a is placed in Stage I, when these occur with T1, despite a note to the classification which states that N1a has similar prognosis as N0.

As for M stage, metastases to *ipsilateral supraclavicular* nodes are staged as N3 in the 3rd edition, (together with *infraclavicular* nodes, as well as edema of the arm, and **Table**

---

* Union Internationale contre le cancer, April 2000, www3.uicc.org/programmes/detection/ tnm/tnmfaqs

C.3: Comparison of T, N, and M components for the 3rd and 4th revision of the TNM pre-treatment clinical classification

TNM	3rd	4th
T is assessed using:	***clinical exam; mammography admissible but not mandatory; mammographic measurement takes precedence***	***physical exam and imaging (eg, mammography)***
Tx	Primary tumour cannot be assessed	
T0	No evidence of primary tumour	
Tis	Pre-invasive carcinoma (carcinoma in situ), non-infiltrating intraductal carcinoma, or Paget's disease of the nipple, with no demonstrable tumour	
T1	Tumour of 2 cm or less in greatest dimension	
T1a	with no fixation to the underlying pectoral fascia and/or muscle	0.5 cm or less in greatest dimension
T1b	with fixation to underlying pectoral fascia and/or muscle	more than 0.5 cm, but not more than 1 cm in greatest dimension
T1c	--	more than 1 cm but not more than 2 cm in greatest dimension
T2	Tumour more than 2 cm, but not more than 5 cm in greatest dimension	
T2a	with no fixation...as for T1a	--
T2b	with fixation.... as for T1b	--
T3	Tumour more than 5 cm in greatest dimension	
T3a	with no fixation....as for T1a	--
T3b	with fixation.... as for T1b	--
T4	Tumour of any size with direct extension to chest wall or skin	
T4a	with fixation to chest wall	Extension to chest wall
T4b	Oedema (including peau d'orange), *infiltration (3rd only)* or ulceration of the skin of the breast, or satellite skin nodules confined to the same breast	
T4c	Both of the above (T4a and T4b)	
T4d	--	Inflammatory carcinoma

Staging System	3rd	4th
N assessed:	***clinical exam***	***physical examination and imaging***
Nx	The minimum requirement to assess the regional lymph nodes can not be met	Regional lymph nodes cannot be assessed (e.g., previously removed)
N0	No palpable homolateral **correct** axillary lymph nodes	No regional lymph node metastases
N1	Movable homolateral axillary lymph nodes	Metastasis to movable ipsilateral axillary node(s) **see picture**
N1a	nodes not considered to have growth	
N1b	nodes considered to have growth	
N2	Homolateral axillary lymph nodes fixed to one another or to other structures and considered to contain growth	Metastasis to movable ipsilateral axillary nodes
N3	Homolateral supraclavicular or infraclavicular lymph nodes considered to contain growth or oedoma of the arm	Metastasis to ipsilateral internal mammary lymph nodes fixed to one another or to other structures
M assessed	***clinical examination and radiography***	***physical examination and imaging***
Mx	The minimum requirements to assess the presence of distant metastases can not be met.	Presence of distant metastases can not be assessed
M0	No evidence of distant metastases	No distant metastases
M1	Evidence of distant metastases	Distant metastases (includes supra-clavicular, cervical or contralateral mammary lymph nodes.)

invasion of homolateral internal mammary lymph nodes). However, in the 4th edition, ipsilateral supraclavicular nodes are staged as M1, distant metastases, while N3 refers only to ipsilateral internal mammary lymph nodes. This makes it impossible to separate N3 between the two staging systems to convert from one to the other regarding summary stage.

Summary Stage differs somewhat between the 3rd and 4th editions as noted above, and it is therefore not possible to convert completely between the 3rd and 4th revisions. However,

if analyses are restricted to Stage 0, I, II, III, and IV, then there are only a few differences, making comparisons possible. First, Stage II is coded as a single entity in the 3rd, but is divided into Stage IIA and IIB in the 4th. Second, N1a is included as Stage I in 3rd but Stage IIA in 4th. Third, some entities assigned to N3 in Stage IIIB in the 3rd edition are classified in the 4th as M1 and therefore appear in Stage IV. Fourth, T3 N0 is classified to Stage IIIA in the 3rd and Stage IIB in 4th, and fifth, T1, T2, T3, N3 are assigned to Stage IIIA in the 3rd but to Stage IIIB in the 4th. Finally T0, no evidence of primary tumour, is handled inconsistently in the two systems, when assigning summary stage.

C.3.2 Development of T Stage Data

In addition to recorded TNM, pathology size of the tumour is often recorded as measured by the pathologist based on the excised tumour. For analyses in this thesis, pathology size and recorded TNM stage were used together to reduce the percentage of cases with missing T stage. Comparisons were made for Canadian data using the T component of the stage distribution (important for the MISCAN model) and to compare the stage distributions between the 50-69 and 70-79 age groups.

T stage data were derived from the Canadian data sources to explore stage shifts and to validate the MISCAN model input and output for the Canadian population. MISCAN output is presented in terms of T stage, as follows:

Tis(DCIS): ductal carcinoma in situ (99% of breast *in situ* cancers are DCIS);

T1a: invasive cancer, 0-5 mm;

T1b: invasive cancer, 6-10 mm;

T1c: invasive cancer, 11-20mm; and

T2+*: invasive cancer, 21+ mm and no metastases or extension to chest wall)

(*T2 = 21-50 mm, T3 = 51+ mm, T4 = tumour of any size with direct extension to chest wall or skin)

Using data in various formats from Northern Alberta, Saskatchewan, and the Canadian Breast Cancer Screening data bases, the pathologic tumour size and recorded TNM stage were combined to minimize the amount of missing information on T stage. First, T stage was recoded as T0, T1, T2, T3, T4, Tis and Tx, and including T1a, T1b, T1c, and T1x where available. Pathological size in mm was also recoded as 0-5, 6-10, 11-20, 21-50, 50+, and unknown, to correspond to T sizes. Next, RT stage (recoded T stage) was created using the following logic:

If T stage = Tis or T4, then RT stage = *in situ* or T4 respectively

For the remaining T stages,

1) First use the pathology size to assign T stage as T1a, T1b, T1c, T2 or T3. (this is the only way to identify T1a, T1b and T1c in Alberta;

2) Next use recorded T stage to assign RT stage as T1x, T2, or T3, where pathology size is missing. Use T1x to retain as much information as possible; and

3) Finally, assign Tx (unknown) where path size is unknown and T = T0 or Tx.

REFERENCES

1. Chen VS, Wu C, Andrews PA, (eds.). Cancer in North American 1991-1995. Volume One: Incidence. Sacramento, CA: North American Association of Central Cancer Registries, 1999.

2. Band PR, Gaudette LA, Hill GB *et al.* The Making of the Canadian Cancer Registry: Cancer Incidence in Canada and its Regions, 1969 to 1988. Health and Welfare Canada, Canadian Council of Cancer Registries, Statistics Canada. Ottawa: Minister of Supply and Services Canada, 1993.

3. Gaudette LA, Lee J. Cancer Incidence in Canada, 1969-1993. Statistics Canada cat. no. 82-566-XPB. Ottawa: Minister of Industry, 1997.

4. Gaudette LA, Silberberger C, Altmayer CA, Gao RN. Trends in breast cancer incidence and mortality. Health Rep 1996; 8(2):29-37(Eng); 31-40(Fre).

5. Health Canada. Organized Breast Cancer Screening Programs in Canada - 1996 Report. cat. no. H1-9/13-1999. Ottawa: Minister of Public Works and Government Services, 1999.

6. Health Canada. Organized Breast Cancer Screening Programs in Canada- 1997 and 1998 Report. cat. no. H1-9/13-1998. Ottawa. Minister of Public Works and Government Services Canada, 2001.

7. Gaudette LA, Altmayer CA, Nobrega KM, Lee J. Trends in mammography utilization, 1981 to 1994. Health Rep 1996; 8(3):17-27(Eng); 19-30(Fre).

8. O'Connor A. Women's cancer prevention practices. In Stephens T, Graham DF, eds. Canada's Health Promotion Survey. Ottawa: Minister of Supply and Services, 1993.

9. Burns PE, Kredentser J, Grace M, Hanson J. Breast cancer in northern Alberta: pilot study in computerized registration. Can Med Assoc J 1977; 116(10):1131-5.

10. Burns PE, Kredentser S, Grace M. Restaging of breast cancer in a northern Alberta registry. Can Med Assoc J 1976; 115(6):516-8.

11. Brower ST, Cunningham JD, Tartter PI. Staging and assessing prognosis, Chapter 17. in: Roses DF, editor. Breast Cancer. New York: Churchill Livingstone, 1999: 273-86.

APPENDIX D

Life table calculations for cumulative mortality of death before a given age, for all causes of death and all causes other than breast cancer death, Canada, females, 1995-1997

Age	All Deaths	Breast Cancer Deaths	All - Breast Cancer	Population	Rate, all causes, mx	other causes, mx(i)	qx all	qx(i) other	(expected deaths after age x) all causes	No. of "deaths" dx all causes	No. of "deaths" dx(i) other	deaths after age x causes	Cumulative Probability of Death before age x all causes	other causes	Age
(x)	1995-1997	1995-1997	1995-1997	1995-1997	1995-1997	1995-1997			((lx-1)-(dx-1)	(lx*qx)	lx*qx(i)	Wx(i)		Fx(i) =1-(Wx(i)/W0(i)	(x)
00-04	3,367	0	3,367	2,867,920	0.00117	0.00117	0.00585	0.00585	100,000.0	585.3	585.3	96,087.5	0.0000	0.0000	00-04
05-09	415	0	415	2,954,700	0.00014	0.00014	0.00070	0.00070	99,414.7	69.8	69.8	95,502.3	0.0059	0.0061	05-09
10-14	471	0	471	2,932,826	0.00016	0.00016	0.00080	0.00080	99,344.9	79.7	79.7	95,432.5	0.0066	0.0068	10-14
15-19	967	0	967	2,928,032	0.00033	0.00033	0.00165	0.00165	99,265.2	163.8	163.8	95,352.7	0.0073	0.0076	15-19
20-24	955	1	954	2,968,044	0.00032	0.00032	0.00161	0.00161	99,101.4	159.3	159.1	95,188.9	0.0090	0.0094	20-24
25-29	1,235	27	1,208	3,214,964	0.00038	0.00038	0.00192	0.00188	98,942.1	189.9	185.7	95,029.8	0.0106	0.0110	25-29
30-34	1,955	133	1,822	3,823,281	0.00051	0.00048	0.00255	0.00238	98,752.2	252.2	235.0	94,844.1	0.0125	0.0129	30-34
35-39	2,963	372	2,591	3,918,201	0.00076	0.00066	0.00377	0.00330	98,500.1	371.7	325.1	94,609.1	0.0150	0.0154	35-39
40-44	4,357	727	3,630	3,578,349	0.00122	0.00101	0.00607	0.00506	98,128.3	595.6	496.2	94,284.0	0.0187	0.0188	40-44
45-49	6,092	993	5,099	3,176,650	0.00192	0.00161	0.00954	0.00799	97,532.7	930.8	779.0	93,787.8	0.0247	0.0239	45-49
50-54	7,567	1,189	6,378	2,515,236	0.00301	0.00254	0.01493	0.01258	96,602.0	1,442.3	1,215.6	93,008.8	0.0340	0.0320	50-54
55-59	10,052	1,314	8,738	2,021,110	0.00497	0.00432	0.02456	0.02135	95,159.7	2,337.3	2,031.8	91,793.1	0.0484	0.0447	55-59
60	2,375	253	2,122	374,073	0.00635	0.00567	0.00633	0.00565	92,822.4	587.5	524.9	89,761.3	0.0718	0.0658	60
61	2,579	277	2,302	368,785	0.00699	0.00624	0.00697	0.00622	92,234.9	642.8	573.7	89,236.5	0.0777	0.0713	61
62	2,840	298	2,542	365,846	0.00776	0.00695	0.00773	0.00692	91,592.2	708.3	633.9	88,662.7	0.0841	0.0773	62
63	3,079	280	2,799	365,851	0.00842	0.00765	0.00838	0.00762	90,883.9	761.7	692.4	88,028.8	0.0912	0.0839	63
64	3,386	253	3,133	368,901	0.00918	0.00849	0.00914	0.00845	90,122.2	823.4	761.9	87,336.4	0.0988	0.0911	64
65	3,771	319	3,452	367,028	0.01027	0.00941	0.01022	0.00936	89,298.8	912.8	835.6	86,574.5	0.1070	0.0990	65
66	4,108	289	3,819	360,855	0.01138	0.01058	0.01132	0.01052	88,386.0	1,000.5	930.1	85,738.9	0.1161	0.1077	66
67	4,338	318	4,020	353,831	0.01226	0.01136	0.01219	0.01129	87,385.5	1,064.8	986.8	84,808.8	0.1261	0.1174	67
68	4,739	353	4,386	346,396	0.01368	0.01266	0.01359	0.01258	86,320.7	1,172.9	1,085.6	83,822.0	0.1368	0.1276	68
69	5,135	340	4,795	343,613	0.01494	0.01395	0.01483	0.01385	85,147.8	1,263.0	1,179.4	82,736.5	0.1485	0.1389	69
70	5,636	377	5,259	338,833	0.01663	0.01552	0.01650	0.01539	83,884.7	1,383.8	1,291.2	81,557.1	0.1612	0.1512	70
71	6,062	399	5,663	333,837	0.01816	0.01696	0.01800	0.01681	82,500.9	1,484.6	1,386.9	80,265.8	0.1750	0.1647	71
72	6,452	424	6,028	324,888	0.01986	0.01855	0.01966	0.01837	81,016.3	1,593.1	1,488.4	78,878.9	0.1898	0.1791	72
73	6,936	382	6,554	316,403	0.02192	0.02071	0.02168	0.02049	79,423.2	1,722.2	1,627.3	77,390.5	0.2058	0.1946	73
74	7,428	362	7,066	304,934	0.02436	0.02317	0.02407	0.02289	77,701.0	1,870.0	1,778.8	75,763.2	0.2230	0.2115	74
75	8,079	400	7,679	291,008	0.02776	0.02639	0.02738	0.02603	75,831.1	2,076.4	1,973.6	73,984.3	0.2417	0.2300	75
76	8,116	385	7,731	266,011	0.03051	0.02906	0.03005	0.02863	73,754.7	2,216.4	2,111.3	72,010.7	0.2625	0.2506	76
77	8,245	335	7,910	241,369	0.03416	0.03277	0.03359	0.03222	71,538.2	2,402.7	2,305.0	69,899.4	0.2846	0.2725	77
78	8,247	330	7,917	219,670	0.03754	0.03604	0.03685	0.03538	69,135.5	2,547.7	2,445.8	67,594.4	0.3086	0.2965	78
79	8,881	356	8,525	207,298	0.04284	0.04112	0.04194	0.04026	66,587.8	2,792.9	2,681.0	65,148.6	0.3341	0.3220	79

Age	All Deaths	Breast Cancer Deaths	All - Breast Cancer	Population	Rate, all causes, mx	other causes, mx(i)	qx all	qx(i) other	(expected deaths after age x) all causes ((lx-1)-(dx-1)	No. of "deaths" dx all causes (lx*qx)	No. of "deaths" dx(i) other lx*qx(i)	deaths after age x causes Wx(i)	Cumulative Probability of Death before age x all causes	other causes Fx(i)	Age
(x)	1995-1997	1995-1997	1995-1997	1995-1997	1995-1997	1995-1997									(x)
80	9,511	309	9,202	199,263	0.04773	0.04618	0.04662	0.04510	63,794.9	2,974.0	2,877.4	62,467.7	0.3621	0.3499	80
81	10,017	329	9,688	188,763	0.05307	0.05132	0.05169	0.05000	60,820.9	3,144.1	3,040.9	59,590.3	0.3918	0.3798	81
82	10,540	307	10,233	175,462	0.06007	0.05832	0.05832	0.05662	57,676.8	3,363.6	3,265.6	56,549.4	0.4232	0.4115	82
83	10,725	314	10,411	157,677	0.06802	0.06603	0.06578	0.06386	54,313.2	3,572.8	3,468.2	53,283.8	0.4569	0.4455	83
84	10,753	337	10,416	140,270	0.07666	0.07426	0.07383	0.07152	50,740.4	3,746.1	3,628.7	49,815.6	0.4926	0.4816	84
85	10,464	286	10,178	123,517	0.08472	0.08240	0.08127	0.07905	46,994.2	3,819.4	3,715.0	46,186.8	0.5301	0.5193	85
86	10,277	228	10,049	108,109	0.09506	0.09295	0.09075	0.08873	43,174.8	3,918.0	3,831.1	42,471.8	0.5683	0.5580	86
87	10,031	217	9,814	93,956	0.10676	0.10445	0.10135	0.09916	39,256.8	3,978.8	3,892.7	38,640.7	0.6074	0.5979	87
88	9,538	200	9,338	80,439	0.11857	0.11609	0.11194	0.10959	35,278.0	3,948.9	3,866.1	34,748.0	0.6472	0.6384	88
89	9,278	168	9,110	67,963	0.13652	0.13404	0.12779	0.12548	31,329.1	4,003.6	3,931.1	30,881.9	0.6867	0.6786	89
90	8,553	136	8,417	57,073	0.14986	0.14748	0.13941	0.13720	27,325.4	3,809.6	3,749.0	26,950.7	0.7267	0.7195	90
91	7,975	133	7,842	46,302	0.17224	0.16937	0.15858	0.15594	23,515.9	3,729.2	3,667.0	23,201.8	0.7648	0.7585	91
92	6,968	105	6,863	38,209	0.18237	0.17962	0.16713	0.16461	19,786.7	3,306.9	3,257.0	19,534.8	0.8021	0.7967	92
93	6,038	83	5,955	29,836	0.20237	0.19959	0.18378	0.18125	16,479.8	3,028.6	2,987.0	16,277.7	0.8352	0.8306	93
94	5,100	66	5,034	22,997	0.22177	0.21890	0.19963	0.19705	13,451.2	2,685.3	2,650.5	13,290.7	0.8655	0.8617	94
95	4,302	70	4,232	17,967	0.23944	0.23554	0.21384	0.21036	10,765.9	2,302.2	2,264.7	10,640.2	0.8923	0.8893	95
96	3,441	50	3,391	12,845	0.26789	0.26399	0.23624	0.23281	8,463.7	1,999.5	1,970.4	8,375.5	0.9154	0.9128	96
97	2,711	47	2,664	8,975	0.30206	0.29682	0.26243	0.25788	6,464.2	1,696.4	1,667.0	6,405.0	0.9354	0.9333	97
98	2,179	23	2,156	6,406	0.34015	0.33656	0.29071	0.28764	4,767.9	1,386.1	1,371.4	4,738.1	0.9523	0.9507	98
99	1,556	11	1,545	4,503	0.34555	0.34310	0.29464	0.29256	3,381.8	996.4	989.4	3,366.6	0.9662	0.9650	99
100+	3,382	33	3,349	8,015	0.42196	0.41784	0.34844	0.34504	2,385.4	831.2	823.1	2,377.3	0.9761	0.9753	100+
101+							1.00000	0.00000	1,554.2	1,554.2	1,554.2	1,554.2	0.9845	0.9838	
									0.0						
0-130	304,167	14,938	289,229	44,398,008		0.00651				100,000.0	96,087.5		1.0000		0-130
	0	0													

Notes: Average annual rates are calculated by dividing the three-year mortality counts by the three-year total population

qx = [2*a*mx]/[2+a*mx] where a = width of age group (ie, 1 or 5) and Mx = the mortality rate for each age or age group

qx(i) = [2*a*mx(i)]/[2+a*mx]

APPENDIX E

Life table calculations for cumulative probability of birth before a given year, Canada, females, 1896-1996

Age	All Deaths		Population		Rate, all causes, mx	qx	No. of "deaths" dx	Cumulative probability of death before age x, all deaths	Births needed= popn at age(x)/ (1-F(x+1))	Cumulative Probability of Birth by Year		Birth Year
(x)	1995-1997	average	1995-1997	average	1995-1997		(lx*qx)	Fx		No.	Ratio	
00-04	3,367	1,122.3	2,867,920	955,973	0.00117	0.00585	585.3	0.0000	961,602	17,043,061	1.000	1996
05-09	415	138.3	2,954,700	984,900	0.00014	0.00070	69.8	0.0059	991,394	16,081,459	0.944	1991
10-14	471	157.0	2,932,826	977,609	0.00016	0.00080	79.7	0.0066	984,846	15,090,065	0.885	1986
15-19	967	322.3	2,928,032	976,011	0.00033	0.00165	163.8	0.0073	984,861	14,105,219	0.828	1981
20-24	955	318.3	2,968,044	989,348	0.00032	0.00161	159.3	0.0090	999,926	13,120,358	0.770	1976
25-29	1,235	411.7	3,214,964	1,071,655	0.00038	0.00192	189.9	0.0106	1,085,195	12,120,432	0.711	1971
30-34	1,955	651.7	3,823,281	1,274,427	0.00051	0.00255	252.2	0.0125	1,293,834	11,035,237	0.647	1966
35-39	2,963	987.7	3,918,201	1,306,067	0.00076	0.00377	371.7	0.0150	1,330,978	9,741,403	0.572	1961
40-44	4,357	1,452.3	3,578,349	1,192,783	0.00122	0.00607	595.6	0.0187	1,222,956	8,410,425	0.493	1956
45-49	6,092	2,030.7	3,176,650	1,058,883	0.00192	0.00954	930.8	0.0247	1,096,130	7,187,468	0.422	1951
50-54	7,567	2,522.3	2,515,236	838,412	0.00301	0.01493	1,442.3	0.0340	881,058	6,091,339	0.357	1946
55-59	10,052	3,350.7	2,021,110	673,703	0.00497	0.02456	2,337.3	0.0484	725,798	5,210,281	0.306	1941
60	2,375	791.7	374,073	124,691	0.00635	0.00633	587.5	0.0718	135,188	4,484,483	0.263	1936
61	2,579	859.7	368,785	122,928	0.00699	0.00697	642.8	0.0777	134,213	4,349,294	0.255	1935
62	2,840	946.7	365,846	121,949	0.00776	0.00773	708.3	0.0841	134,181	4,215,082	0.247	1934
63	3,079	1,026.3	365,851	121,950	0.00842	0.00838	761.7	0.0912	135,317	4,080,901	0.239	1933
64	3,386	1,128.7	368,901	122,967	0.00918	0.00914	823.4	0.0988	137,703	3,945,584	0.232	1932
65	3,771	1,257.0	367,028	122,343	0.01027	0.01022	912.8	0.1070	138,419	3,807,881	0.223	1931
66	4,108	1,369.3	360,855	120,285	0.01138	0.01132	1,000.5	0.1161	137,649	3,669,463	0.215	1930
67	4,338	1,446.0	353,831	117,944	0.01226	0.01219	1,064.8	0.1261	136,634	3,531,814	0.207	1929
68	4,739	1,579.7	346,396	115,465	0.01368	0.01359	1,172.9	0.1368	135,606	3,395,180	0.199	1928
69	5,135	1,711.7	343,613	114,538	0.01494	0.01483	1,263.0	0.1485	136,542	3,259,574	0.191	1927
70	5,636	1,878.7	338,833	112,944	0.01663	0.01650	1,383.8	0.1612	136,901	3,123,032	0.183	1926
71	6,062	2,020.7	333,837	111,279	0.01816	0.01800	1,484.6	0.1750	137,354	2,986,132	0.175	1925
72	6,452	2,150.7	324,888	108,296	0.01986	0.01966	1,593.1	0 1898	136,353	2,848,778	0.167	1924
73	6,936	2,312.0	316,403	105,468	0.02192	0.02168	1,722.2	0.2058	135,735	2,712,425	0.159	1923
74	7,428	2,476.0	304,934	101,645	0.02436	0.02407	1,870.0	0.2230	134,041	2,576,690	0.151	1922
75	8,079	2,693.0	291,008	97,003	0.02776	0.02738	2,076.4	0.2417	131,521	2,442,649	0.143	1921
76	8,116	2,705.3	266,011	88,670	0.03051	0.03005	2,216.4	0.2625	123,948	2,311,128	0.136	1920
77	8,245	2,748.3	241,369	80,456	0.03416	0.03359	2,402.7	0 2846	116,375	2,187,180	0.128	1919
78	8,247	2,749.0	219,670	73,223	0.03754	0.03685	2,547.7	0 3086	109,965	2,070,805	0.122	1918
79	8,881	2,960.3	207,298	69,099	0.04284	0 04194	2,792 9	0 3341	108,315	1,960,840	0.115	1917

Age	All Deaths		Population		Rate, all causes, mx	qx	No. of "deaths" dx	Cumulative probability of death before age x, all deaths	Births needed= popn at age(x)/ (1-F(x+1))	Cumulative Probability of Birth by Year		
(x)	1995-1997	average	1995-1997	average	1995-1997		(lx*qx)	Fx		No.	Ratio	Birth Year
80	9,511	3,170.3	199,263	66,421	0.04773	0.04662	2,974.0	0.3621	109,207	1,852,525	0.109	1916
81	10,017	3,339.0	188,763	62,921	0.05307	0.05169	3,144.1	0.3918	109,092	1,743,318	0.102	1915
82	10,540	3,513.3	175,462	58,487	0.06007	0.05832	3,363.6	0.4232	107,685	1,634,225	0.096	1914
83	10,725	3,575.0	157,677	52,559	0.06802	0.06578	3,572.8	0.4569	103,584	1,526,540	0.090	1913
84	10,753	3,584.3	140,270	46,757	0.07666	0.07383	3,746.1	0.4926	99,494	1,422,956	0.083	1912
85	10,464	3,488.0	123,517	41,172	0.08472	0.08127	3,819.4	0.5301	95,362	1,323,461	0.078	1911
86	10,277	3,425.7	108,109	36,036	0.09506	0.09075	3,918.0	0.5683	91,796	1,228,099	0.072	1910
87	10,031	3,343.7	93,956	31,319	0.10676	0.10135	3,978.8	0.6074	88,777	1,136,303	0.067	1909
88	9,538	3,179.3	80,439	26,813	0.11857	0.11194	3,948.9	0.6472	85,585	1,047,526	0.061	1908
89	9,278	3,092.7	67,963	22,654	0.13652	0.12779	4,003.6	0.6867	82,906	961,941	0.056	1907
90	8,553	2,851.0	57,073	19,024	0.14986	0.13941	3,809.6	0.7267	80,900	879,035	0.052	1906
91	7,975	2,658.3	46,302	15,434	0.17224	0.15858	3,729.2	0.7648	78,002	798,135	0.047	1905
92	6,968	2,322.7	38,209	12,736	0.18237	0.16713	3,306.9	0.8021	77,284	720,133	0.042	1904
93	6,038	2,012.7	29,836	9,945	0.20237	0.18378	3,028.6	0.8352	73,936	642,849	0.038	1903
94	5,100	1,700.0	22,997	7,666	0.22177	0.19963	2,685.3	0.8655	71,203	568,912	0.033	1902
95	4,302	1,434.0	17,967	5,989	0.23944	0.21384	2,302.2	0.8923	70,761	497,709	0.029	1901
96	3,441	1,147.0	12,845	4,282	0.26789	0.23624	1,999.5	0.9154	66,236	426,948	0.025	1900
97	2,711	903.7	8,975	2,992	0.30206	0.26243	1,696.4	0.9354	62,747	360,712	0.021	1899
98	2,179	726.3	6,406	2,135	0.34015	0.29071	1,386.1	0.9523	63,142	297,966	0.017	1898
99	1,556	518.7	4,503	1,501	0.34555	0.29464	996.4	0.9662	62,925	234,824	0.014	1897
100+	3,382	1,127.3	8,015	2,672	0.42196	0.34844	831.2	0.9761	171,899	171,899	0.010	1896
101+		0.0		0		1.00000	1,554.2	0.9845	0	0	0.000	
0-130	304,167		44,398,008	14,799,336			100,000.0	1.0000	17,043,061			

MQ

66042

UMI

MICROFILMED 2002

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

PERMISSION DE REPRODUIRE ET DE DISTRIBUER LA THÈSE

PERMISSION TO REPRODUCE AND DISTRIBUTE THE THESIS

NOM DE L'AUTEUR / NAME OF AUTHOR:	GAUTHIER, Eric
ADRESSE POSTALE / MAILING ADDRESS:	1-35 St-Dominique Hull, Québec J9A 1A2
GRADE / DEGREE:	**ANNÉE D'OBTENTION / YEAR GRANTED**
M.A.SC. GÉNIE ÉLECTRIQUE	2001
TITRE DE LA THÈSE / TITLE OF THESIS: CAPACITÉ DES SYSTÈMES DE TÉLÉCOMMUNICATION À MULTIPLES ANTENNES DANS LES CANAUX À ÉVANOUISSEMENTS LENTS DE RAYLEIGH.	

L'auteur permet, par la présente, la consultation et le prêt de cette thèse en conformité avec les règlements établis par le bibliothécaire en chef de l'Université d'Ottawa. L'auteur autorise aussi l'Université d'Ottawa, ses sumccesseurs et cessionnaires, à reproduire cet exemplaire par photograhie ou photocopie pour fins de prêt ou de vente au prix coûtant aux bibliothèques ou aux chercheurs qui en feront la demande.

Les droits de publication par tout autre moyen et pour vente au public demeureront la propriété de l'auteur de la thèse sous réserve des règlements de l'Université d'Ottawa en matière de publication de thèses.

The author hereby permits the consultation and the lending of this thesis pursuant to the regulations established by the Chief Librarian of the University of Ottawa. The author also authorizes the University of Ottawa, its successors and assignees, to make reproductions of this copy by photographic means or by photocopying and to lend or sell such reproductions at cost to libraries and to scholars requesting them.

The right to publish the thesis by other means and to sell it to the public is reserved to the author, subject to the regulations of the University of Ottawa governing the publication of theses.

N.B. LE MASCULIN COMPREND ÉGALEMENT LE FÉMININ

19/11/01

DATE

E. Gauthier

(AUTEUR) SIGNATURE (AUTHOR)



Université d'Ottawa • University of Ottawa

Université d'Ottawa • University of Ottawa

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

GAUTHIER, Éric

AUTEUR DE LA THÈSE - AUTHOR OF THESIS

M.ASc. (Génie électrique)

GRADE - DEGREE

École d'ingénierie et de technologie de l'information

FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

Capacité des systèmes de télécommunication à antennes multiples dans les canaux à évanouissements lents de Rayleigh

Jean-Yves Chouinard et Abbas Yongaçoglu

DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

C. D'Amours

F. Danilo-Lemoine

J.-M. De Koninck, Ph.D.

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

SIGNATURE

DEAN OF THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

CAPACITÉ DES SYSTÈMES DE TÉLÉCOMMUNICATION À ANTENNES MULTIPLES DANS LES CANAUX À ÉVANOUISSEMENTS LENTS DE RAYLEIGH

par

ÉRIC GAUTHIER

Thèse soumise à la
Faculté des études supérieures et postdoctorales
conformément aux exigences du grade de
Maître es Sciences Appliquées
en génie électrique

Thèse de maîtrise

École d'ingénierie et de technologie de l'information
Département de génie électrique
Faculté de génie
Université d'Ottawa
Ottawa, Ontario

© Eric Gauthier, Ottawa, Canada, Septembre 2001

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-66042-7

Canada

Résumé

Cette thèse porte sur la capacité des systèmes de télécommunication dans les canaux à évanouissements lents de Rayleigh. On s'intéresse plus particulièrement aux systèmes comportant plusieurs antennes à l'émetteur et au récepteur et à l'effet de cette diversité spatiale sur la capacité. La capacité de systèmes à antennes multiples est d'abord étudiée pour le cas où l'état du canal n'est connu qu'au récepteur. Ensuite, nous étudions le cas où l'émetteur connaît aussi l'état du canal, ce qui permet alors d'adapter les symboles transmis selon la réalisation du canal. Nous présentons alors un algorithme afin de déterminer le prétraitement à l'émetteur qui maximise la capacité. Finalement, nous approfondissons le cas où l'émetteur connaît aussi l'état du canal en simulant des erreurs d'estimation du canal à l'émetteur, et nous présentons l'effet de ces estimations imparfaites sur la capacité.

Remerciements

L'idée de faire une maîtrise m'est venue après avoir complété ma spécialisation en théorie des communications dans le cadre de mon baccalauréat en génie électrique. J'avais l'impression que je n'avais vu que la pointe de l'iceberg et il me fallait en savoir plus sur ce sujet passionnant. J'ai eu l'opportunité ces dernières années de réaliser cette thèse de maîtrise grâce aux recommendations et à la supervision de mes directeurs de thèse, le Dr Jean-Yves Chouinard et le Dr Abbas Yongaçoglu, que je remercie pour leur support tout au long des trois dernières années.

Aussi, je remercie le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) et l'Université d'Ottawa pour leur appui financier, sans lequel cette thèse n'aurait pas vu le jour.

Table des matières

Table des matières v

Liste des figures vii

Liste des symboles viii

1 Introduction 1

1.1 Capacité des systèmes de télécommunication à antennes multiples 2

1.1.1 Définition de la capacité 2

1.1.2 Canal à évanouissements lents de Rayleigh 3

1.1.3 Diversité de transmission et réception 3

1.2 Objectif et organisation de la thèse 5

1.3 Contributions originales 6

2 Capacité du canal à antennes multiples: notions de base 8

2.1 Introduction 8

2.2 Calcul de la capacité 9

2.2.1 Canal additif gaussien sans diversité ($m = n = 1$): 10

2.2.2 Cas où m et n sont des entiers quelconques et le canal est additif gaussien 11

2.2.3 Cas où m et n sont des entiers quelconques et le canal est à évanouissements lents de Rayleigh avec bruit additif gaussien 13

2.3 Conclusion 14

3 Capacité lorsque l'état du canal est connu au récepteur seulement **15**

3.1 Introduction 15

3.2 Capacité asymptotique 16

3.2.1 Configuration d'antennes $(m.n) = (1.1)$ (sans diversité) 19

3.2.2 Configuration d'antennes $min(m,n) = 1$, $(m.n) = t$ (diversité d'ordre t à l'émetteur ou au récepteur) 20

3.2.3 Diversité d'ordre supérieur à 1, $(m.n) > 1$ (cas général) 23

3.2.4 Cas limite lorsque le nombre d'antennes est très large 25

3.3 Capacité asymptotique avec corrélation parfaite entre les antennes 29

3.3.1 Effet de la corrélation pour $(m,n) = 1$. $(m.n) = t$ 29

3.3.2 Effet de la corrélation pour $(m.n) > 1$ 34

3.4 Fonction de répartition de la capacité 38

3.5 Capacité pour un canal à large bande 45

3.6 Conclusion 46

4 Canal connu au récepteur et à l'émetteur **49**

4.1 Introduction 49

4.2 Expression générale de la capacité avec prétraitement 50

4.2.1 Cas particulier de la contrainte sur la puissance 52

4.3 Maximisation de l'information mutuelle 53

4.3.1 Maximisation sur la matrice des vecteurs propres $\mathbf{Q}$ 54

4.3.2 Maximisation sur la matrice des valeurs propres $\mathbf{D}$ 55

4.3.3 Transformation linéaire $\mathbf{G}$ 59

4.4 Capacité asymptotique 61

4.4.1 Asymptote pour $n \geq m$ 62

4.5 Gain obtenu suite à la transformation linéaire $\mathbf{G}$ pour $\rho \gg 1$ 65

4.6 Symmétrie de la capacité 66

4.7 Conclusions 67

5 Erreur d'estimation du canal à l'émetteur **69**

5.1 Introduction 69

5.2 Modélisation de l'erreur d'estimation 70

5.3 Résultats de simulation 71

5.3.1 Capacité en fonction de la corrélation 74

5.3.2 Capacité en fonction du rapport signal-sur-bruit 85

5.4 Conclusion 93

6 Conclusions et suggestions de travaux de recherche **94**

6.1 Sommaire 94

6.2 Suggestions de travaux de recherche 96

A Calculs alternatifs des asymptotes pour $(m, n) = 1$ **97**

A.1 Cas à antennes uniques de transmission et de réception $(m = n = 1)$ 98

A.2 Cas avec une antenne émettrice et n antennes réceptrices $(1, n)$ 99

A.3 Cas à m antennes émettrices et une antenne réceptrice $(m, 1)$ 100

B Les valeurs propres non nulles des matrices $\mathbf{H}\mathbf{H}^H$ et $\mathbf{H}^\Gamma\mathbf{H}^{\Gamma H}$ sont identiques (preuves) **102**

C Calcul de $\sum_{i=1}^{(m,n)} E_{\mathbf{H}}[\log_2 \lambda_i]$ **104**

D Calcul de $\sum_{i=1}^{(m,n)} E_{\mathbf{H}}[\lambda_i]$ **106**

Bibliographie **109**

Liste des figures

1.1 Modèle du canal multiplicatif et additif gaussien. 4

1.2 Évanouissements de Rayleigh typique 5

2.1 Système de télécommunications à antennes multiples. 9

3.1 Capacité pour quelques valeurs de (m, n) 17

3.2 Gain sur la capacité en fonction du nombre d'antennes: 22

3.3 Asymptotes de la capacité $(m, n) = (1, 1)$, $(2, 2)$, $(4, 4)$ et $(8, 8)$ 24

3.4 Comparaison de la capacité du canal de Rayleigh et du canal avec bruit blanc additif gaussien pour $m = 1$, 4 ou 16 antennes émettrices et une antenne réceptrice $(n = 1)$ 27

3.5 Comparaison de la capacité du canal de Rayleigh et du canal avec bruit blanc additif gaussien pour une antenne émettrice $(m = 1)$ et $n = 1$, 4 ou 16 antennes réceptrices 28

3.6 Capacité asymptotique avec et sans corrélation entre les antennes réceptrices pour $(m, n) = (1, 2)$, $(1, 4)$ et $(1, 8)$ 32

3.7 Capacité asymptotique avec et sans corrélation entre les antennes émettrices pour $(m, n) = (2, 1)$, $(4, 1)$ et $(8, 1)$ 33

3.8 Capacité asymptotique avec et sans corrélation entre les antennes émettrices pour $(m, n) = (8, 2)$, $(2, 8)$ et $(8, 8)$ 36

3.9 Capacité asymptotique avec et sans corrélation entre les antennes réceptrices pour $(m, n) = (8, 2)$, $(2, 8)$ et $(8, 8)$ 37

3.10 Fonction de répartition de la capacité 39
3.11 Fonction de répartition de la capacité pour différentes valeurs de m 41
3.12 Fonction de répartition de la capacité pour différentes valeurs de n 42
3.13 Fonction de densité de probabilité de la capacité pour différentes valeurs de m 43
3.14 Fonction de densité de probabilité de la capacité pour différentes valeurs de n 44
3.15 Capacité pour un canal à large bande 47

4.1 Système à antennes multiples avec prétraitement. 50
4.2 Principe du remplissage ("waterfilling"). 56
4.3 Capacité avec prétraitement optimal 60
4.4 Comparaison de la capacité: prétraitement optimal et aucun prétraitement . 63
4.5 Comparaison de la capacité: prétraitement optimal et aucun prétraitement . 64

5.1 Corrélation entre $\mathbf{H}$ et $\tilde{\mathbf{H}}$ en fonction de la variance de l'erreur d'estimation . 72
5.2 Capacité en fonction de la corrélation. $\rho = -15$ dB et (m, n)=(3.1) 76
5.3 Capacité en fonction de la corrélation. $\rho = 3$ dB et (m, n)=(3.1) 77
5.4 Capacité en fonction de la corrélation. $\rho = 21$ dB et (m, n)=(3.1) 78
5.5 Capacité en fonction de la corrélation. $\rho = -15$ dB et (m, n)=(3.3) 79
5.6 Capacité en fonction de la corrélation. $\rho = 3$ dB et (m, n)=(3.3) 80
5.7 Capacité en fonction de la corrélation. $\rho = 21$ dB et (m, n)=(3.3) 81
5.8 Capacité en fonction de la corrélation. $\rho = -15$ dB et (m, n)=(3.5) 82
5.9 Capacité en fonction de la corrélation. $\rho = 3$ dB et (m, n)=(3.5) 83
5.10 Capacité en fonction de la corrélation. $\rho = 21$ dB et (m, n)=(3.5) 84
5.11 Capacité pour différentes valeurs de corrélation. (m, n)=(3.1) 88
5.12 Capacité pour différentes valeurs de corrélation. (m, n)=(3.3) 89
5.13 Capacité pour différentes valeurs de corrélation. (m, n)=(3.3) (agrandissement) 90
5.14 Capacité pour différentes valeurs de corrélation. (m, n)=(3.5) 91
5.15 Capacité pour différentes valeurs de corrélation. (m, n)=(3.5) (agrandissement) 92

Liste des symboles

Notation générale

$\mathbf{x}$	Vecteur
X	Variable aléatoire
$\mathbf{A}$	Matrice, ou vecteur/matrice de variables aléatoires
$\mathbf{A}^T$	Transposé de $\mathbf{A}$
$\mathbf{A}^H$	Transconjugué complexe de $\mathbf{A}$
$\mathrm{tr}(\mathbf{A})$	Trace de $\mathbf{A}$
$\det(\mathbf{A})$	Déterminant de $\mathbf{A}$
I_k	Matrice identité d'ordre k
$E_X[\cdot]$	Espérance mathématique selon la variable aléatoire X

Symboles

m	Nombre d'antennes émettrices
n	Nombre d'antennes réceptrices
ρ	Rapport signal-sur-bruit
$\rho_{X,Y}$	Coefficient de corrélation entre les variables aléatoires X et Y
C	Capacité
C_{asympt}	Capacité asymptotique ($\rho \gg 1$)
$\mathbf{H}^{n\times m}$	Matrice des gains complexes du canal

Constantes

- γ Constante d'Euler-Mascheroni $\approx 0.57721566490153329$

Chapitre 1

Introduction

On assiste ces dernières années à une croissance de la popularité des systèmes de communications sans fil. De plus, le nombre d'applications pour ces systèmes se multiplie et n'est maintenant plus limité à la transmission de la voix. En effet, les systèmes sans fil de troisième génération devront être capable de supporter des applications multimédia utilisant les images, vidéos, sons et voix; tout en répondant aux attentes des utilisateurs en terme de qualité d'information. Techniquement, cette nouvelle demande se traduit par un débit de transmission élevé, donc une plus grande capacité et une meilleure efficacité spectrale sont nécessaires par rapport aux systèmes actuels. Plusieurs techniques ont été proposées afin d'améliorer l'efficacité spectrale des systèmes sans fil de troisième génération, l'une d'elle étant la diversité spatiale, ou en d'autres mots l'utilisation de plusieurs antennes de transmission et/ou de réception. Jusqu'à récemment, la diversité spatiale n'était qu'une notion théorique puisque les moyens technologiques ne permettaient pas sa réalisation. En effet, les réseaux d'antennes occupaient un espace trop important et la puissance de calcul requise afin de contrôler ces antennes ne pouvait être obtenue que par de volumineux processeurs. Cependant, avec les progrès technologiques récents, dont l'augmentation toujours croissante de la puissance de calcul des processeurs et l'utilisation de fréquences plus élevées, la diversité spatiale émetteur-récepteur semble être une technique maintenant réalisable.

1.1 Capacité des systèmes de télécommunication à antennes multiples

1.1.1 Définition de la capacité

La capacité d'un canal de télécommunications peut être définie de plusieurs façons. Fondamentalement, la capacité est définie comme le maximum de l'information mutuelle entre deux variables aléatoires, X et Y [CT91]:

$$C = \max_{f_X(x)} I(X;Y) \tag{1.1}$$

où $f_X(x)$ est la fonction de densité de probabilité de X. Dans un système de communications, ces deux variables X et Y sont les données envoyées par l'émetteur et celles reçues au récepteur respectivement. La maximisation de l'information mutuelle se fait sur la distribution des données transmises et dépend du canal. De plus, la distribution des données reçues dépend de celle des données transmises et des paramètres du canal. La capacité est donc purement fonction des paramètres du canal. Pour un système à largeur de bande restreinte, la capacité peut aussi se traduire par un débit de transmission maximum avec une probabilité d'erreur arbitrairement basse. Il n'est donc pas possible de construire un code correcteur d'erreur qui pourra émettre à un débit plus élevé que la capacité tout en assurant un taux d'erreur négligeable. Un des buts poursuivis par les ingénieurs de communications est de développer des techniques permettant l'atteinte de cette capacité. Il faut comprendre que le calcul de capacité donne une limite supérieure à ce qu'un système réel pourra atteindre et donne peu d'indices sur la façon d'y parvenir. On peut donc diviser le travail de recherche de ces codes en deux parties: une première phase théorique qui consiste à calculer cette capacité, ou limite supérieure, pour un canal donné, et une deuxième plus pratique qui consiste à concevoir un système réel pour ce canal qui s'approchera le plus possible de cette limite. Dans cette thèse, nous utiliserons aussi le terme "capacité" pour désigner le maximum de l'information mutuelle entre deux variables lorsque des contraintes sur la distribution des symboles transmis sont imposées. Nous utiliserons donc le terme capacité

pour décrire l'efficacité spectrale d'un système dans un canal donné, ce qui est une pratique courante dans la littérature récente sur la théorie de l'information.

1.1.2 Canal à évanouissements lents de Rayleigh

Comme mentionné auparavant, la capacité dépend des paramètres du canal dans lequel le système sera utilisé. Un modèle de canal très répandu pour les communications sans fil est le canal avec évanouissement de Rayleigh corrompu de bruit additif gaussien [Rap96]. Le modèle de canal utilisé tout au long de cette thèse est non-sélectif en fréquence et varie lentement par rapport au débit de transmission dans le canal. Le canal est donc simplement multiplicatif: les signaux transmis sont multipliés par un coefficient h complexe circulaire gaussien, dont les partie réelle et imaginaire sont non-corrélées. Le coefficient h peut s'exprimer par une somme de variable aléatoire normales non-corrélées, centrées à zéro avec une variance de valeur $1/2$, c'est-à-dire que $h = N(0, 1/2) + jN(0, 1/2)$. L'enveloppe de h, illustrée à la figure 1.2 en fonction du temps, est une variable aléatoire de Rayleigh, d'où le nom du canal. De plus, le canal est corrompu par un bruit blanc additif, z, qui possède une distribution gaussienne complexe centrée à zéro avec variance $= \sigma_z^2$: on a donc $z = N(0, \sigma_z^2/2) + jN(0, \sigma_z^2/2)$. Le modèle du canal multiplicatif avec bruit additif pour une entrée $x(t)$ est illustré à la figure 1.1, où les signaux $x(t)$ et $y(t)$ sont représentés par leur enveloppe complexe et $h(t)$ et $z(t)$ sont des coefficients complexes tels que définis auparavant et dont la valeur change avec le temps.

1.1.3 Diversité de transmission et réception

À la figure 1.2, on remarque que le canal de Rayleigh cause de profonds évanouissements. Le principe de diversité est basé sur le fait que les erreurs arrivent lors de ces évanouissements sévères. En émettant sur plusieurs canaux indépendants, créant ainsi une diversité, la probabilité que tous les canaux subissent un évanouissement prononcé simultanément est faible, d'où l'augmentation de capacité du canal avec diversité. La diversité consiste donc à ajouter de la redondance au message transmis afin de combattre les effets des évanouissements. Il

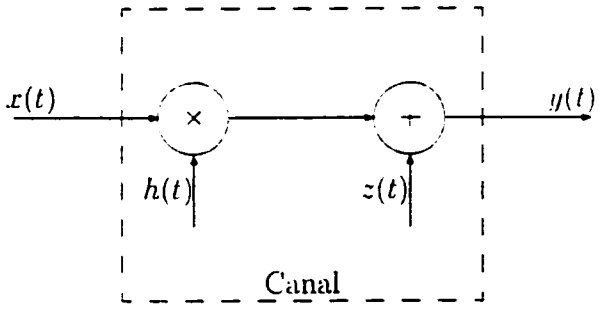


Figure 1.1: Modèle du canal multiplicatif et additif gaussien.

existe plusieurs types de diversité [Rap96]:

- diversité de fréquence: l'information est envoyée sur différentes ondes porteuses où la séparation entre chacune des porteuses est supérieure à la largeur de bande de cohérence du canal, ce qui assure l'indépendance des évanouissements de chaque porteuse. La largeur de bande de cohérence correspond à l'inverse de l'étalement en temps des réflections du signal dans le canal [Pro95].

- diversité temporelle: on obtient ici la redondance en émettant l'information dans plusieurs intervalles de temps séparés par au moins le temps de cohérence du canal, assurant ainsi l'indépendance des signaux reçus. Le temps de cohérence correspond à l'inverse de l'étalement en fréquence du signal dans le canal [Pro95].

- diversité d'espace: il est aussi possible d'obtenir plusieurs canaux indépendants en utilisant plusieurs antennes séparées par une distance suffisamment grande pour assurer des trajets indépendants, et donc des évanouissements indépendants. Cette distance minimale dépend de l'arrangement géométrique des antennes et peut aller d'une demie longueur d'onde pour un environnement très dense à plusieurs dizaines de longueurs d'ondes pour un espace plus vaste.

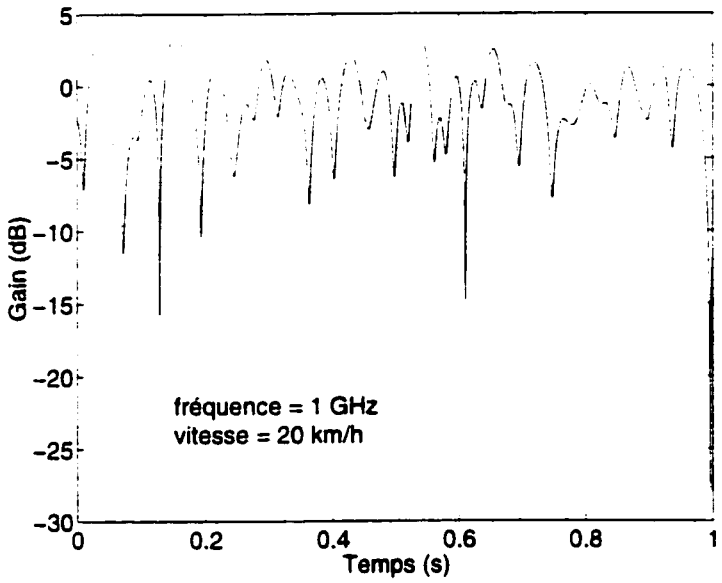


Figure 1.2: Évanouissements de Rayleigh typique (f_c = 1 GHz, v = 20 km/h).

1.2 Objectif et organisation de la thèse

En premier lieu, cette thèse s'intéresse à l'aspect théorique de l'information, c'est-à-dire aux calculs de capacité. La capacité est étudiée dans le cas où la diversité spatiale est utilisée à l'émetteur et au récepteur. La diversité spatiale à l'émetteur et au récepteur est un sujet de recherche nouveau puisque la technologie la rendait jusqu'à tout récemment impraticable. Ces dernières années, plusieurs chercheurs ont commencé l'étude de la capacité de systèmes profitant de la diversité spatiale. La plupart des travaux effectués sur le sujet ne considèrent que le cas où l'émetteur ne connaît pas l'état du canal. Notre objectif principal est de continuer les recherches sur la capacité de canaux avec diversité spatiale, en particulier pour le cas où le canal est connu de l'émetteur.

Le chapitre deux de cette thèse discute de la capacité de systèmes à antennes multiples lorsque le canal n'est pas connu à l'émetteur. Ce chapitre constitue un rappel de quelques résultats de la théorie de l'information, allant du cas le plus simple qui consiste en un système à une antenne émettrice et une antenne réceptrice, jusqu'au cas général à m antennes émettrices et n antennes réceptrices.

Au chapitre trois, l'étude du cas où l'état du canal n'est pas connu à l'émetteur est approfondi. Nous dérivons d'abord les expressions asymptotiques de capacité pour un grand rapport signal-sur-bruit; cette capacité asymptotique constitue aussi une limite inférieure de capacité. Ensuite, la fonction de densité de probabilité de la capacité prise comme variable aléatoire est aussi dérivée pour le cas où le minimum du nombre d'antennes émettrices et réceptrices, $\min(m, n)$, est égal à un. Finalement, la capacité d'un canal à antennes multiples à large bande est dérivée.

Le chapitre quatre est la suite logique du chapitre précédent et discute de la capacité lorsque le canal est aussi connu à l'émetteur. Dans ce cas, on a le problème intéressant de déterminer la distribution optimale des symboles transmis afin d'atteindre la capacité réelle du canal selon sa définition fondamentale. La matrice de covariance optimale des symboles transmis est donc dérivée et un algorithme permettant son calcul est présenté. Ensuite, les capacités asymptotiques sont aussi calculées tout comme au chapitre trois.

Finalement, le chapitre cinq traite du cas où l'estimation du canal par l'émetteur n'est pas parfaite et de l'effet de cette erreur d'estimation sur la capacité du canal.

1.3 Contributions originales

Le chapitre deux est une revue de littérature et présente quelques notions de base sur la capacité et se termine en présentant une expression générale de capacité pour les systèmes à antennes multiples obtenue par Foschini et Gans [FG98].

Les premières contributions originales de cette thèse se retrouvent au chapitre trois où les expressions asymptotiques de capacité pour un grand rapport signal-sur-bruit sont dérivées. L'expression de capacité asymptotique pour la configuration d'antennes générale (m, n) est donnée par l'équation (3.23) et quelques asymptotes pour différentes configurations (m, n) sont tracées à la figure 3.3. Ensuite, la capacité asymptotique est obtenue pour les cas où les antennes émettrices ou les antennes réceptrices sont parfaitement corrélées, dont les expressions sont données aux équations (3.40) et (3.42). Aussi, une expression générale

de capacité pour un canal à large bande est dérivée (3.52).

Au chapitre quatre, nous présentons à la page 49 un algorithme permettant de déterminer le prétraitement optimal lorsque l'émetteur connaît l'état du canal. Tout comme au chapitre trois, l'expression de capacité asymptotique pour la configuration d'antennes générale (m, n) est présentée à l'équation (4.48). Finalement, le gain sur le rapport signal-sur-bruit obtenu du prétraitement est présenté à l'équation (4.55) pour un grand rapport signal-sur-bruit.

La plus importante contribution de cette thèse se retrouve au chapitre cinq, où l'effet sur la capacité d'une erreur d'estimation du canal à l'émetteur est étudié. Ce chapitre est à notre avis entièrement original, puisque cet aspect ne semble pas avoir été étudié auparavant.

Chapitre 2

Capacité du canal à antennes multiples: notions de base

2.1 Introduction

Dans ce chapitre, nous sommes intéressés à calculer la capacité du système de communications illustré à la figure 2.1. Ce calcul a été effectué par Foschini et Gans [FG98] d'après les travaux de Kullback [Kul68]. Dans les prochaines sections, les dérivations de ces auteurs sont rassemblées. Le système de communications est composé de m émetteurs et n récepteurs. Le canal est à évanouissements lents de Rayleigh et est corrompu de bruit additif gaussien, tel que décrit au chapitre précédent, avec la seule différence que les variables $x(t)$, $y(t)$ et $z(t)$ sont maintenant considérées constantes durant chaque utilisation du canal et peuvent donc simplement s'exprimer sous la forme de variables aléatoires. Aussi, on suppose que le canal change lentement, ce qui permet de faire une analyse appelée quasi-statique, dans laquelle la capacité est calculée à partir d'une réalisation aléatoire du canal qui reste constante pour la durée de la transmission. De plus, on suppose que le canal est connu au récepteur seulement. À chaque utilisation du canal, les symboles transmis et reçus sont considérées comme des variables aléatoires définies par les vecteurs $\mathbf{X} = (X_1, X_2, \ldots, X_m)^T$ et $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$

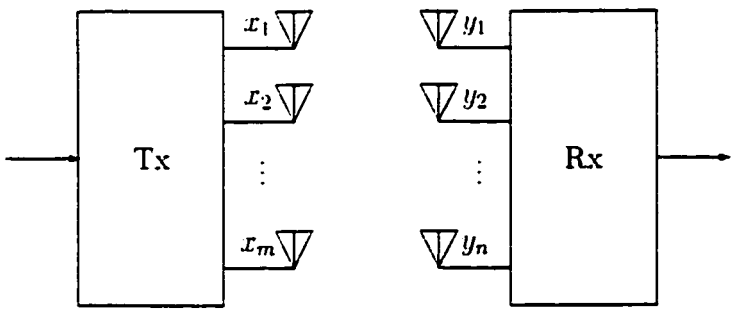


Figure 2.1: Système de télécommunications à antennes multiples.

respectivement. L'équation décrivant le canal est:

$$\mathbf{Y} = \mathbf{HX} + \mathbf{Z} \tag{2.1}$$

où $\mathbf{H}^{n\times m}$ est la matrice de la fonction de transfert du canal à m antennes de transmission et n antennes de réception. $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)^T$ est le vecteur du bruit gaussien, qui est par hypothèse statistiquement indépendant de $\mathbf{X}$. La relation entre $\mathbf{X}$ et $\mathbf{Y}$ est définie par la fonction densité de probabilité conjointe $f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})$, où $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ et $\mathbf{y} = (y_1, y_2, \ldots, y_n)$.

2.2 Calcul de la capacité

La capacité est définie comme étant le maximum de l'information mutuelle entre les vecteurs $\mathbf{X}$ et $\mathbf{Y}$ [CT91]:

$$C = \max_{f_{\mathbf{X}}(x)} I(\mathbf{X};\mathbf{Y}) \tag{2.2}$$

où la maximisation se fait sur l'ensemble des distributions possibles de la source de transmission X.

Cette section se veut être un rappel de résultats importants de la théorie de l'information. Les dérivations du calcul de la capacité sont présentées pour le cas le plus simple, le canal additif gaussien sans diversité, jusqu'au système à antennes multiples.

2.2.1 Canal additif gaussien sans diversité ($m = n = 1$):

L'information mutuelle dans le cas où $m = n = 1$ est définie comme suit [CT91]:

$$I(X;Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\log_2\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}dydx \tag{2.3}$$

L'information mutuelle peut cependant être calculée plus facilement en utilisant la relation existante entre elle et les entropies, $H(X)$ et $H(Y)$, et équivocations, $H(X|Y)$ et $H(Y|X)$ [CT91]:

$$I(X;Y) = H(Y) - H(Y|X) = H(X) - H(X|Y) \tag{2.4}$$

où $H(X) = -\int_{-\infty}^{\infty} f_X(x)\log_2 f_X(x)dx$ est l'entropie différentielle de X. Celle-ci est utilisée puisque l'entropie d'une variable aléatoire continue est infinie. L'équivocation de X étant donné Y est donnée par: $H(X|Y) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\log_2 f_{XY}(x|y)dydx$. On trouve que la valeur maximale de $I(X;Y)$ est obtenue lorsque X suit une distribution normale centrée à zéro avec variance σ_x^2 [CT91]. Comme $Y = X + Z$, où Z est le bruit gaussien de moyenne zéro avec variance σ_z^2, statistiquement indépendant de **X**, il est facile de montrer que Y est aussi gaussienne de moyenne nulle et de variance $\sigma_y^2 = \sigma_x^2 - \sigma_z^2$. Utilisant l'équation (2.4), on obtient:

$$\begin{aligned} C &= \frac{1}{2}\log_2\left(1 + \frac{\sigma_x^2}{\sigma_z^2}\right) \\ &= \frac{1}{2}\log_2\left(1 + \frac{P_{signal}}{P_{bruit}}\right) \end{aligned} \tag{2.5}$$

L'équation (2.5) donne la capacité en bits par utilisation du canal (si la base 2 est utilisée pour le logarithme) lorsque la puissance moyenne est limitée à P_{signal}. Dans un canal à bande étroite avec largeur de bande égale à B Hertz, $2B$ échantillons indépendants par seconde (ou $2B$ utilisations du canal par seconde) sont nécessaires [CT91]. La capacité possède la propriété d'addition pour des réalisations indépendantes et peut donc être exprimée de la façon suivante:

$$C = 2B\frac{1}{2}\log_2\left(1 + \frac{P_{signal}}{P_{bruit}}\right) \quad \text{bps} \tag{2.6}$$

$$= \log_2\left(1 + \frac{P_{signal}}{P_{bruit}}\right) \quad \text{bps/Hz} \tag{2.7}$$

2.2.2 Cas où m et n sont des entiers quelconques et le canal est additif gaussien

La capacité lorsque m et n sont des entiers quelconques est calculée en utilisant une variante plus générale de la formule (2.3). En effet, l'équation (2.3) peut être obtenue en utilisant une définition très générale de l'information relative. Kullback [Kul68] appelait cette définition *l'information de discrimination en faveur de l'hypothèse H_1 sur H_2*. Supposons deux hypothèses H_1 et H_2, l'information moyenne en faveur de H_1 sur H_2 est donné par [Kul68]:

$$I(1:2) = \int f_1(u) \log_2 \frac{f_1(u)}{f_2(u)} du \tag{2.8}$$

où $f_1(u)$ et $f_2(u)$ sont des fonctions de densité de probabilité de la variable aléatoire U. Afin de calculer l'information mutuelle entre X et Y, il faut poser deux hypothèses où f_1 représente la situation réelle et f_2 est un cas où Y ne donne aucune information sur X. Par exemple, dans le cas du canal additif gaussien où $Y = X + Z$, l'information mutuelle peut être calculée en posant les hypothèses suivantes:

H_1 : les variables aléatoires X et Y sont dépendantes et $f_1(u) = f_{X,Y}(x, y)$.

H_2 : les variables aléatoires X et Y sont indépendantes et $f_2(u) = f_X(x) f_Y(y)$.

Dans ce cas, on trouve que les équations (2.3) et (2.8) sont équivalentes. Cette technique peut s'étendre à un plus grand nombre de variables ([Kul68] chap 9), comme dans le cas qui nous intéresse, c'est-à-dire où m et n sont des entiers quelconques. Formons d'abord un nouveau vecteur $\mathbf{U} = (\mathbf{X}^T, \mathbf{Y}^T)^T$ incluant toutes les variables où $\mathbf{X} = (X_1, X_2, \ldots, X_m)^T$ et $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$. Le nouveau vecteur $\mathbf{U}$ possède donc $k = m + n$ éléments et les hypothèses sont les suivantes:

H_1 : les variables aléatoires $U_1, U_2, \ldots, U_k$ sont conjointement dépendantes selon une fonction de densité de probabilité gaussienne.

H_2 : les variables aléatoires $U_1, U_2, \ldots, U_m$ sont conjointement dépendantes selon une fonction de densité de probabilité gaussienne, tout comme les variables $U_{m+1}, U_{m+2}, \ldots, U_k$, mais les deux groupes sont indépendants.

La forme générale de la fonction de densité de probabilité pour ces deux hypothèses est:

$$f_i(u_1, u_2, \ldots, u_k) = \frac{1}{|\pi \mathbf{K}_i|} \exp\left(-(\mathbf{u} - \mu_i)^H \mathbf{K}_i^{-1} (\mathbf{u} - \mu_i)\right). \qquad \text{avec } i = 1, 2 \tag{2.9}$$

où μ_i et $\mathbf{K}_i$ sont respectivement le vecteur des moyennes et la matrice des covariances. Noter que les dérivations de Kullback sont pour le cas réel seulement et que l'expression (2.9) et les dérivations qui suivent ont été modifiées afin de tenir compte du cas complexe. En utilisant l'identité $\mathbf{u}^H \mathbf{V} \mathbf{u} = tr(\mathbf{V} \mathbf{u} \mathbf{u}^H)$, on obtient:

$$\begin{aligned} \log_2 \frac{f_1(u_1, u_2, \ldots, u_k)}{f_2(u_1, u_2, \ldots, u_k)} &= \log_2 \frac{|\mathbf{K}_2|}{|\mathbf{K}_1|} - \mathrm{tr}\left[\mathbf{K}_1^{-1} (\mathbf{u} - \mu_1)(\mathbf{u} - \mu_1)^H\right] \\ &\quad - \mathrm{tr}\left[\mathbf{K}_2^{-1} (\mathbf{u} - \mu_2)(\mathbf{u} - \mu_2)^H\right] \end{aligned} \tag{2.10}$$

où $|\mathbf{K}_i|$ représente le déterminant de la matrice de covariance $\mathbf{K}_i$. L'information mutuelle est alors donnée par:

$$\begin{aligned} I(1:2) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_1(u_1, u_2, \ldots, u_k) \log_2 \frac{f_1(u_1, u_2, \ldots, u_k)}{f_2(u_1, u_2, \ldots, u_k)} du_1 \ldots du_k \\ &= \log_2 \frac{|\mathbf{K}_2|}{|\mathbf{K}_1|} + \mathrm{tr}\left[\mathbf{K}_1 (\mathbf{K}_2^{-1} - \mathbf{K}_1^{-1})\right] \\ &\quad + \mathrm{tr}\left[\mathbf{K}_2^{-1} (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T\right] \end{aligned} \tag{2.11}$$

Dans le cas du canal avec bruit additif gaussien:

$\mu_1 = \mu_2 = 0$ (Les variables sont centrées à zéro)

$\mathbf{K}_1 = \begin{pmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{pmatrix}$ (Matrice de covariance de l'hypothèse H_1)

$\mathbf{K}_2 = \begin{pmatrix} \mathbf{K}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{22} \end{pmatrix}$ (Matrice de covariance de l'hypothèse H_2)

où

$\mathbf{K}_{11} = (\sigma_{ij})$ $\quad i, j = 1, 2, \ldots, m$

$\mathbf{K}_{22} = (\sigma_{uv})$ $\quad u, v = m+1, m+2, \ldots, m+n = k$

$$\mathbf{K}_{12} = (\sigma_{iv}) = \mathbf{K}_{21}^{H}$$

où $\mathbf{K}_{21}^{H}$ est le transconjugué de la matrice $\mathbf{K}_{21}$. En remplaçant ces matrices dans l'équation (2.11), l'information moyenne en faveur de H_1 sur H_2 devient:

$$I(1:2) = \log_2 \frac{|\mathbf{K}_{11}|\,|\mathbf{K}_{22}|}{\begin{vmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{vmatrix}} \tag{2.12}$$

Finalement, utilisant l'identité $\begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{vmatrix} = |\mathbf{A}|\,|\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}|$, on obtient:

$$I(1:2) = \log_2 \frac{|\mathbf{K}_{22}|}{|\mathbf{K}_{22} - \mathbf{K}_{21}\mathbf{K}_{11}^{-1}\mathbf{K}_{12}|} \tag{2.13}$$

2.2.3 Cas où m et n sont des entiers quelconques et le canal est à évanouissements lents de Rayleigh avec bruit additif gaussien

Les dérivations et les résultats obtenus dans les sections précédentes ne peuvent s'appliquer directement au cas où le canal est à évanouissements de Rayleigh car les symboles reçus, $\mathbf{Y}$, ne suivent plus une distribution gaussienne. En effet, on a $\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{Z}$, où $\mathbf{H}$ est la matrice des coefficients d'évanouissement. Ces coefficients sont de la forme $H_R + jH_I$, où H_R et H_I sont des variables aléatoires suivant une distribution gaussienne. $\mathbf{HX}$ n'est pas gaussien et par conséquent, $\mathbf{Y}$ n'est pas gaussien. Par contre, comme les évanouissements dans le canal sont lents, on peut faire les suppositions suivantes:

1. La matrice des coefficients $\mathbf{H}$ est constante pendant la durée de transmission.

2. La réalisation de $\mathbf{H}$ pour la transmission courante est parfaitement connue au récepteur.

Dans ce cas, $\mathbf{HX}$ et $\mathbf{Y}$ sont gaussiens et les résultats des sections précédentes peuvent être utilisés. Par contre, la capacité obtenue est une variable aléatoire dépendante de la réalisation

du canal. Les matrices de covariance de l'équation (2.13) sont les suivantes:

$$\mathbf{K}_{11} = E_{\mathbf{H}}\left[\mathbf{X}\mathbf{X}^H\right] = E_{\mathbf{H}}\left[\begin{pmatrix} |X_1|^2 & & 0 \\ & \ddots & \\ 0 & & |X_m|^2 \end{pmatrix}\right] = \frac{P}{m}\mathbf{I}_m \tag{2.14}$$

$$\mathbf{K}_{22} = E_{\mathbf{H}}\left[\mathbf{Y}\mathbf{Y}^H\right] = E_{\mathbf{H}}\left[(\mathbf{H}\mathbf{X}+\mathbf{Z})(\mathbf{H}\mathbf{X}+\mathbf{Z})^H\right] = \frac{P}{m}\mathbf{H}\mathbf{H}^H + \sigma_z^2\mathbf{I}_n \tag{2.15}$$

$$\mathbf{K}_{12} = E_{\mathbf{H}}\left[\mathbf{X}\mathbf{Y}^H\right] = E_{\mathbf{H}}\left[\mathbf{X}(\mathbf{H}\mathbf{X}+\mathbf{Z})^H\right] = \frac{P}{m}\mathbf{H}^H = \mathbf{K}_{21}^H \tag{2.16}$$

Le dénominateur de l'équation (2.13) devient donc:

$$\mathbf{K}_{22} - \mathbf{K}_{21}\mathbf{K}_{11}^{-1}\mathbf{K}_{12} = \frac{P}{m}\mathbf{H}\mathbf{H}^H + \sigma_z^2\mathbf{I}_n - \frac{P}{m}\mathbf{H}\mathbf{H}^H = \sigma_z^2\mathbf{I}_n \tag{2.17}$$

La capacité par transmission, $C_{\mathbf{H}}$, telle qu'obtenue par Foschini et Gans [FG98], exprimée sous la forme d'une variable aléatoire dépendante de la matrice $\mathbf{H}$, est donnée par:

$$\begin{aligned} C_{\mathbf{H}} &= \log_2 \frac{\left|\frac{P}{m}\mathbf{H}\mathbf{H}^H - \sigma_z^2\mathbf{I}_n\right|}{|\sigma_z^2\mathbf{I}_n|} \\ &= \log_2 \left|\mathbf{I}_n + \frac{\rho}{m}\mathbf{H}\mathbf{H}^H\right| \end{aligned} \tag{2.18}$$

où $\rho = P/\sigma_z^2$ est le rapport signal-sur-bruit. Finalement, la capacité, C, du canal est donnée par l'espérance de la capacité par transmission par rapport à la matrice du canal $\mathbf{H}$:

$$C = E_{\mathbf{H}}\left[\log_2\left|\mathbf{I}_n - \frac{\rho}{m}\mathbf{H}\mathbf{H}^H\right|\right] \qquad \text{bps/Hz} \tag{2.19}$$

2.3 Conclusion

Dans ce chapitre, nous avons fait une revue des dérivations permettant d'arriver à une expression générale de la capacité pour les systèmes à antennes multiples lorsque l'état du canal n'est connu qu'au récepteur. L'expression finale de la capacité d'un canal à antennes multiples (2.19) a été obtenue par Foschini et Gans, qui ont utilisé les définitions de base de l'information relative de Kullback. Cette expression et les concepts permettant sa dérivation sont à la base des deux chapitres suivants, d'où l'importance de cette revue de littérature.

Chapitre 3

Capacité lorsque l'état du canal est connu au récepteur seulement

3.1 Introduction

Dans la majorité des systèmes de télécommunication, le canal n'est estimé qu'au récepteur et l'émetteur ne connaît donc pas l'état du canal. Ceci constitue le cas étudié au chapitre précédent et la capacité d'un tel système est alors donné par l'expression (2.19). En se basant sur cette expression, nous obtenons dans ce chapitre quelques conclusions intéressantes sur la capacité lorsque l'état du canal n'est connu qu'au récepteur. À la section 3.2, nous calculons l'asymptote de la capacité pour différentes valeurs de (m, n). Ensuite, nous étudions à la section 3.3 l'effet de la corrélation entre les antennes sur les asymptotes de capacité. Puis, à la section 3.4, nous dérivons la fonction de répartition de la capacité, où nous définissons la probabilité de coupure d'un canal à antennes multiples. Finalement, nous redérivons l'expression de la capacité à la section 3.5 pour le cas du canal à large bande.

3.2 Capacité asymptotique

La figure 3.1 trace la capacité pour différentes configurations d'antennes (m, n). Chaque point de la courbe a été obtenue en faisant la moyenne de la capacité calculée à l'aide de l'équation (2.18) pour 10.000 réalisations aléatoires de la matrice du canal $\mathbf{H}$. Plus le nombre de réalisations du canal utilisé lors du calcul de la moyenne est élevé, plus cette moyenne se rapproche de l'espérance mathématique. En observant les graphiques de capacité présentés à la figure 3.1, on remarque que la valeur de la capacité varie linéairement avec le rapport signal-sur-bruit (en dB) lorsque celui-ci est très grand. Dans cette section, on dérive la capacité asymptotique pour plusieurs configurations (m, n) des antennes de transmission et de réception. On peut démontrer que la capacité asymptotique obtenue correspond aussi à une limite inférieure de la capacité. Afin de simplifier le calcul de l'asymptote, l'équation (2.19) peut être exprimée sous une forme plus pratique. On remarque tout d'abord que le déterminant de l'équation (2.19) est de la forme $|\mathbf{B}| = |\mathbf{I}_n + \mathbf{A}^{n\times n}|$. Il est donc possible d'obtenir le déterminant de la matrice $\mathbf{B}$ en fonction des valeurs propres de $\mathbf{A}$ en utilisant quelques théorèmes matriciels [Lan69]:

Théorème 1.1: Toute matrice hermitienne est une matrice simple ([Lan69], équation (2.9.4)) et par conséquent, les matrices A et B sont simples. Or, toute matrice simple $\mathbf{A}$ peut s'exprimer sous la forme:

$$\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{-1} \tag{3.1}$$

où $\mathbf{Q}$ et $\mathbf{D}$ sont les matrices des vecteurs propres et des valeurs propres de $\mathbf{A}$ respectivement [Lan69](2.4.3). Notez que la matrice $\mathbf{Q}$ est unitaire, c'est-à-dire que $\mathbf{Q}\mathbf{Q}^H = \mathbf{I}$.

Théorème 1.2: Le déterminant de toute matrice $\mathbf{A}$ est le produit de ses valeurs propres ([Lan69], équation (2.1.2)):

$$\det(\mathbf{A}) = \prod_{i=1}^{n} d_i \tag{3.2}$$

où $\{d_i\}$ représente l'ensemble des valeurs propres de $\mathbf{A}$.

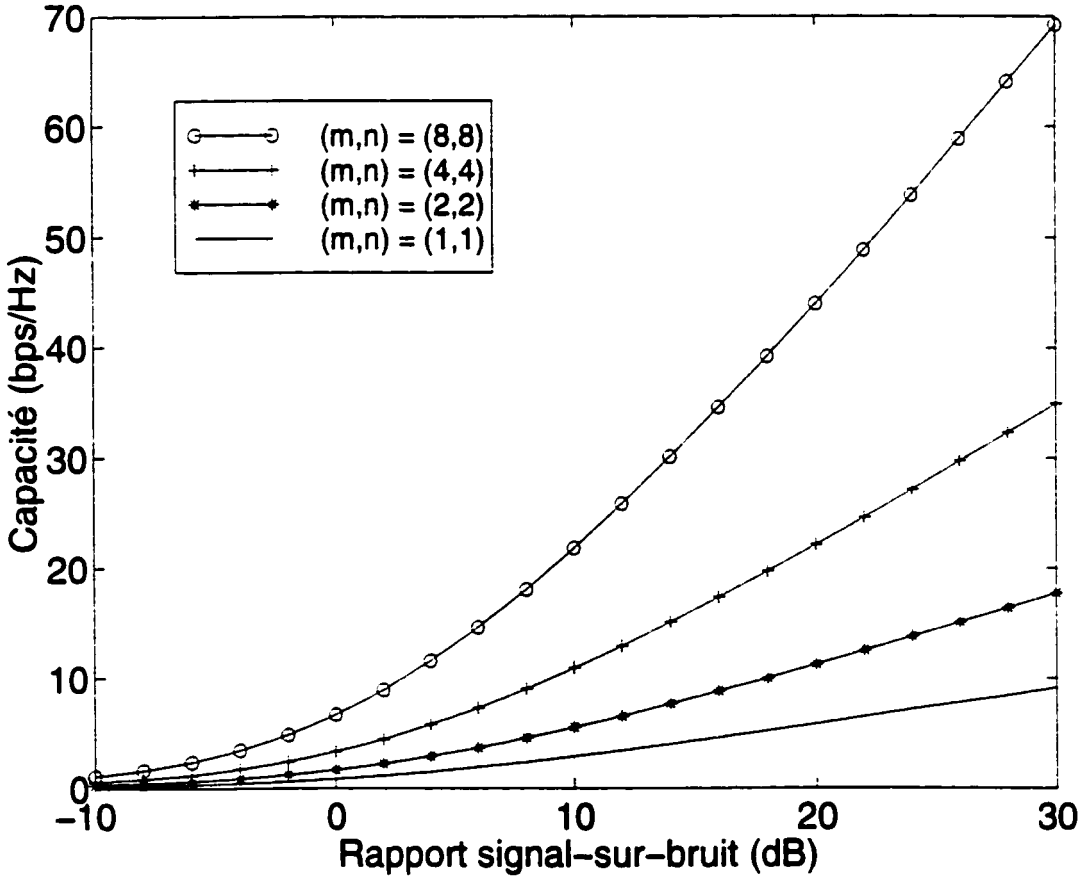


Figure 3.1: Capacité pour quelques valeurs de $(m.n)$.

En appliquant ces théorèmes, le déterminant de l'équation (2.19) devient:

$$\begin{aligned} |\mathbf{B}| &= |\mathbf{I}_n + \mathbf{A}| \\ &= |\mathbf{Q}\mathbf{I}_n\mathbf{Q}^{-1} + \mathbf{Q}\mathbf{D}\mathbf{Q}^{-1}| \\ &= |\mathbf{Q}(\mathbf{I}_n + \mathbf{D})\mathbf{Q}^{-1}| \\ &= |\mathbf{Q}\mathbf{V}\mathbf{Q}^{-1}| \end{aligned} \tag{3.3}$$

où $\mathbf{V}$ est la matrice des valeurs propres de $\mathbf{B}$. Comme $\mathbf{Q}$ est unitaire, le déterminant de $\mathbf{B}$ est alors donné par:

$$\begin{aligned} \det(\mathbf{B}) &= \det(\mathbf{V}) \\ &= \prod_{i=1}^{n}(1 - d_i) \end{aligned} \tag{3.4}$$

Or, comme le nombre de valeurs propres non nulles de $\mathbf{A}$ est égal au rang de $\mathbf{A}$, on obtient finalement:

$$\det(\mathbf{B}) = \prod_{i=1}^{rang(\mathbf{A})} (1 - d_i) \tag{3.5}$$

où les d_i, $i = 1, \ldots, rang(\mathbf{A})$, sont les valeurs propres non nulles de $\mathbf{A}$ ordonnées en valeurs décroissantes. En remplaçant $\mathbf{A}$ par $\frac{\rho}{m}\mathbf{H}\mathbf{H}^H$, la capacité devient:

$$C = E_{\mathbf{H}}\left[\log_2 \prod_{i=1}^{\min(m,n)} \left(1 - \frac{\rho}{m}\lambda_i\right)\right] \tag{3.6}$$

$$C = \sum_{i=1}^{\min(m,n)} E_{\mathbf{H}}\left[\log_2\left(1 - \frac{\rho}{m}\lambda_i\right)\right] \tag{3.7}$$

Un résultat similaire a été obtenu par Telatar [Tel99]. L'équation (3.7) montre que le système est équivalent à un système avec $\min(m, n)$ canaux parallèles de puissance P/m et de bruit équivalent σ_z^2/λ_i. La capacité asymptotique est donc ($\rho \gg 1$):

$$C_{asympt} = \sum_{i=1}^{\min(m,n)} E_{\mathbf{H}}\left[\log_2\left(\frac{\rho}{m}\lambda_i\right)\right] \tag{3.8}$$

où les λ_i sont les valeurs propres non nulles de la matrice $\mathbf{HH}^H$. On remarque que (3.8) sera toujours inférieure à (3.7) et que l'asymptote (3.8) corrrespond donc à une limite inférieure de la capacité C du système:

$$C_{asympt} < C \tag{3.9}$$

En utilisant (3.8), la capacité asymptotique est donc:

$$C_{asympt} = \min(m,n) \cdot \log_2\left(\frac{\rho}{m}\right) + \sum_{i=1}^{\min(m,n)} E_{\mathbf{H}}\left[\log_2 \lambda_i\right] \tag{3.10}$$

Cette équation correspond à une limite inférieure de la capacité dans le cas général (m,n). Afin de calculer l'espérance de (3.10), on devra utiliser la distribution des valeurs propres λ_i de $\mathbf{HH}^H$ telle qu'obtenue par Edelman [Ede89], avec $\lambda_1 > \lambda_2 \ldots > \lambda_{\min(m,n)} > 0$:

$$f_\Lambda(\lambda_1, \lambda_2, \ldots, \lambda_{\min(m,n)}) = \frac{\exp(-\sum_{i=1}^{min(m,n)} \lambda_i) \prod_{i=1}^{min(m,n)} \lambda_i^{|m-n|} \prod_{i=1, i<j}^{min(m,n)} (\lambda_i - \lambda_j)^2}{\prod_{i=1}^{\min(m,n)-1} i! \prod_{i=1}^{\min(m,n)} (\max(m,n) - i)!} \tag{3.11}$$

L'expression de capacité (3.10) et la distribution des valeurs propres (3.11) seront maintenant utilisées pour calculer la capacité asymptotique pour différentes configurations d'antennes.

3.2.1 Configuration d'antennes $(m,n) = (1,1)$ (sans diversité)

Lorsque la paire $(m,n) = (1,1)$, nous avons le cas où aucune diversité n'est utilisée ni à l'émetteur ni au récepteur. La matrice du canal n'est donc caractérisée que par un seul coefficient complexe et la distribution de la valeur propre devient simplement:

$$f_\Lambda(\lambda) = e^{-\lambda} \qquad \lambda > 0 \tag{3.12}$$

L'espérance mathématique du terme $\log_2 \lambda$ dans (3.10) devient donc:

$$E_{\mathbf{H}}\left[\log_2 \lambda\right] = \frac{1}{\ln(2)} \int_0^\infty \ln(\lambda) e^{-\lambda} d\lambda = -\frac{\gamma}{\ln(2)} \tag{3.13}$$

où $\gamma = 0.5772$ est la constante d'Euler. La capacité asymptotique est donnée par:

$$C_{asympt} = \log_2(\rho) - \frac{\gamma}{\ln(2)} \tag{3.14}$$

On constate donc que l'asymptote de la capacité d'un canal à évanouissement Rayleigh est égale à celle d'un canal AWGN ($C \approx \log_2(\rho)$) moins une constante égale à environ 0.83 bps/Hz.

3.2.2 Configuration d'antennes $min(m.n) = 1$, $\max(m.n) = t$ (diversité d'ordre t à l'émetteur ou au récepteur)

Dans ce cas, le système est équivalent à un seul canal et n'a donc qu'une seule valeur propre avec une distribution égale à:

$$f_\Lambda(\lambda) = \frac{e^{-\lambda}\lambda^{t-1}}{(t-1)!} \qquad \lambda > 0 \tag{3.15}$$

L'espérance de $\log_2 \lambda$ est donnée par:

$$\begin{aligned} E_{\mathbf{H}}[\log_2 \lambda] &= \frac{1}{\ln(2)} \int_0^\infty \ln(\lambda) \frac{e^{-\lambda}\lambda^{t-1}}{(t-1)!} d\lambda \\ &= \frac{1}{\ln(2)} \left[\sum_{i=1}^{t-1} \frac{1}{i} - \gamma \right] \end{aligned} \tag{3.16}$$

Nous allons maintenant considérer individuellement les dérivations pour les deux cas possibles: diversité de réception et diversité de transmission d'ordre t.

Cas $(1.n)$ (diversité de réception d'ordre n)

L'asymptote de la capacité pour le cas $(1.n)$ est obtenue en remplaçant (3.16) dans (3.10) avec $t = n$:

$$C_{asympt} = \frac{1}{\ln(2)} \left[\sum_{i=1}^{n-1} \frac{1}{i} + \ln(\rho) - \gamma \right] \tag{3.17}$$

L'équation (3.17) est un cas particulier du résultat obtenu par Alouini et Goldsmith ([AG99], équation (42)). L'équation (3.17) montre que l'effet du rapport signal-sur-bruit ρ et du nombre d'antennes réceptrices n sur la capacité sont indépendants pour $\rho \gg 1$. Donc, l'ajout d'antennes à un tel système aura toujours le même effet peu importe la valeur de ρ; par conséquent, le gain relatif obtenu en augmentant le nombre d'antennes de réception diminue

avec ρ. De plus, on constate que chaque antenne supplémentaire apporte un gain inférieur à la précédente. En effet, l'ajout d'une antenne apporte un gain proportionnel à $1/n$, donc plus le nombre d'antennes est élevé, plus le gain différentiel d'une antenne supplémentaire est faible. L'étude da la diversité sur la probabilité d'erreur de canaux à évanouisements de Rayleigh montre un résultat semblable [Rap96].

Cas $(m, 1)$ (diversité de transmission d'ordre m)

De la même façon, l'asymptote de la capacité pour le cas $(m, 1)$ est

$$C_{asympt} = \frac{1}{\ln(2)} \left[\sum_{i=1}^{m-1} \frac{1}{i} - \ln(m) + \ln(\rho) - \gamma \right] \tag{3.18}$$

On constate que, dû au terme $-\ln(m)$ de l'équation (3.18), le gain de capacité obtenu par l'ajout d'antennes émettrices est moindre que celui obtenu lorsque le nombre d'antennes réceptrices est augmenté. La figure 3.2 montre le gain obtenu selon le nombre d'antennes émettrices et réceptrices comparativement au cas $(m, n) = (1, 1)$. On constate que le gain tend rapidement vers une constante lorsque le nombre d'antennes émettrices augmente. En effet, lorsque le nombre d'antennes émettrices tend vers l'infini, on a:

$$\lim_{m \to \infty} \left[-\ln(m) + \sum_{i=1}^{m-1} \frac{1}{i} \right] = \gamma \approx 0.577 \tag{3.19}$$

L'asymptote de la capacité est alors donnée par:

$$\lim_{m \to \infty} C_{asympt} = \log_2(\rho) \tag{3.20}$$

On constate donc qu'un canal à évanouissement de Rayleigh est équivalent à un canal AWGN lorsque le nombre d'antennes émettrices est infini. D'après la figure 3.2, on voit que le gain atteint 90% de la limite $(\gamma / \ln(2))$ avec un nombre d'antennes émettrices relativement peu élevé $(m = 8)$. Il est donc clair que lorsque l'état du canal n'est connu qu'au récepteur, la diversité de réception offre un avantage marqué par rapport à la diversité de transmission.

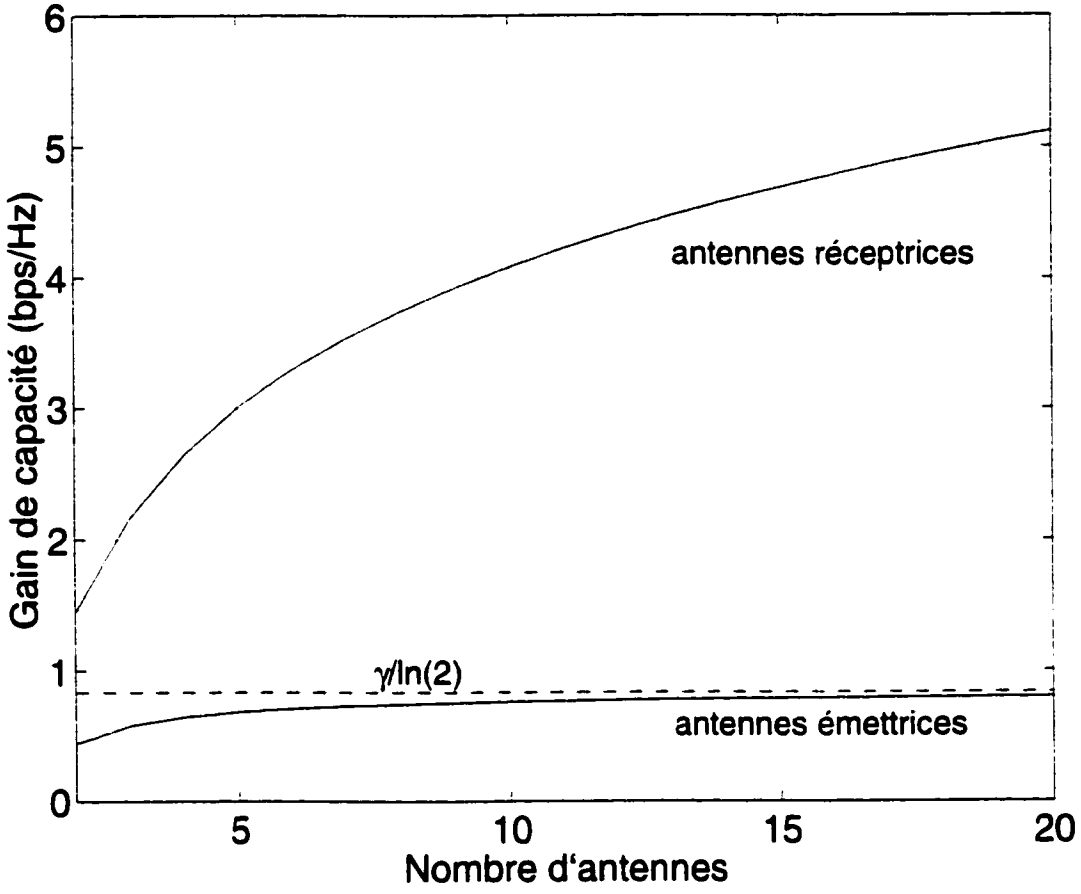


Figure 3.2: Gain sur la capacité en fonction du nombre d'antennes.

3.2.3 Diversité d'ordre supérieur à 1, $\min(m, n) > 1$ (cas général)

Afin de trouver la capacité asymptotique pour le cas $\min(m, n) > 1$, l'espérance de l'équation (3.10) a été calculée pour les cas spécifiques $\min(m, n) = 2$ et 3 (voir les calculs à l'annexe C). Une expression générale a été déterminée, puis vérifiée numériquement pour $\min(m, n) > 1$:

$$\sum_{i=1}^{\min(m,n)} E_{\mathbf{H}}\left[\log_2 \lambda_i\right] = \frac{1}{\ln(2)}\left[l - \min(m, n)\gamma\right] \tag{3.21}$$

où l est une constante dépendante des valeurs de m et de n:

$$l = \min(m, n)\sum_{i=1}^{|m-n|}\frac{1}{i} + \sum_{i=1}^{\min(m,n)-1}\frac{i}{\max(m, n) - i} \tag{3.22}$$

La capacité asymptotique est donc donnée par:

$$C_{asympt} = \min(m, n)\log_2\left(\frac{\rho}{m}\right) + \frac{1}{\ln(2)}(l - \min(m, n)\gamma) \tag{3.23}$$

où l est donnée par (3.22). Le tableau 3.1 indique quelques valeurs de l en fonction des nombres minimal et maximal d'antennes. Quelques asymptotes sont tracées à la figure 3.3 pour $(m, n) = (1, 1)$, (2,2), (4,4) et (8,8).

	$\max(m, n)$						
$\min(m, n)$	2	3	4	5	6	7	8
2	1.00	2.50	3.33	3.92	4.37	4.73	5.04
3	-	2.50	4.33	5.42	6.20	6.82	7.33
4	-	-	4.33	6.42	7.70	8.65	9.41

Tableau 3.1: Valeurs de la constante l en fonction des nombres minimal et maximal d'antennes.

Comme les valeurs propres non nulles des matrices $\mathbf{H}\mathbf{H}^H$ et $\mathbf{H}^T\mathbf{H}^{T^H}$ sont identiques (voir la preuve à l'annexe B), il est possible de calculer la différence de capacité entre les cas

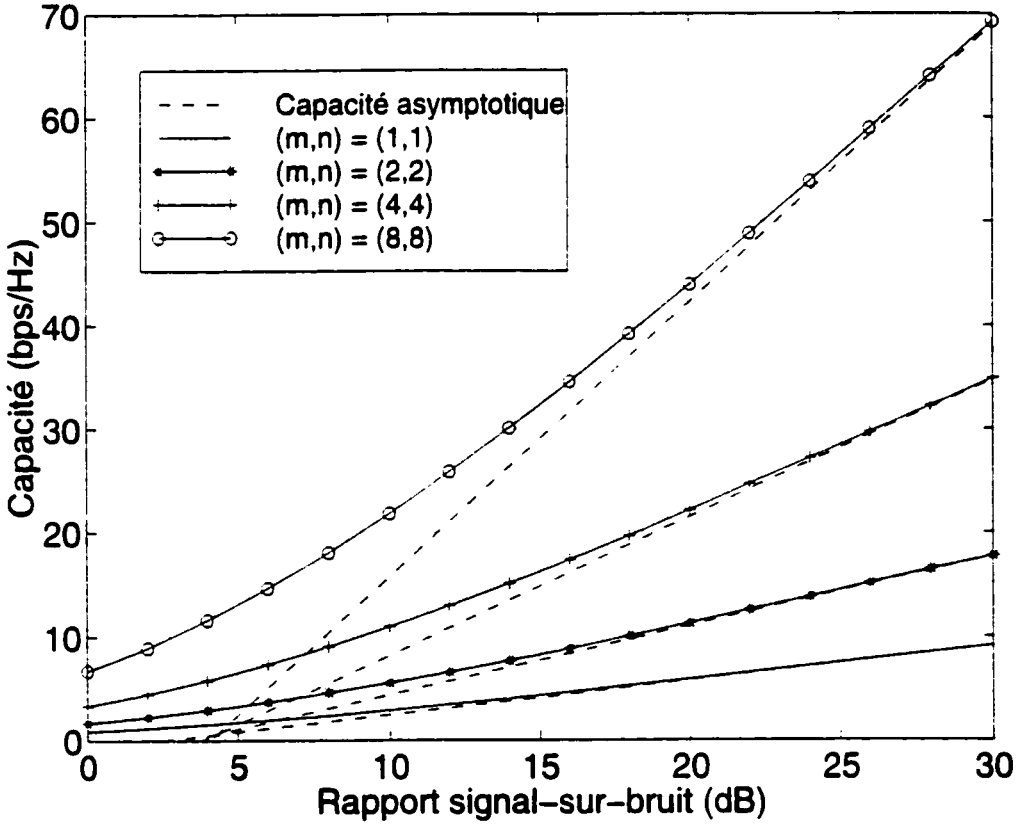


Figure 3.3: Asymptotes de la capacité $(m, n) = (1, 1)$, $(2, 2)$, $(4, 4)$ et $(8, 8)$.

$(m = t_1, n = t_2)$ et $(m = t_2, n = t_1)$ pour $\rho/m \gg 1$:

$$\begin{aligned}\Delta C_{asympt} &= \min(m,n)\cdot\log_2\left(\frac{\rho}{t_1}\right) + \sum_{i=1}^{\min(m,n)} E_{\mathbf{H}}\left[\log_2\lambda_i\right] \\ &\quad - \left[\min(m,n)\cdot\log_2\left(\frac{\rho}{t_2}\right) + \sum_{i=1}^{\min(m,n)} E_{\mathbf{H}}\left[\log_2\lambda_i\right]\right] \\ &= \min(m,n)\cdot\log_2\left(\frac{t_2}{t_1}\right) \qquad (3.24)\end{aligned}$$

Ceci confirme que l'ajout d'antennes réceptrices augmente davantage la valeur de la capacité que lorsque l'on accroît le nombre d'antennes émettrices.

3.2.4 Cas limite lorsque le nombre d'antennes est très large

D'après la loi des grands nombres [Ede89, p.79], lorsque le nombre d'antennes émettrices est très grand, la matrice $\frac{1}{m}\mathbf{H}\mathbf{H}^H$ tend vers une matrice identité $\mathbf{I}_n$ d'ordre n et ses valeurs propres tendent vers l'unité. En remplaçant $\lambda_i = m$, $i = 1, \ldots, n$ dans (3.7), on obtient donc [Tel99].(6):

$$C = n\log_2(1+\rho) \qquad (3.25)$$

Aussi, lorsque le nombre d'antennes réceptrices est très élevé, la matrice $\mathbf{H}^H\mathbf{H}$ tend vers la matrice $n\mathbf{I}_m$ et chacune de ses valeurs propres tend vers n. En remplaçant $\lambda_i = n$ pour $i = 1, \ldots, m$ dans (3.7), on obtient donc:

$$C = m\log_2\left(1+\frac{n\rho}{m}\right) \qquad (3.26)$$

En général, si le rapport $\frac{\max(m,n)}{\min(m,n)}$ est élevé, on a:

$$C = \min(m,n)\log_2\left(1+\frac{\max(m,n)\rho}{m}\right) \qquad (3.27)$$

On remarque que lorsque $\min(m,n) = 1$, l'équation (3.27) représente la capacité d'un canal avec bruit additif gaussien avec diversité soit au récepteur ou soit à l'émetteur. L'augmentation

du nombre d'antennes dans un canal avec évanouissements de Rayleigh permet donc de combattre ces évanouissements, ce qu'illustre les figures 3.4 et 3.5. Les courbes de capacité avec évanouissement de Rayleigh des figures 3.4 et 3.5 ont été obtenues par simulation tel que décrit au début du chapitre alors que les courbes de capacité pour le canal additif gaussien ont été obtenues avec l'expression (3.27).

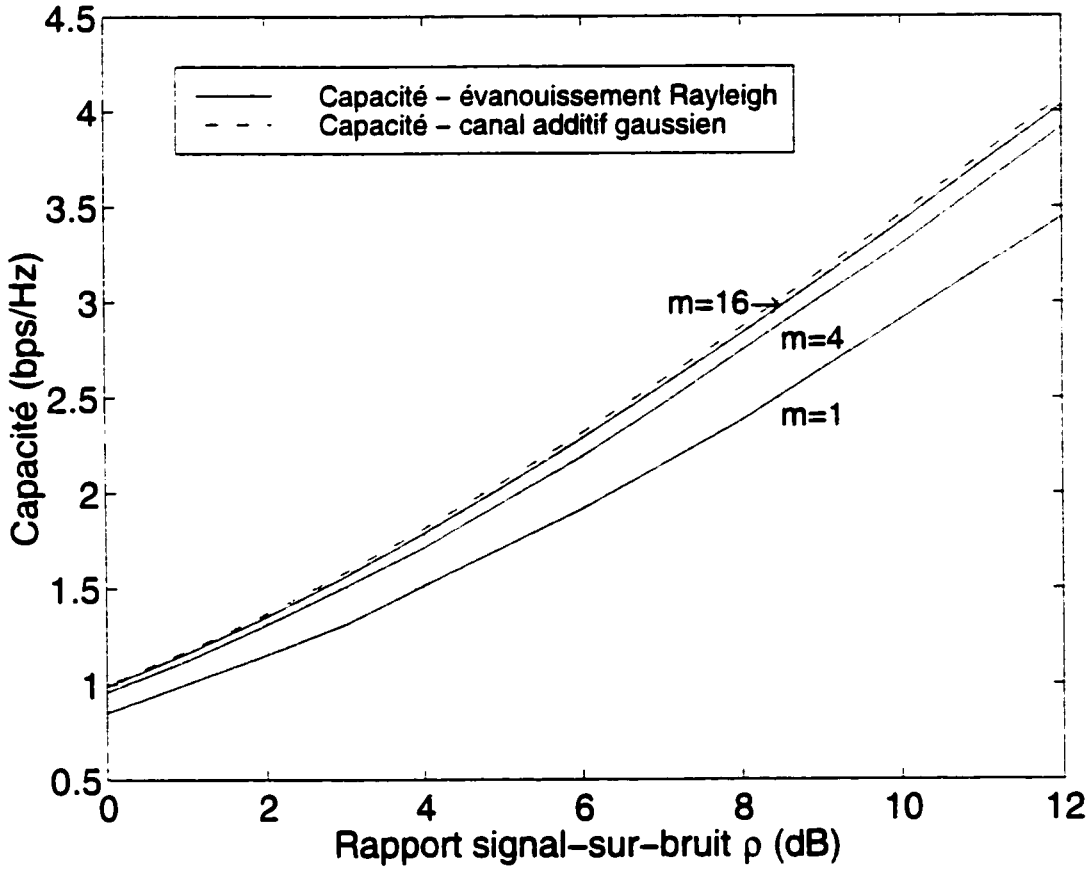


Figure 3.4: Comparaison de la capacité du canal de Rayleigh et du canal avec bruit blanc additif gaussien pour $m = 1, 4$ ou 16 antennes émettrices et une antenne réceptrice ($n = 1$).

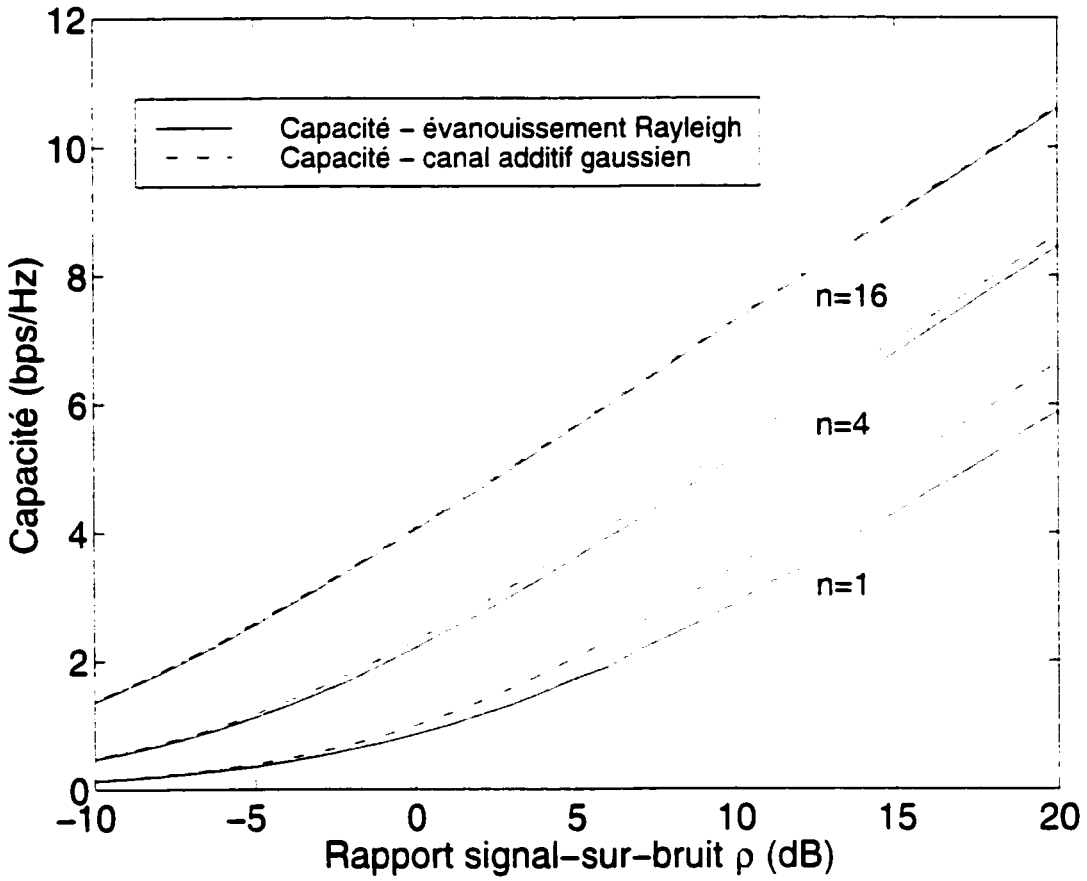


Figure 3.5: Comparaison de la capacité du canal de Rayleigh et du canal avec bruit blanc additif gaussien pour une antenne émettrice ($m = 1$) et $n = 1$, 4 ou 16 antennes réceptrices.

3.3 Capacité asymptotique avec corrélation parfaite entre les antennes

Les dérivations des expressions des asymptotes de capacité précédentes supposent que les coefficients d'évanouissement du canal sont non corrélés. En pratique, ces coefficients seront plus ou moins corrélés selon la séparation physique entre les antennes émettrices ainsi que les antennes réceptrices. Dans le cas où la distance séparant les antennes est très petite, on suppose que chaque antenne reçoit un signal subissant la même atténuation. Les expressions des asymptotes de capacité sont redérivées dans cette section pour une corrélation parfaite entre les antennes. On peut considérer que la corrélation parfaite entre les signaux provenant d'antennes distinctes constitue un pire cas pour la diversité de transmission et de réception. Comme le calcul se fait directement à partir de la matrice des coefficients d'évanouissement, l'équation générale de la capacité (2.19) est utilisée et les résultats de l'annexe A sont donc utilisés dans cette section.

3.3.1 Effet de la corrélation pour $\min(m, n) = 1$, $\max(m, n) = t$

Dans le cas où $\min(m, n) = 1$ et $\max(m, n) = t$, une corrélation parfaite entre les antennes implique que les t coefficients d'évanouissement sont tous égaux.

Capacité avec corrélation parfaite au récepteur

En posant les coefficients d'évanouissement égal à h, l'équation (2.19) devient:

$$\begin{aligned} C &= E_{\mathbf{H}}\left[\log_2\left(1+\rho n|h|^2\right)\right] \\ &= E_{\mathbf{H}}\left[\log_2\left(1+n\rho Y\right)\right] \end{aligned} \tag{3.28}$$

En remplaçant ρ par $n\rho$ dans (A.3) et (A.12), on obtient:

$$C_{asympt} = \frac{1}{\ln(2)}\left[\ln(\rho) + \ln(n) - \gamma\right] \tag{3.29}$$

La diminution de capacité due à une corrélation parfaite au récepteur est donc ((3.17)-(3.29)):

$$\Delta C_{asympt} = \frac{1}{\ln(2)} \left[\sum_{i=1}^{n-1} \frac{1}{i} - \ln(n) \right] \tag{3.30}$$

Aussi, lorsque le nombre d'antennes tend vers l'infini, cette diminution devient:

$$\lim_{n \to \infty} \Delta C_{asympt} = \frac{\gamma}{\ln(2)} \approx 0.83 \tag{3.31}$$

Suite à ces résultats, on remarque que l'augmentation de capacité causée par l'ajout d'antennes au récepteur peut être divisée en deux parties: un gain d'amplification de puissance, G_A, et un gain de diversité réelle, G_D. Comme n antennes identiques capteront une puissance n fois plus élevée qu'une seule antenne, le rapport signal-sur-bruit sera augmenté proportionellement et le gain de capacité ainsi obtenu, que l'on défini par gain d'amplification de puissance, est:

$$G_A = \log_2(n) \tag{3.32}$$

De plus, en utilisant plusieurs antennes, on réduit la probabilité que chaque antenne reçoive un signal fortement atténué, c'est ce qu'on appelle le gain de diversité réelle G_D:

$$G_D = \frac{1}{\ln(2)} \left[\sum_{i=1}^{n-1} \frac{1}{i} - \ln(n) \right] \tag{3.33}$$

L'asymptote de la capacité pour le système $(1, n)$ est donc égale à la capacité obtenue pour le cas $(1, 1)$ plus les deux types de diversité (3.17):

$$C_{asympt} = \frac{1}{\ln(2)} [\ln(\rho) - \gamma] - G_A - G_D \tag{3.34}$$

Avec corrélation parfaite au récepteur, on perd la diversité réelle mais le gain d'amplification demeure et la capacité asymptotique devient (3.29):

$$C_{asympt} = \frac{1}{\ln(2)} [\ln(\rho) - \gamma] - G_A \tag{3.35}$$

La capacité asymptotique avec et sans corrélation au récepteur pour $\min(m, n) = 1$ est tracée à la figure 3.6. On remarque que l'effet de la corrélation est croissant avec le nombre d'antennes réceptrices n. Par contre, comme il a été démontré auparavant, la réduction de capacité due à la corrélation sera au maximum égale à 0.83 bps/Hz.

Capacité avec corrélation parfaite à l'émetteur

Dans le cas où la corrélation est parfaite entre les antennes à l'émetteur, la capacité est donnée par:

$$C = E_{\mathbf{H}} \left[\log_2 \left(1 + \rho |h|^2 \right) \right] \tag{3.36}$$

La capacité dans ce cas particulier est donc la même que pour le cas $(m.n) = (1.1)$. La diminution de capacité due à la corrélation est alors ((3.18)-(3.36)):

$$\Delta C_{asympt} = \frac{1}{\ln(2)} \left[\sum_{i=1}^{m-1} \frac{1}{i} - \ln(m) \right] \tag{3.37}$$

Cette diminution est donc causée par la perte de diversité réelle et le résultat pouvait être trouvé intuitivement suite aux discussions précédentes. De plus, il n'est plus possible d'augmenter le nombre d'antennes émettrices afin de convertir le canal Rayleigh en canal gaussien puisqu'il n'existe plus de diversité réelle. En effet, l'asymptote de la capacité sans corrélation est donnée par (3.18):

$$C_{asympt} = \frac{1}{\ln(2)} \left[\ln(\rho) - \gamma \right] + G_D \tag{3.38}$$

Donc, pour les cas où $(m.1)$ et $(1.n)$, une corrélation parfaite entre les antennes réduira la capacité asymptotique au maximum de 0.83 bps/Hz. On peut donc conclure que cette diminution est relativement peu élevée pour un rapport signal-sur-bruit $\rho \gg 1$. La capacité asymptotique avec et sans corrélation à l'émetteur pour $\min(m.n) = 1$ est tracée à la figure 3.7. On remarque que l'effet de la corrélation entre les antennes émettrices croit avec m mais que la réduction de capacité due à la corrélation sera au maximum égale à 0.83 bps/Hz.

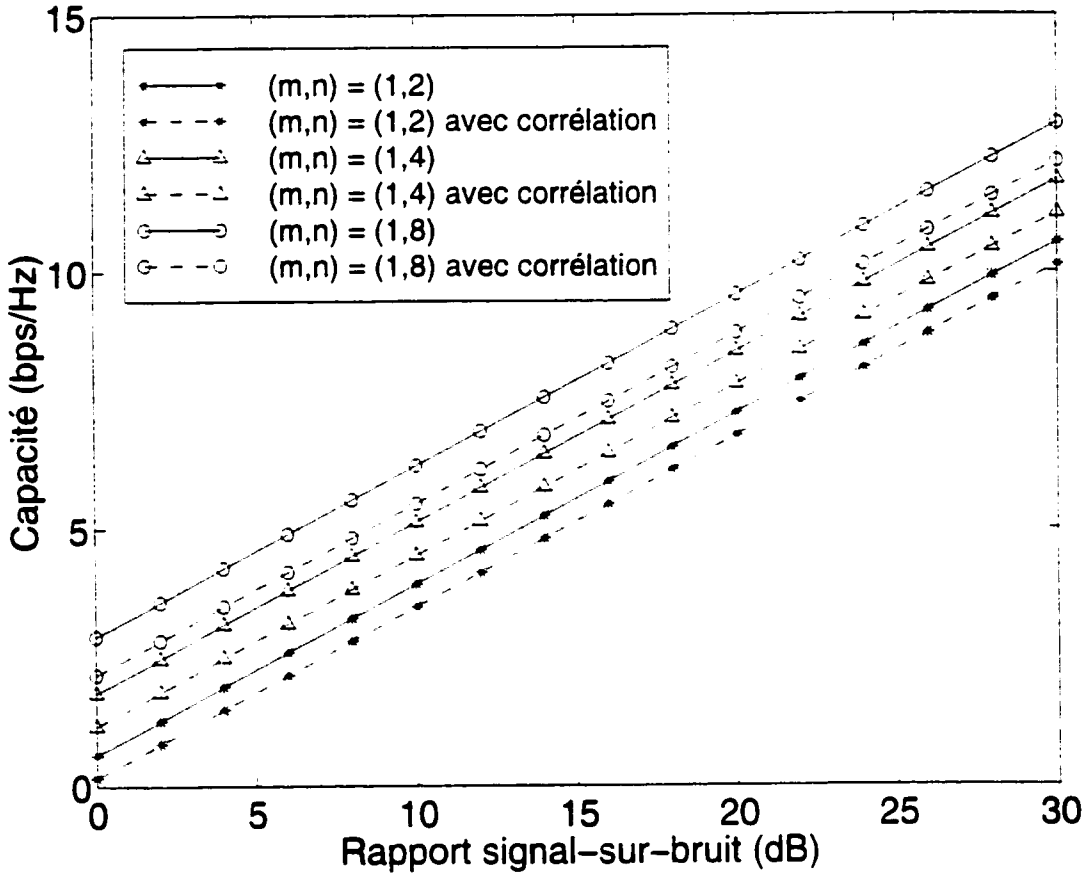


Figure 3.6: Capacité asymptotique avec et sans corrélation entre les antennes réceptrices pour $(m, n) = (1, 2)$, $(1, 4)$ et $(1, 8)$.

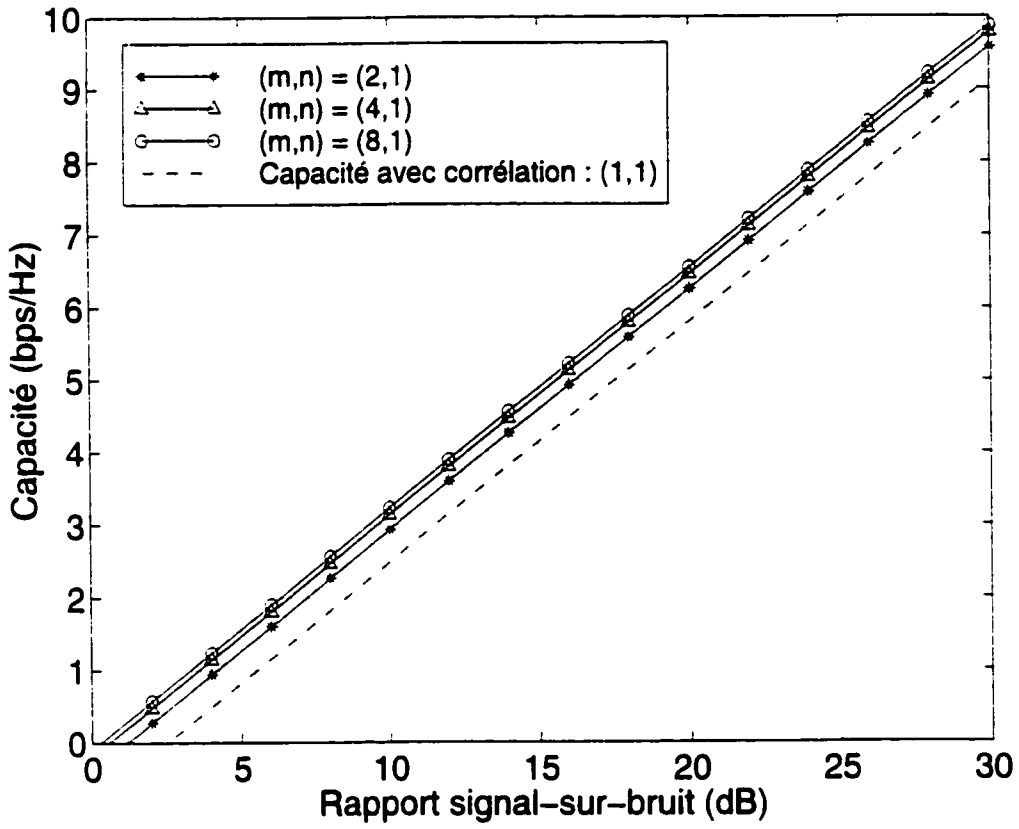


Figure 3.7: Capacité asymptotique avec et sans corrélation entre les antennes émettrices pour $(m.n) = (2.1)$. (4.1) et (8.1).

3.3.2 Effet de la corrélation pour $\min(m, n) > 1$

Corrélation à l'émetteur:

Lorsque les antennes émettrices sont parfaitement corrélées, la matrice du canal devient:

$$\mathbf{H}^{n\times m} = \begin{pmatrix} h_1 & \dots & h_1 \\ \vdots & \ddots & \vdots \\ h_n & \dots & h_n \end{pmatrix}$$

Le déterminant de l'expression (2.19) est alors:

$$\left|\mathbf{I}_n + \frac{\rho}{m}\mathbf{H}\mathbf{H}^H\right| = 1 - \rho\sum_{i=1}^{n}|h_i|^2 \tag{3.39}$$

On remarque que le déterminant est identique à l'expression (A.13). La capacité est donc donnée par (A.14) et la capacité asymptotique est alors (A.20):

$$C_{asympt} = \frac{1}{\ln(2)}\left[\sum_{i=1}^{n-1}\frac{1}{i} + \ln(\rho) - \gamma\right] \tag{3.40}$$

On remarque que l'équation (3.40) est identique à l'équation (3.17) pour le cas où il n'y a aucune diversité à l'émetteur.

Corrélation au récepteur:

Dans le cas où les antennes réceptrices sont parfaitement corrélées, la matrice du canal est donnée par

$$\mathbf{H}^{n\times m} = \begin{pmatrix} h_1 & \dots & h_m \\ \vdots & \ddots & \vdots \\ h_1 & \dots & h_m \end{pmatrix}$$

Le déterminant de l'expression (2.19) est alors:

$$\left|\mathbf{I}_n + \frac{\rho}{m}\mathbf{H}\mathbf{H}^H\right| = 1 + \frac{n\rho}{m}\sum_{i=1}^{m}|h_i|^2 \tag{3.41}$$

La capacité est obtenue en remplaçant ρ par $n\rho$ dans l'expression (A.21) et la capacité asymptotique est alors ((A.22)):

$$C_{asympt} = \frac{1}{\ln(2)} \left[\sum_{i=1}^{m-1} \frac{1}{i} + \ln\left(\frac{n}{m}\right) + \ln(\rho) - \gamma \right] \tag{3.42}$$

Dans ce cas, l'asymptote est différente de celle sans diversité au récepteur (3.18). La diminution de capacité est causée par la perte de diversité réelle alors que l'amplification de puissance reçue, $\ln(n)$ est toujours présente. La capacité asymptotique avec et sans corrélation pour $\min(m, n) > 1$ est tracée aux figures 3.8 et 3.9 pour $(m, n) = (8, 2)$, $(2, 8)$ et $(8, 8)$. On remarque que l'effet négatif de la corrélation entre antennes est plus important lorsque $m = n$. En effet, comme la pente de chaque asymptote sans corrélation est proportionnelle à $\min(m, n)\log_2(\rho)$ alors que les asymptotes avec corrélation croissent avec $\log_2(\rho)$, la plus grande différence entre les asymptotes est obtenue en maximisant $\min(m, n) - 1$, ce qui donne $m = n$.

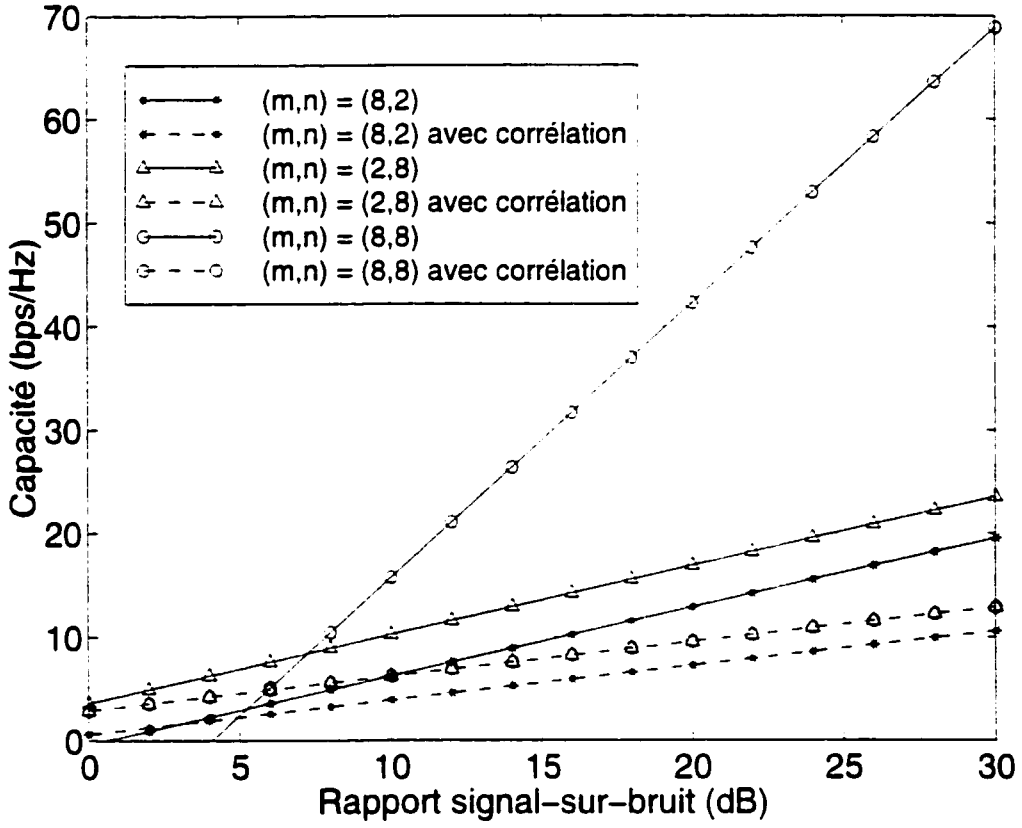


Figure 3.8: Capacité asymptotique avec et sans corrélation entre les antennes émettrices pour $(m, n) = (8, 2)$, $(2, 8)$ et $(8, 8)$.

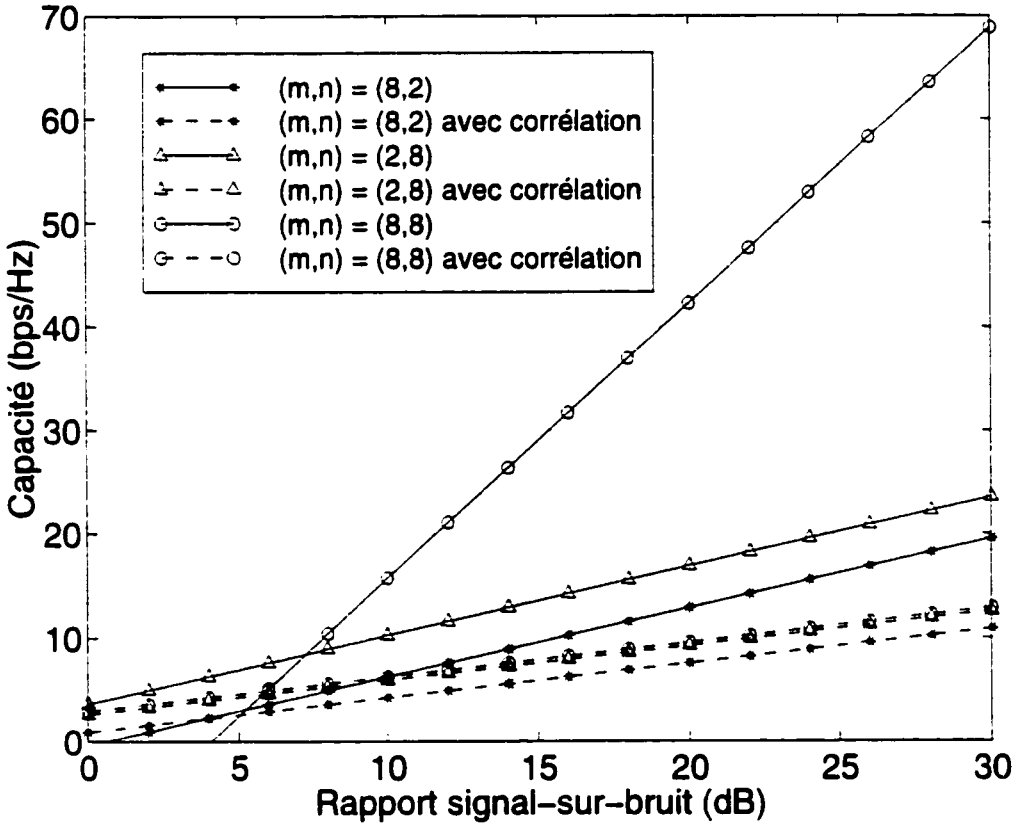


Figure 3.9: Capacité asymptotique avec et sans corrélation entre les antennes réceptrices pour $(m, n) = (8, 2)$, $(2, 8)$ et $(8, 8)$.

3.4 Fonction de répartition de la capacité

Comme discuté à la section 3.3, l'avantage de la diversité peut être décomposé en un gain d'amplification de puissance G_A et un gain de diversité G_D. On a remarqué que pour $\min(m,n) = 1$, le gain de diversité G_D sera toujours inférieur à 0.83 bps/Hz. Il semble donc que la diversité, soit à l'émetteur ou au récepteur, n'offre qu'un avantage limité. Par contre, cette conclusion ne s'applique qu'à la valeur de la capacité et non au comportement de la capacité vue comme une variable aléatoire. Nous étudions dans cette section l'effet de la diversité sur la distribution de la capacité vue comme une variable aléatoire. Les fonctions de répartition de la capacité pour différents cas peuvent être obtenues de l'équation (2.18). La capacité a été calculée pour 10000 matrices de gains complexes $\mathbf{H}$ pour les cas $(m,n) = (1,1)$, $(1,8)$, $(8,1)$ et $(8,8)$, et les distributions obtenues sont tracées à la figure 3.10 (de façon semblable à [FG98]. L'avantage d'utiliser la diversité, soit à l'émetteur ou au récepteur, est bien connu et est illustré par les distributions cumulatives $(1,8)$ et $(8,1)$. Par contre, l'amélioration obtenue par une telle diversité par rapport au cas $(1,1)$ est minime comparativement aux gains obtenus par une diversité à l'émetteur et au récepteur (cas $(8,8)$).

Les figures 3.11 et 3.12 montrent les fonctions de répartition de la capacité dans les cas $(m,1)$ et $(1,n)$ pour différentes valeurs de m et de n. On remarque que la pente de ces fonctions de répartition devient de plus en plus abrupte avec un nombre d'antennes croissant. Cela signifie que lorsque le nombre d'antennes tend vers l'infini, la capacité devient une variable déterministe. Il est en effet bien connu qu'un canal Rayleigh peut être converti en canal gaussien avec un nombre d'antennes infini. Le gain de capacité causé par l'augmentation du nombre d'antennes émettrices est relativement faible. L'avantage de l'augmentation du nombre d'antennes émettrices réside plutôt sur son effet sur la probabilité de coupure (P_c). On définit la probabilité de coupure, P_c, comme la probabilité que le débit de transmission possible (capacité) pour une réalisation du canal donnée, soit plus petit que le débit de transmission requis. En effet, on remarque (figure 3.11) que la probabilité de coupure diminue drastiquement avec l'augmentation du nombre d'antennes émettrices malgré le fait que la

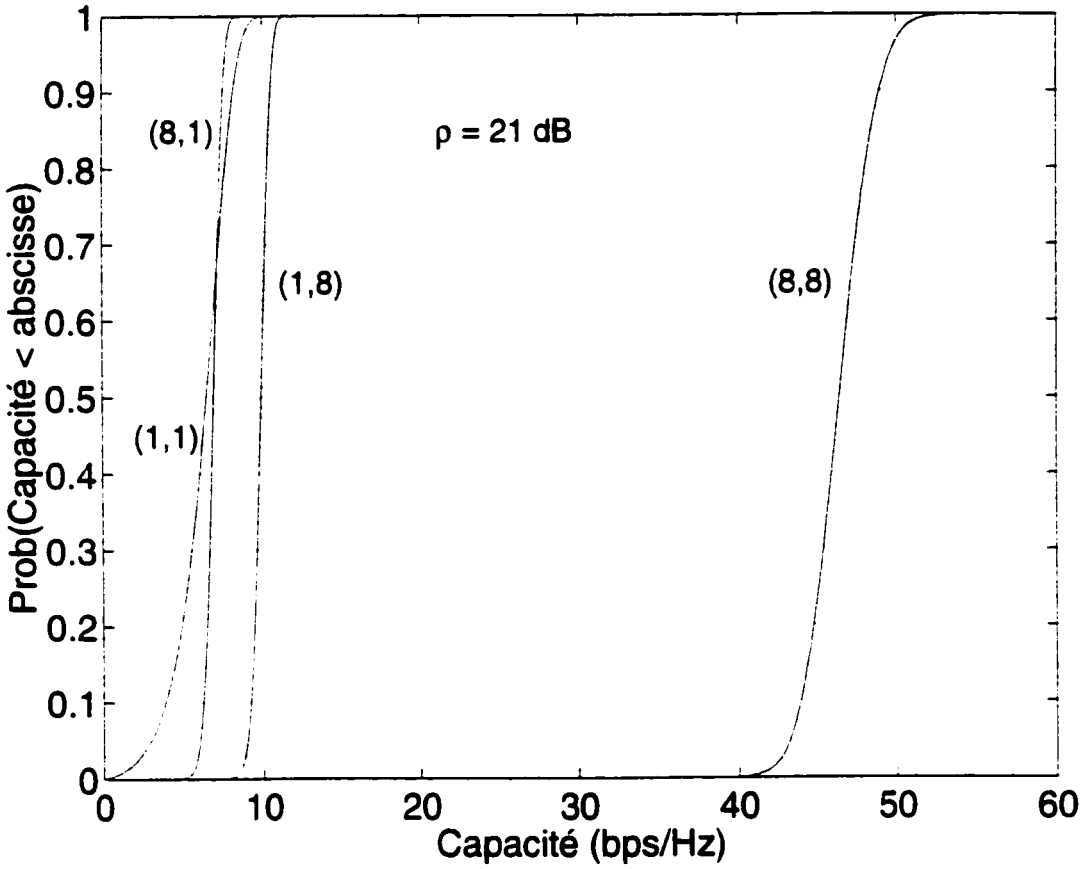


Figure 3.10: Fonction de répartition de la capacité en fonction du nombre d'antennes émettrices et réceptrices.

capacité n'augmente pas beaucoup. Par exemple, lorsque le nombre d'antennes émettrices passe de m=1 à m=16, la capacité augmente seulement de 6.5 à 7 bps/Hz. Aussi, supposons que le débit de transmission requis pour une communication soit de 6 bps/Hz. On remarque que la probabilité de coupure à 6 bps/Hz est de 40% pour m=1 alors qu'elle est seulement de 1% pour m=16.

La fonction de répartition peut être calculée facilement pour le cas $\min(m, n) = 1$, en utilisant la distribution des valeurs propres de Edelman [Ede89]. D'après l'équation (3.7), la capacité prise comme variable aléatoire est donnée par:

$$C = \log_2\left(1 + \frac{\rho}{m}\lambda\right) \tag{3.43}$$

En posant $\max(m, n) = t$, la distribution de la valeur propre λ est donnée par:

$$f_\Lambda(\lambda) = \frac{e^{-\lambda}\lambda^{t-1}}{(t-1)!} \qquad \lambda > 0 \tag{3.44}$$

La fonction de répartition de la capacité $F_C(c)$ est alors:

$$\begin{aligned} F_C(c) &= \text{Prob}(C < c) = \text{Prob}\left(\log_2(1 - \frac{\rho}{m}\lambda) < c\right) \\ &= \text{Prob}\left(\lambda < \frac{m}{\rho}(2^c - 1)\right) \\ &= \int_0^x \frac{e^{-\lambda}\lambda^{t-1}}{(t-1)!}d\lambda \quad \text{où} \quad x = \frac{m}{\rho}(2^c - 1) \\ &= 1 - e^{-x}\sum_{i=1}^{t}\frac{x^{t-i}}{(t-i)!} \\ &= 1 - e^{-\frac{m}{\rho}(2^c-1)}\sum_{i=1}^{\max(m,n)}\frac{\left(\frac{m}{\rho}(2^c-1)\right)^{\max(m,n)-i}}{(\max(m,n)-i)!} \end{aligned} \tag{3.45}$$

La fonction de densité de probabilité $f_C(c)$ est donnée par la dérivée de la fonction de répartition $F_C(c)$:

$$\begin{aligned} f_C(c) &= \frac{dF_C(c)}{dc} \\ &= \frac{m2^c\ln(2)}{\rho}\frac{\left[\frac{m}{\rho}(2^c-1)\right]^{\max(m,n)-1}}{(\max(m,n)-1)!}\exp\left[-\frac{m}{\rho}(2^c-1)\right] \end{aligned} \tag{3.46}$$

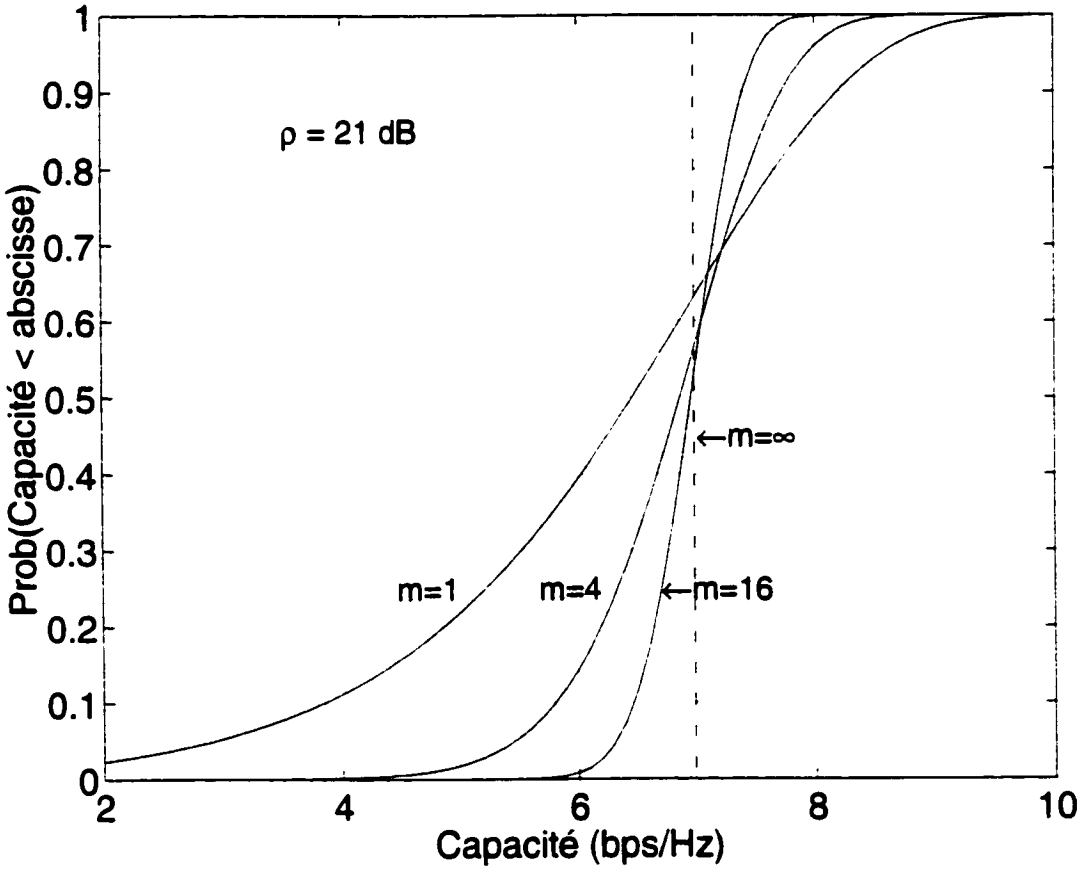


Figure 3.11: Fonction de répartition de la capacité pour $m = 1$, 4 et 16 antennes émettrices.

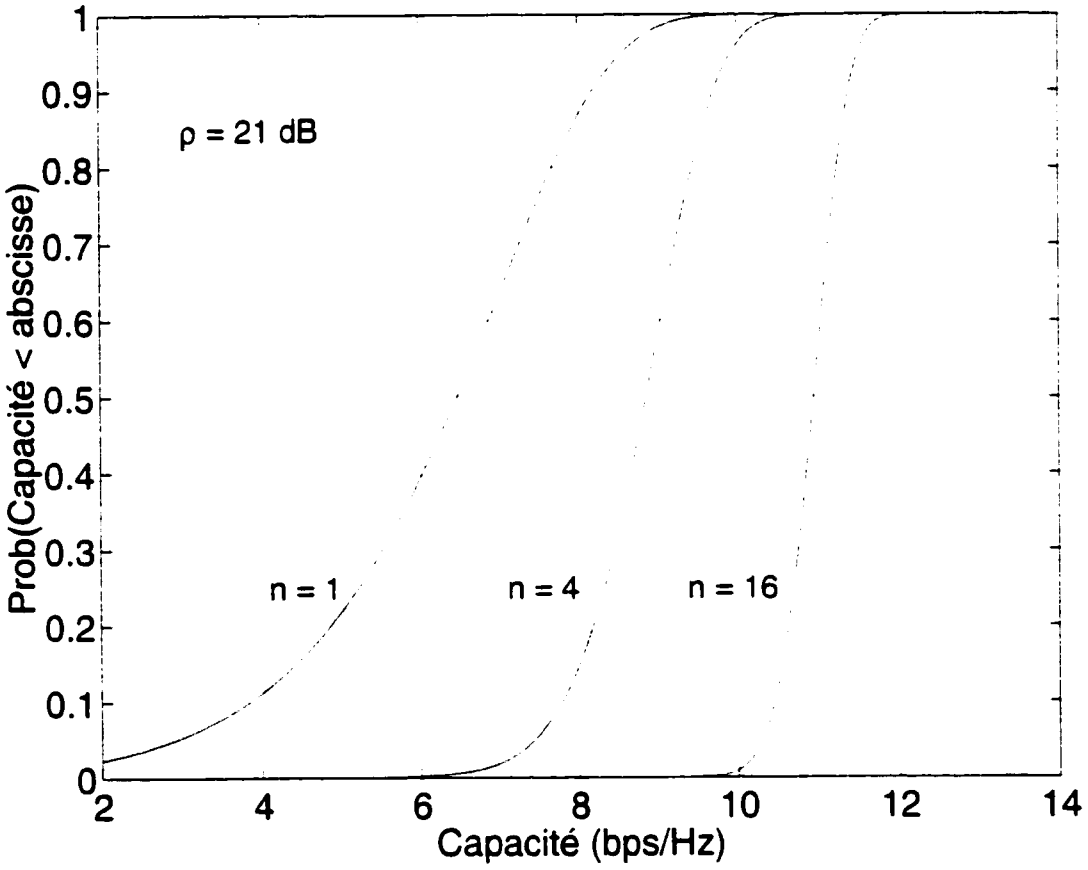


Figure 3.12: Fonction de répartition de la capacité pour $n = 1$, 4 et 16 antennes réceptrices.

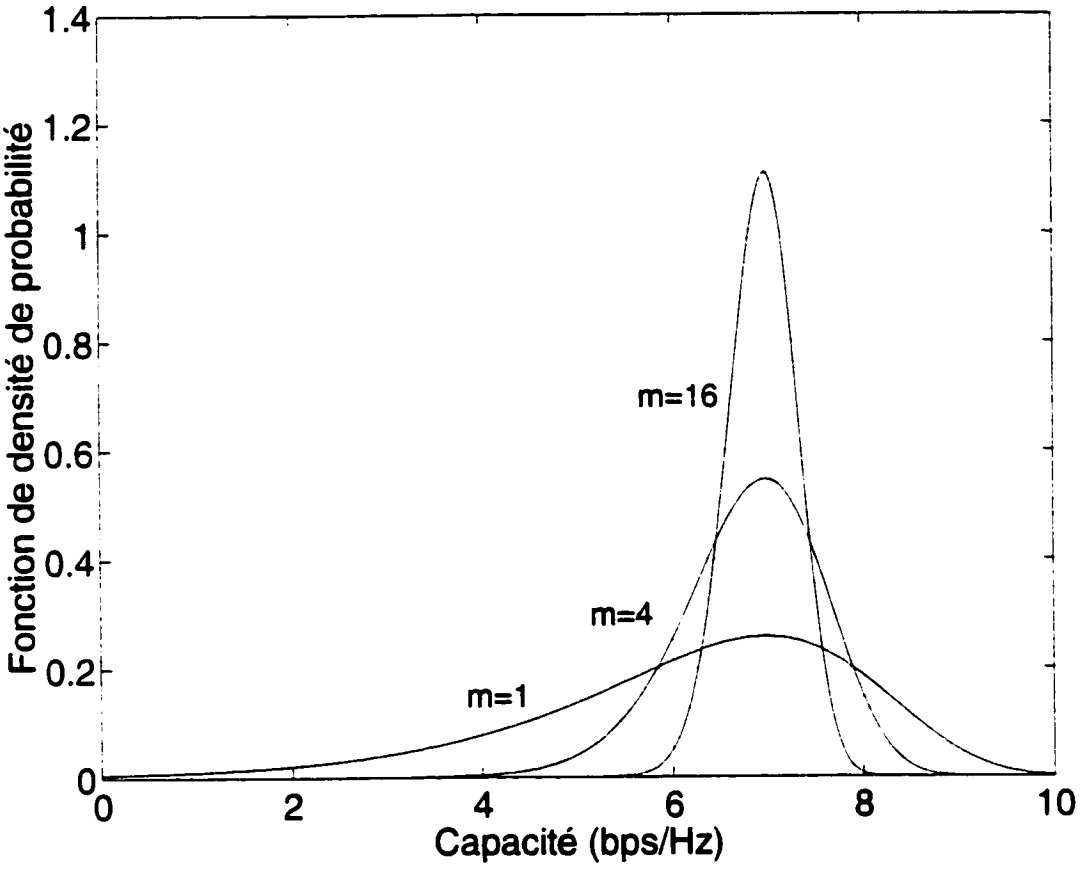


Figure 3.13: Fonction de densité de probabilité de la capacité lorsque $m = 1, 4$ et 16 antennes émettrices.

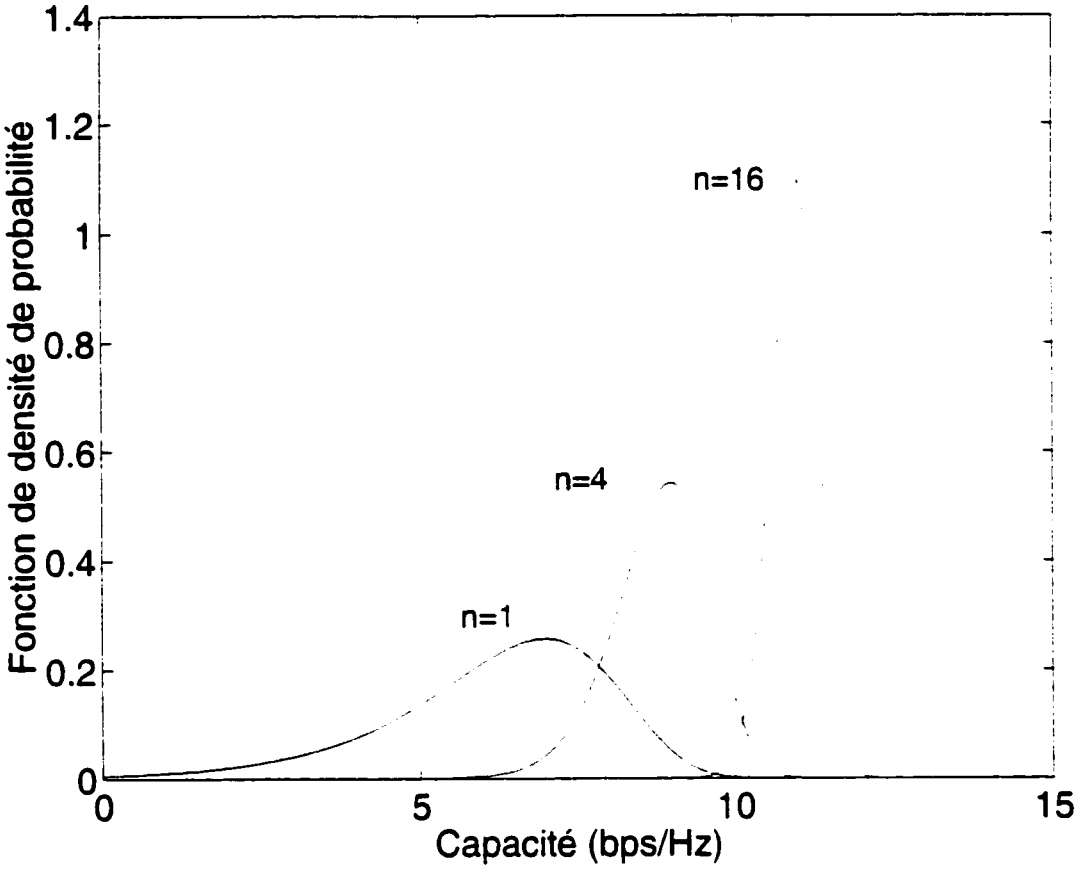


Figure 3.14: Fonction de densité de probabilité de la capacité lorsque $n = 1$, 4 et 16 antennes réceptrices.

3.5 Capacité pour un canal à large bande

Les dérivations de la capacité pour le canal à large bande sont similaires à celles effectuées dans les sections précédentes. Pour un canal à large bande, la différence se trouve dans le fait que l'on retrouve plusieurs copies retardées du même signal au récepteur. En supposant que ces copies ont chacune leur propre matrice d'évanouissement $\mathbf{H}_i$ et que les retards (ou délais) sont des multiples de T_c, la période d'utilisation du canal, on obtient:

$$\mathbf{Y}_n = \sum_{i=0}^{L} \mathbf{H}_i \mathbf{X}_{n-d_i} + \mathbf{Z}_n \tag{3.47}$$

où $L+1$ est le nombre de copies du signal et d_i est le délai. Comme les matrices $\mathbf{H}_i$ sont supposées constantes durant la transmission et que les vecteurs $\mathbf{X}_{n-d_i}$ sont des variables aléatoires gaussiennes centrées à zéro. $\mathbf{X}'_i = \mathbf{H}_i \mathbf{X}_{n-d_i}$ est aussi gaussien. De plus, comme les $\mathbf{X}_n$ sont indépendants, les $\mathbf{X}'_i$ sont aussi indépendants. Il est possible de construire un nouveau vecteur gaussien $\mathbf{W}_n = \sum_i \mathbf{X}'_i$ et l'équation (3.47) devient:

$$\mathbf{Y}_n = \mathbf{W}_n + \mathbf{Z}_n \tag{3.48}$$

Le reste des dérivations est alors très similaire à ce qui a été fait auparavant. En posant un nouveau vecteur $\mathbf{U} = (\mathbf{W}^T, \mathbf{Y}^T)^T$, on trouve que les matrices de covariance sont:

$$\begin{aligned}
\mathbf{K}_{11} &= E_{\mathbf{H}}\left[\mathbf{W}_n \mathbf{W}_n^H\right] = E_{\mathbf{H}}\left[(\mathbf{X}'_0 + \mathbf{X}'_1 + \ldots + \mathbf{X}'_L)(\mathbf{X}'_0 + \mathbf{X}'_1 + \ldots + \mathbf{X}'_L)^H\right] \\
&= \sum_{i=0}^{L} E_{\mathbf{H}}\left[\mathbf{X}'_i \mathbf{X}'^H_i\right] \\
&= \sum_{i=0}^{L} E_{\mathbf{H}}\left[(\mathbf{H}_i \mathbf{X}_{n-d_i})(\mathbf{H}_i \mathbf{X}_{n-d_i})^H\right] \\
&= \frac{P}{m} \sum_{i=0}^{L} \mathbf{H}_i \mathbf{H}_i^H
\end{aligned} \tag{3.49}$$

$$\mathbf{K}_{22} = E_{\mathbf{H}}\left[\mathbf{Y}_n \mathbf{Y}_n^H\right] = E_{\mathbf{H}}\left[(\mathbf{W}_n + \mathbf{Z}_n)(\mathbf{W}_n + \mathbf{Z}_n)^H\right] = \frac{P}{m} \sum_{i=0}^{L} \mathbf{H}_i \mathbf{H}_i^H + \sigma_z^2 \mathbf{I}_N \tag{3.50}$$

$$\mathbf{K}_{12} = E_{\mathbf{H}}\left[\mathbf{W}_n \mathbf{Y}_n^H\right] = E_{\mathbf{H}}\left[\mathbf{W}_n (\mathbf{W}_n + \mathbf{Z})^H\right] = \frac{P}{m} \sum_{i=0}^{L} \mathbf{H}_i \mathbf{H}_i^H = \mathbf{K}_{21}^H \tag{3.51}$$

Finalement, en remplaçant les matrices de covariance dans l'équation (2.13), on obtient:

$$C = E_{\mathbf{H}} \left[\log_2 \left| \mathbf{I}_n + \frac{\rho}{m} \sum_{i=0}^{L} \mathbf{H}_i \mathbf{H}_i^H \right| \right] \tag{3.52}$$

La figure 3.15 montre la capacité d'un système $(m, n) = (2, 2)$ pour les cas où le nombre de copies est 1, 2, 4 et 8. Les courbes de capacité ont été obtenues en faisant la moyenne de la capacité calculée à l'aide de l'équation (3.52) pour 10.000 réalisations aléatoires de la matrice du canal $\mathbf{H}$. La puissance moyenne est supposée la même pour toutes les copies. Aussi, le système équivalent à bande étroite est indiqué pour des fins de comparaison. Par exemple, on constate qu'un système (8,8) à large bande où le nombre de copies est 4 est équivalent à un système (8,32) à bande étroite. Il faut cependant noter que ceci est dû au fait que les copies ont toutes la même puissance moyenne.

3.6 Conclusion

En se basant sur les conclusions du chapitre deux sur les notions de base de la capacité, nous avons tout d'abord dérivé la capacité asymptotique des systèmes à antennes multiples. Les expressions de capacité asymptotique nous ont permis d'arriver à quelques conclusions intéressantes. Tout d'abord, le gain de capacité obtenu en augmentant le nombre d'antennes peut être décomposé en un gain d'amplification et un gain de diversité. Le gain de diversité obtenu en ajoutant une antenne supplémentaire n'est pas constant mais diminue avec le nombre d'antennes. Donc, pratiquement, l'ajout d'antennes devient moins efficace à mesure que le nombre d'antennes augmente. Ce phénomène est facilement identifiable lorsque le nombre d'antennes émettrices seulement est augmenté, puisque dans ce cas on a démontré qu'il n'y pas de gain d'amplification. En effet, la capacité atteint 90% de son maximum avec seulement 8 antennes émettrices. Comme l'ajout d'antennes réceptrices cause un gain d'amplification en plus du gain de diversité, on a aussi pu démontrer qu'il est plus efficace d'augmenter le nombre d'antennes réceptrices que le nombre d'antennes émettrices lorsque le canal n'est pas connu à l'émetteur. Nous avons ensuite étudié l'effet de la corrélation

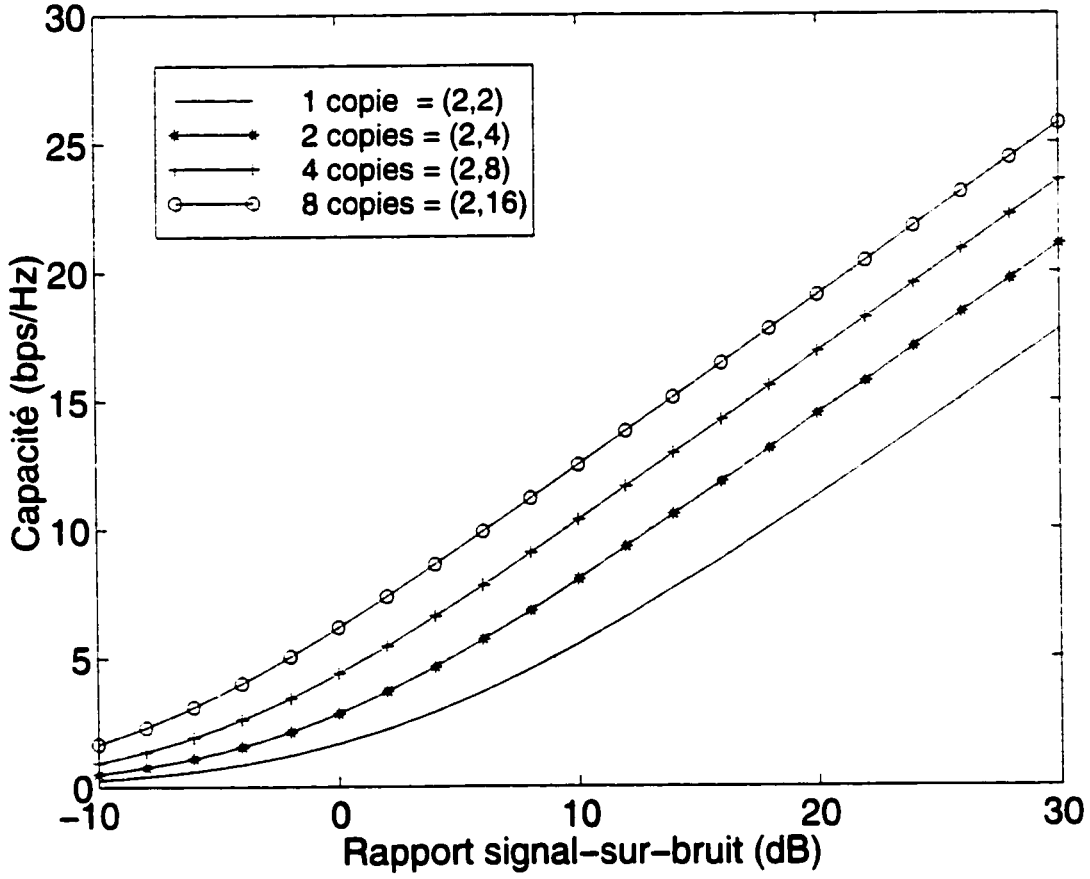


Figure 3.15: Capacité pour un canal à large bande pour un système $(m.n) = (2.2)$ avec 1, 2, 4 et 8 copies du signal.

entre antennes sur la capacité, ce qui nous a révélé que seulement le gain de diversité est affecté par cette corrélation. La corrélation entraine donc une diminution de capacité mineure lorsque la diversité spatiale est utilisée à l'émetteur ou au récepteur seulement, puisque le gain d'amplification au récepteur reste. En effet, on a démontré que le gain de diversité sera au maximum de 0.83 bps/Hz dans ce cas. Par contre, lorsque la diversité est utilisée à l'émetteur et au récepteur, la corrélation aura un effet important et la perte de capacité résultante croît rapidement avec le rapport signal-sur-bruit et le nombre d'antennes. Puis, nous avons dérivé la fonction de répartition de la capacité pour le cas où $\min(m, n) = 1$, ce qui nous a permis de définir la probabilité de coupure d'un canal comme la probabilité que le débit de transmission possible pour une réalisation du canal donnée soit plus petit que le débit de transmission requis. Nous avons alors remarqué que la diversité réduisait de façon considérable cette probabilité de coupure. Finalement, les dérivations du calcul de la capacité ont été reprises pour le canal à bande large. En supposant que les l copies d'un même signal d'un canal à large bande possède la même puissance, on arrive à la conclusion qu'un canal (m, n) à large bande est équivalent à un canal $(m, l \times n)$ à bande étroite.

Chapitre 4

Capacité lorsque le canal est connu au récepteur et à l'émetteur

4.1 Introduction

Dans le chapitre précédent, des expressions de capacité ont été présentées pour des canaux à évanouissements lents de Rayleigh dans la situation où la matrice des gains complexes du canal $\mathbf{H}$ était connue au récepteur. Dans certaines situations, il est possible que l'état du canal soit aussi connu de l'émetteur. Par exemple, lorsque le multiplexage temporel est utilisé entre deux stations et que le temps de transmission alloué à chaque station est assez court comparé au taux de changement du canal, l'estimation de l'état du canal lors de la réception peut être utilisée lors de la transmission subséquente. Comme le canal est connu à l'émetteur, celui-ci peut utiliser cette information à son avantage afin de modifier le vecteur transmis en l'adaptant aux conditions présentes du canal.

Selon la théorie sur les variables gaussiennes, il est possible d'appliquer une transformation linéaire à un vecteur de variables gaussiennes indépendantes afin d'obtenir un nouveau vecteur ayant une matrice de covariance désirée [LGon]. Le principe consiste à démultiplexer la séquence des symboles à émettre en un vecteur $\mathbf{X}$. On effectue ensuite une adaptation au canal sous la forme d'une transformation linéaire du vecteur $\mathbf{X}$ en un nouveau

vecteur $\mathbf{W}$, tel qu'illustré à la figure 4.1. Le vecteur $\mathbf{W}$ est alors une variable gaussienne complexe de dimension m possédant une matrice de covariance $\mathbf{C_W}$. Le problème se pose alors comme suit: comment peut-on choisir la matrice $\mathbf{C_W}$ afin de maximiser l'information mutuelle entre $\mathbf{W}$ et $\mathbf{Y}$?

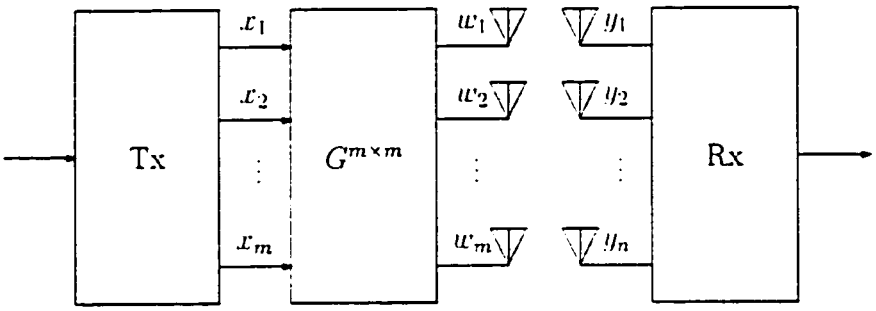


Figure 4.1: Système à antennes multiples avec prétraitement.

Nous répondons à cette question dans ce chapitre en présentant tout d'abord l'expression de capacité avec prétraitement à la section 4.2. Ensuite, la maximisation de cette expression est dérivée, ce qui nous permet alors d'obtenir à la section 4.3 la matrice de covariance $\mathbf{C_W}$ et la transformation linéaire $\mathbf{G}$ correspondante. Puis, tout comme au chapitre 3, les asymptotes de capacité sont dérivées à la section 4.4. On calcule ensuite à la section 4.5 le gain obtenu du prétraitement à l'émetteur. Finalement, une preuve de la symmétrie de la capacité est dérivée à la section 4.6.

4.2 Expression générale de la capacité avec prétraitement

En modifiant les expressions de capacité dérivées au chapitre précédent, il est possible d'exprimer la capacité avec pré-traitement en fonction de $\mathbf{C_W}$. Les nouvelles matrices de

covariance correspondantes sont:

$$\mathbf{K}_{11} = E_{\mathbf{H}}\left[\mathbf{W}\mathbf{W}^H\right] = \mathbf{C}_{\mathbf{W}} \tag{4.1}$$

$$\mathbf{K}_{22} = E_{\mathbf{H}}\left[\mathbf{Y}\mathbf{Y}^H\right] = E_{\mathbf{H}}\left[(\mathbf{H}\mathbf{W}+\mathbf{Z})(\mathbf{H}\mathbf{W}+\mathbf{Z})^H\right] = \mathbf{H}\mathbf{C}_{\mathbf{W}}\mathbf{H}^H + \sigma_z^2\mathbf{I}_n \tag{4.2}$$

$$\mathbf{K}_{12} = E_{\mathbf{H}}\left[\mathbf{W}\mathbf{Y}^H\right] = E_{\mathbf{H}}\left[\mathbf{W}(\mathbf{H}\mathbf{W}+\mathbf{Z})^H\right] = \mathbf{C}_{\mathbf{W}}\mathbf{H}^H = \mathbf{K}_{21}^H \tag{4.3}$$

La capacité est alors donnée par:

$$C = \max_{\mathbf{C}_{\mathbf{W}}} E_{\mathbf{H}}\left[\log_2\left|\mathbf{I}_n + \frac{1}{\sigma_z^2}\mathbf{H}\mathbf{C}_{\mathbf{W}}\mathbf{H}^H\right|\right] \qquad \text{(bps/Hz)} \tag{4.4}$$

Le problème du calcul de la capacité se résume donc à maximiser l'information mutuelle en fonction de la matrice de covariance $\mathbf{C}_{\mathbf{W}}$, avec la contrainte que la puissance moyenne de $\mathbf{W}$ doit être égale à celle de $\mathbf{X}$, c'est-à-dire:

$$P = E\left[\mathbf{X}^H\mathbf{X}\right] = E\left[\mathbf{W}^H\mathbf{W}\right] = \sum_{i=1}^{m} E\left[|W_i|^2\right] = \mathrm{tr}(\mathbf{C}_{\mathbf{W}}) \tag{4.5}$$

En représentant la transformation linéaire par une matrice $\mathbf{G}$, le vecteur transmis est $\mathbf{W} = \mathbf{G}\mathbf{X}$ et on obtient comme matrice de covariance $\mathbf{C}_{\mathbf{W}}$:

$$\mathbf{C}_{\mathbf{W}} = E\left[\mathbf{W}\mathbf{W}^H\right] = E\left[\mathbf{G}\mathbf{X}\mathbf{X}^H\mathbf{G}^H\right] = \frac{P}{m}\mathbf{G}\mathbf{G}^H \tag{4.6}$$

On trouve alors que la capacité s'exprime:

$$C = \max_{\mathbf{C}_{\mathbf{W}}} E_{\mathbf{H}}\left[\log_2\left|\mathbf{I}_n + \frac{\rho}{m}\mathbf{H}\mathbf{G}\mathbf{G}^H\mathbf{H}^H\right|\right] \tag{4.7}$$

Dans ce cas, la contrainte de puissance égale peut s'exprimer comme suit:

$$P = E\left[\mathbf{X}^H\mathbf{X}\right] = E\left[\mathbf{W}^H\mathbf{W}\right] = E\left[\mathbf{X}^H\mathbf{G}^H\mathbf{G}\mathbf{X}\right] \tag{4.8}$$

En posant que les symboles X_i sont indépendants et de puissance égale.

$$E\left[X_i^* X_j\right] = \begin{cases} \frac{P}{m} & i = j \\ 0 & i \neq j \end{cases} \tag{4.9}$$

on obtient:

$$E\left[\mathbf{X}^H\mathbf{G}^H\mathbf{G}\mathbf{X}\right] = \frac{P}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}|g_{ij}|^2 \tag{4.10}$$

La contrainte équivalente est donc:

$$\sum_{i=1}^{m}\sum_{j=1}^{m}|g_{ij}|^2 = \|\mathbf{G}\|_f^2 = m \tag{4.11}$$

où $\|\cdot\|_f$ est la norme de Frobinius.

4.2.1 Cas particulier de la contrainte sur la puissance

Un cas particulier de la contrainte donnée par (4.11) est obtenu en forçant la puissance de $\mathbf{W}$ à être égale à celle de $\mathbf{X}$ en tout temps, c'est-à-dire:

$$\mathbf{X}^H\mathbf{X} = \mathbf{W}^H\mathbf{W} \tag{4.12}$$

Dans ce cas, la contrainte est équivalente à dire que $\mathbf{G}$ doit être une matrice unitaire:

$$\mathbf{X}^H\mathbf{X} = \mathbf{W}^H\mathbf{W} = \mathbf{X}^H\mathbf{G}^H\mathbf{G}\mathbf{X} \tag{4.13}$$

et donc,

$$\mathbf{G}^H\mathbf{G} = \mathbf{I}_m = \mathbf{G}\mathbf{G}^H \tag{4.14}$$

En remplaçant l'équation (4.14) dans l'équation de capacité (4.7), on obtient:

$$C = E_{\mathbf{H}}\left[\log_2\left|\mathbf{I}_n + \frac{\rho}{m}\mathbf{H}\mathbf{H}^H\right|\right] \tag{4.15}$$

ρ étant le rapport signal-sur-bruit. La capacité est alors la même que si le canal n'est connu qu'au récepteur seulement et n'offre aucun avantage au niveau de la capacité. Cependant, comme la seule contrainte sur la matrice $\mathbf{G}$ est d'être unitaire, il est possible d'appliquer cette transformation sans connaitre le canal. Il serait donc intéressant de vérifier si ce genre de transformation peut offrir un avantage quelconque au récepteur (autre que sur la capacité) lorsque le canal n'est pas connu de l'émetteur.

4.3 Maximisation de l'information mutuelle

Dans cette section, on s'intéresse à déterminer la matrice de covariance $\mathbf{C_W}$, et par conséquent la matrice $\mathbf{G}$, qui maximise l'équation (4.4). Comme le logarithme est une fonction monotonique croissante, la maximisation de l'équation de capacité est équivalente à maximiser le déterminant seulement:

$$\max_{\mathbf{C_W}} \left| \mathbf{I}_n + \frac{1}{\sigma_z^2} \mathbf{H} \mathbf{C_W} \mathbf{H}^H \right| \tag{4.16}$$

avec la contrainte:

$$\mathrm{tr}(\mathbf{C_W}) = P \tag{4.17}$$

Cette complexe maximisation peut être décomposée en deux maximisations plus simples. En effet, comme la matrice $\mathbf{C_W}$ est hermitienne, on peut la décomposer sous la forme:

$$\mathbf{C_W} = \mathbf{Q} \mathbf{D} \mathbf{Q}^{-1} \tag{4.18}$$

où $\mathbf{Q}$ et $\mathbf{D}$ sont les matrices des vecteurs propres et des valeurs propres de $\mathbf{C_W}$ respectivement ([Lan69], équation (2.4.3)). Aussi, la décomposition en valeurs singulières de la matrice du canal $\mathbf{H}$ donne:

$$\mathbf{H} = \mathbf{U} \mathbf{S} \mathbf{V} \tag{4.19}$$

Le déterminant de l'équation (4.16) est alors:

$$\begin{aligned} \left| \mathbf{I}_n + \frac{1}{\sigma_z^2} \mathbf{H} \mathbf{C_W} \mathbf{H}^H \right| &= \left| \mathbf{I}_n + \frac{1}{\sigma_z^2} \mathbf{U} \mathbf{S} \mathbf{V} \mathbf{Q} \mathbf{D} \mathbf{Q}^{-1} (\mathbf{U} \mathbf{S} \mathbf{V})^H \right| \\ &= \left| \mathbf{U} \mathbf{I}_n \mathbf{U}^{-1} + \frac{1}{\sigma_z^2} \mathbf{U} \mathbf{S} \mathbf{V} \mathbf{Q} \mathbf{D} \mathbf{Q}^{-1} \mathbf{V}^{-1} \mathbf{S}^H \mathbf{U}^{-1} \right| \\ &= \left| \mathbf{U} \left(\mathbf{I}_n + \frac{1}{\sigma_z^2} \mathbf{S} \mathbf{V} \mathbf{Q} \mathbf{D} \mathbf{Q}^{-1} \mathbf{V}^{-1} \mathbf{S}^H \right) \mathbf{U}^{-1} \right| \\ &= \left| \mathbf{I}_n + \frac{1}{\sigma_z^2} \mathbf{S} \mathbf{V} \mathbf{Q} \mathbf{D} \mathbf{Q}^{-1} \mathbf{V}^{-1} \mathbf{S}^H \right| \end{aligned} \tag{4.20}$$

La maximisation de l'information mutuelle peut alors s'exprimer de la manière suivante:

$$\max_{\mathbf{C_W}} \left| \mathbf{I}_n + \frac{1}{\sigma_z^2} \mathbf{H} \mathbf{C_W} \mathbf{H}^H \right| = \max_{\mathbf{D},\mathbf{Q}} \left| \mathbf{I}_n + \frac{1}{\sigma_z^2} \mathbf{S} \mathbf{V} \mathbf{Q} \mathbf{D} \mathbf{Q}^{-1} \mathbf{V}^{-1} \mathbf{S}^H \right| \tag{4.21}$$

avec les contraintes:

$$\mathrm{tr}(\mathbf{D}) = P \tag{4.22}$$

et $\mathbf{Q}$ doit être unitaire.

4.3.1 Maximisation sur la matrice des vecteurs propres Q

La maximisation posée à l'équation (4.21) est de la forme:

$$\max_{\mathbf{A}} |\mathbf{I}_n - \mathbf{A}| = \max_{\mathbf{B}} |\mathbf{B}| \tag{4.23}$$

où $\mathbf{A}$ et $\mathbf{B}$ sont hermitiques et positives semi-définies. En effet, la matrice $\mathbf{A}$ est positive semi-définie si elle satisfait la condition:

$$\mathbf{x}\mathbf{A}\mathbf{x}^H \geq 0 \qquad \forall \mathbf{x} \tag{4.24}$$

où $\mathbf{x}$ est un vecteur ligne. Comme $\mathbf{A}$ est de la forme $\mathbf{T}\mathbf{T}^H$, on peut vérifier que:

$$\mathbf{x}\mathbf{A}\mathbf{x}^H = \mathbf{x}\mathbf{T}\mathbf{T}^H\mathbf{x}^H = \mathbf{y}\mathbf{y}^H = \sum_{i=1}^{n} |y_i|^2 \geq 0 \tag{4.25}$$

Par la suite, on peut vérifier que $\mathbf{B}$ est aussi positive semi-définie:

$$\mathbf{x}\mathbf{B}\mathbf{x}^H = \mathbf{x}\mathbf{I}\mathbf{x}^H + \mathbf{x}\mathbf{A}\mathbf{x}^H = \mathbf{x}\mathbf{x}^H - \mathbf{y}\mathbf{y}^H = \sum_{i=1}^{n} \left(|x_i|^2 - |y_i|^2\right) \geq 0 \tag{4.26}$$

Or, l'inégalité d'Hadamard [CT91] affirme que le déterminant d'une matrice positive semi-définie est plus petit ou égal au produit des éléments de sa diagonale:

$$|\mathbf{B}| \leq \prod_{i=1}^{n} b_{ii} \tag{4.27}$$

avec égalité si et seulement si $\mathbf{B}$ est diagonale. Par conséquent, $\mathbf{A}$ doit être diagonale et on doit donc poser $\mathbf{Q} = \mathbf{V}^{-1}$ dans (4.21), ce qui donne:

$$\begin{aligned} \left|\mathbf{I}_n + \frac{1}{\sigma_\xi^2}\mathbf{H}\mathbf{C}_\mathbf{W}\mathbf{H}^H\right| &= \left|\mathbf{I}_n - \frac{1}{\sigma_\xi^2}\mathbf{S}\mathbf{D}\mathbf{S}^H\right| \\ &= \prod_{i=1}^{\min(m,n)} \left(1 - \frac{1}{\sigma_\xi^2}|s_{ii}|^2 d_i\right) \end{aligned} \tag{4.28}$$

où les s_i et d_i sont les éléments de la diagonale principale de $\mathbf{S}$ et $\mathbf{D}$ respectivement. La limite supérieure du produit de l'équation (4.28) correspond au nombre de $|s_i|^2$ non nuls, qui est égal au nombre minimum de colonnes et de lignes de $\mathbf{S}^{n\times m}$, ou $\min(m, n)$. Or, on a [Lan69]:

$$|s_i|^2 = \lambda_i \tag{4.29}$$

où les λ_i sont les valeurs propres non nulles de $\mathbf{H}^H\mathbf{H}$. On obtient donc la maximisation équivalente:

$$\max_{\mathbf{C_W}} \left| \mathbf{I}_n + \frac{1}{\sigma_z^2}\mathbf{H}\mathbf{C_W}\mathbf{H}^H \right| = \max_{\mathbf{D}} \prod_{i=1}^{\min(m,n)} \left(1 + \frac{1}{\sigma_z^2}\lambda_i d_i\right) \tag{4.30}$$

avec la contrainte:

$$\sum_{i=1}^{m} d_i = P \tag{4.31}$$

La capacité est donc donnée par:

$$C = \max_{\mathbf{D}} \left[\sum_{i=1}^{\min(m,n)} E_{\mathbf{H}} \left\{ \log_2 \left(1 + \frac{1}{\sigma_z^2}\lambda_i d_i\right) \right\} \right] \tag{4.32}$$

La matrice de covariance $\mathbf{C_W}$ permettant d'atteindre cette capacité est donnée par:

$$\mathbf{C_W} = \mathbf{V}^{-1}\mathbf{D}\mathbf{V} \tag{4.33}$$

où $\mathbf{V}^{-1}$ est la matrice des vecteurs propres de $\mathbf{H}^H\mathbf{H}$ et $\mathbf{D}$ est la matrice des d_i obtenue après maximisation de (4.32).

4.3.2 Maximisation sur la matrice des valeurs propres D

L'équation (4.32) montre que le système est équivalent à $\min(m, n)$ canaux parallèles avec les puissances $P_i = d_i$, $i = 1, \ldots, \min(m, n)$, et un bruit équivalent $N_i = \sigma_z^2/\lambda_i$. Selon Cover [CT91], la solution de l'équation est obtenue à l'aide de l'algorithme de remplissage («waterfilling»). Comme illustré à la Figure 4.2, le remplissage consiste à augmenter la puissance émise à partir de zéro en l'allouant au canal ayant le moins de bruit (N_1). La

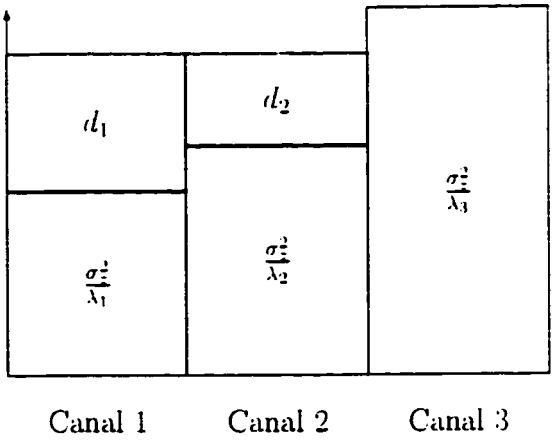


Figure 4.2: Principe du remplissage ("waterfilling").

puissance est ensuite distribuée aux canaux plus bruyants jusqu'à ce que la puissance totale soit atteinte ($d_1 + d_2 = P$). On remarque qu'il est possible que l'on alloue aucune puissance à un canal ($d_3 = 0$). Ce phénomène se produit lorsque la puissance totale transmise n'est pas suffisante pour «remplir» tous les canaux, situation causée par un trop grand écart entre les niveaux de bruit (N_i) des différents canaux. L'écart de bruit entre deux canaux est donné par:

$$\Delta N = \sigma_z^2 \left| \frac{1}{\lambda_i} - \frac{1}{\lambda_j} \right| \tag{4.34}$$

Deux situations peuvent donc causer une grande différence. Premièrement, un écart entre les valeurs propres λ_i et λ_j. Les plus grands écarts seront causés par une des valeurs propres étant très petite due à une corrélation entre les antennes ou à un canal avec une forte atténuation. Deuxièmement, un niveau de bruit σ_z^2 élevé va amplifier les différences entre les λ_i. Par conséquent, on peut s'attendre à ce que le phénomène soit plus fréquent pour un rapport signal-sur-bruit peu élevé.

La solution générale du calcul de capacité s'obtient en utilisant le multiplicateur de Lagrange et en solutionnant le système linéaire. La solution a été obtenue pour quelques valeurs de $\min(m, n)$, ce qui a permis par la suite d'obtenir par déduction une équation

générale. Dans le cas où $\min(m,n) = 2$, il est facile de trouver la solution de (4.32) en utilisant le multiplicateur de Lagrange. On trouve alors que les valeurs propres de $\mathbf{C_W}$ sont:

$$d_1 = \frac{P(\lambda_1\lambda_2 + \lambda_1/\rho - \lambda_2/\rho)}{2\lambda_1\lambda_2} \tag{4.35}$$

$$d_2 = \frac{P(\lambda_1\lambda_2 + \lambda_2/\rho - \lambda_1/\rho)}{2\lambda_1\lambda_2} \tag{4.36}$$

Cette solution peut donner une valeur négative, ce qui en pratique n'est pas possible puisque cela signifierait que l'on doive émettre une puissance négative. Comme les valeurs propres de $\mathbf{C_W}$ ne peuvent être négatives, il faut ajouter la contrainte que $d_1, d_2 \geq 0$, ou de façon équivalente:

$$\rho > \frac{|\lambda_1 - \lambda_2|}{\lambda_1\lambda_2} \tag{4.37}$$

Ou encore, pour faire référence au principe du remplissage(voir (4.34)), on peut exprimer la contrainte de puissance positive par:

$$P > \sigma_z^2 \left| \frac{1}{\lambda_1} - \frac{1}{\lambda_2} \right| \tag{4.38}$$

i.e., la puissance doit être plus grande que l'écart de bruit équivalent entre les deux canaux afin de les «remplir» tous les deux. En supposant que cette contrainte est respectée, le maximum de (4.32) pour le cas où $\min(m,n) = 2$ est donc:

$$\max_{\mathbf{D}} \prod_{i=1}^{2} \left(1 + \frac{1}{\sigma_z^2}\lambda_i d_i\right) = \frac{(\rho\lambda_1\lambda_2 + \lambda_1 + \lambda_2)^2}{4\lambda_1\lambda_2} \tag{4.39}$$

Dans le cas où la valeur de λ_2 est négative (en posant arbitrairement $\lambda_2 \leq \lambda_1$, s'il y a une valeur propre négative, ce sera λ_2), celle-ci est fixée à zéro et λ_1 est donc égale à P. Le maximum est alors simplement $1 + \rho\lambda_1$.

En utilisant la même méthode pour $\min(m,n) = 3, 4, \ldots$, on trouve la solution suivante:

$$d_i = \frac{P}{\rho \min(m,n)} \left[\rho + \sum_{j=1}^{\min(m,n)} \frac{1}{\lambda_j} - \frac{\min(m,n)}{\lambda_i}\right] \tag{4.40}$$

Aussi, le maximum est donné par la formule générale:

$$\max_{\mathbf{D}} \prod_{i=1}^{\min(m,n)} \left(1 + \frac{1}{\sigma_z^2}\lambda_i d_i\right) = \prod_{i=1}^{\min(m,n)} \lambda_i \left[\frac{\rho + \sum_{i=1}^{\min(m,n)} \frac{1}{\lambda_i}}{\min(m,n)}\right]^{\min(m,n)} \tag{4.41}$$

Cependant, tel qu'indiqué auparavant, il est possible qu'une valeur propre soit négative et que la solution ne soit pas valide. Cette valeur propre est alors fixée à zéro et il faut alors vérifier si la solution d'un ordre moins élevé (c'est-à-dire $\min(m,n)-1$) est valide. Afin de vérifier si la solution est valide, on vérifie si la plus petite valeur propre est positive. Ceci peut se faire en utilisant l'équation (4.40) directement ou en vérifiant si la puissance est suffisante pour «remplir» tous les canaux: les deux tests sont équivalents. En effet, l'équation à tester à l'étape 3 de l'algorithme suivant peut être obtenue en utilisant l'équation (4.40) directement:

$$\begin{aligned} d_k &= \frac{P}{\rho k}\left[\rho - \sum_{j=1}^{k}\frac{1}{\lambda_j} - \frac{k}{\lambda_i}\right] \gtrless 0 \\ \rho &\gtrless \frac{k-1}{\lambda_k} - \sum_{i=1}^{k-1}\frac{1}{\lambda_i} \end{aligned}$$

La même équation est aussi obtenue en vérifiant si la puissance est suffisante pour «remplir» les canaux selon le principe du remplissage (voir Figure 4.2), avec $\lambda_1 > \lambda_2 > \ldots > \lambda_k$:

$$\begin{aligned} P &\gtrless \left(\frac{\sigma_z^2}{\lambda_k} - \frac{\sigma_z^2}{\lambda_1}\right) - \left(\frac{\sigma_z^2}{\lambda_k} - \frac{\sigma_z^2}{\lambda_2}\right) - \ldots - \left(\frac{\sigma_z^2}{\lambda_k} - \frac{\sigma_z^2}{\lambda_{k-1}}\right) \\ P &\gtrless \sum_{i=1}^{k-1}\left(\frac{\sigma_z^2}{\lambda_k} - \frac{\sigma_z^2}{\lambda_i}\right) \\ \rho &\gtrless \frac{k-1}{\lambda_k} - \sum_{i=1}^{k-1}\frac{1}{\lambda_i} \end{aligned}$$

Ce processus est répété jusqu'à ce que la solution soit valide, c'est-à-dire jusqu'à ce que toutes les valeurs propres de la solution soient positives. L'algorithme suivant permet de trouver le maximum pour arriver à déterminer la capacité du canal.

Algorithme de calcul de la capacité lorsque l'émetteur connaît l'état du canal

1. $k \leftarrow \min(m, n)$

2. Calculer les valeurs propres non nulles de $\mathbf{H}\mathbf{H}^H$. $\lambda_1 > \lambda_2 > \ldots > \lambda_k$

3. Tester:

$$\rho \underset{d_k<0}{\overset{d_k>0}{\gtrless}} \frac{k-1}{\lambda_k} - \sum_{i=1}^{k-1} \frac{1}{\lambda_i}$$

si $d_k < 0$: $d_k = 0$, $k \leftarrow k-1$, répéter l'étape 3.

si $d_k \geq 0$: la solution est valide:

$$d_i = \begin{cases} \frac{P}{\rho k}\left[\rho + \sum_{j=1}^{k} \frac{1}{\lambda_j} - \frac{k}{\lambda_i}\right], & i \leq k \\ 0. & i > k \end{cases}$$

et le maximum est donné par:

$$\max_{\mathbf{D}} \prod_{i=1}^{\min(m,n)} \left(1 + \frac{1}{\sigma_z^2}\lambda_i d_i\right) = \prod_{i=1}^{k} \lambda_i \left[\frac{\rho + \sum_{i=1}^{k} \frac{1}{\lambda_i}}{k}\right]^k$$

On se souvient que lorsqu'aucun traitement n'est effectué, le système est équivalent à $\min(m, n)$ canaux parallèles. Dans le cas où le traitement est optimal, on remarque que le système peut être réduit à $k \leq \min(m, n)$ canaux indépendants. La solution optimale peut donc consister à réduire le nombre de canaux équivalents en délaissant les pires canaux et en concentrant la puissance vers les meilleurs canaux. L'algorithme a été utilisé pour tracer les graphiques de capacité de la Figure 4.3. Chaque point représente la moyenne de 10000 valeurs de capacité prises comme variables aléatoires, calculées à l'aide de cet algorithme.

4.3.3 Transformation linéaire G

La solution des d_i permet de calculer la matrice de covariance qui atteint la capacité à l'aide de l'équation (4.33). Par la suite, il s'agit de déterminer la transformation linéaire

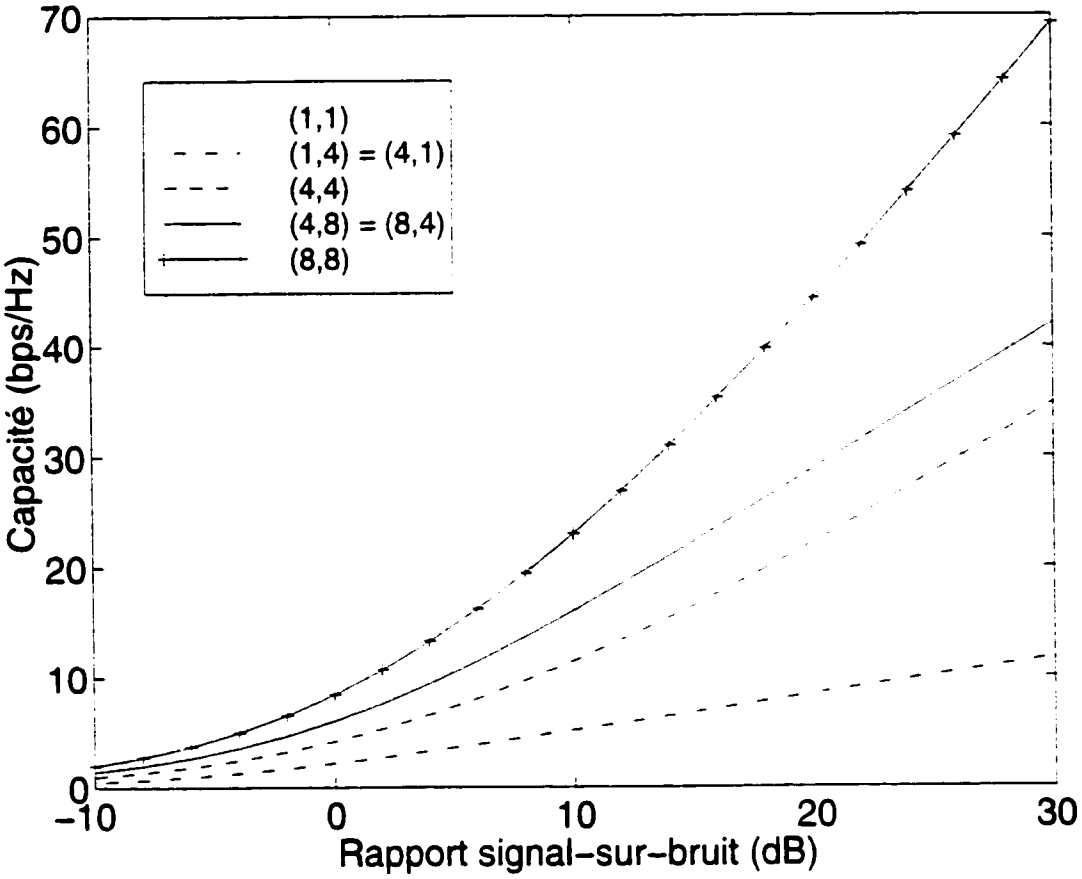


Figure 4.3: Capacité avec prétraitement optimal.

$\mathbf{G}$ correspondante. En combinant les dérivations de la section 4.2 sur la relation entre la matrice de covariance $\mathbf{C_W}$ et la transformation linéaire $\mathbf{G}$, et la solution de maximisation, on obtient:

$$\mathbf{C_W} = \frac{P}{m}\mathbf{G}\mathbf{G}^H = \mathbf{V}^H\mathbf{D}\mathbf{V} \tag{4.42}$$

La matrice $\mathbf{G}$ n'est pas unique mais un choix judicieux est:

$$\mathbf{G} = \sqrt{\frac{m}{P}}\mathbf{V}^H\mathbf{D}^{1/2} \tag{4.43}$$

En fait, la matrice $\mathbf{G}$ doit être de la forme:

$$\mathbf{G} = \sqrt{\frac{m}{P}}\mathbf{V}^H\mathbf{D}^{1/2}\Gamma \tag{4.44}$$

où $\Gamma^{m\times m}$ est une matrice unitaire. Afin de simplifier le traitement, on choisit $\Gamma = \mathbf{I}_m$.

4.4 Capacité asymptotique

La capacité asymptotique avec prétraitement est dérivée en posant $\rho \gg 1$ dans l'équation de maximisation (4.32):

$$\max_{\mathbf{D}} \prod_{i=1}^{\min(m,n)} \frac{1}{\sigma_z^2}\lambda_i d_i = \prod_{i=1}^{\min(m,n)} \left(\frac{1}{\sigma_z^2}\lambda_i\right) \max_{\mathbf{D}} \prod_{i=1}^{\min(m,n)} d_i \tag{4.45}$$

En utilisant le multiplicateur de Lagrange, la solution devient simplement:

$$d_i = \begin{cases} \frac{P}{\min(m,n)} & i \leq \min(m,n) \\ 0 & i > \min(m,n) \end{cases} \tag{4.46}$$

et la capacité asymptotique s'exprime par:

$$C_{asympt} = \min(m,n)\log_2\left(\frac{\rho}{\min(m,n)}\right) + \sum_{i=1}^{\min(m,n)} E_{\mathbf{H}}\left[\log_2 \lambda_i\right] \tag{4.47}$$

L'espérance mathématique a été calculée au chapitre précédent et la capacité asymptotique devient ((3.21) et (3.22)):

$$C_{asympt} = \min(m,n)\log_2\left(\frac{\rho}{\min(m,n)}\right) + \frac{1}{\ln(2)}(l - \min(m,n)\gamma) \tag{4.48}$$

où l pour $\min(m,n) > 1$ est donné par:

$$l = \min(m,n) \sum_{i=1}^{|m-n|} \frac{1}{i} + \sum_{i=1}^{\min(m,n)-1} \frac{i}{\max(m,n)-i} \tag{4.49}$$

4.4.1 Asymptote pour $n \geq m$

On remarque que pour $n \geq m$, $\mathbf{C_W}$ est alors:

$$\mathbf{C_W} = \frac{P}{m}\mathbf{I}_m = \mathbf{C_X} \tag{4.50}$$

Ceci signifie qu'aucun traitement n'est nécessaire à l'émetteur afin d'atteindre la capacité lorsque le nombre d'antennes réceptrices est égal ou supérieur au nombre d'antennes émettrices pour $\rho \gg 1$, c'est-à-dire pour un rapport signal-sur-bruit très grand. En effet, la capacité est alors donnée par:

$$C = E_{\mathbf{H}}\left[\log_2\left|\mathbf{I}_n + \frac{\rho}{m}\mathbf{H}\mathbf{H}^H\right|\right] \tag{4.51}$$

Ceci est le même résultat obtenu que dans le cas où le canal n'est pas connu à l'émetteur. Les Figures 4.4 et 4.5 comparent la capacité avec et sans traitement. On remarque que le gain obtenu du prétraitement lorsque $n \geq m$ est très limité, particulièrement pour un rapport signal-sur-bruit élevé. Cette limitation du prétraitement devra être considérée lors de la conception d'un tel système, et le gain de capacité potentiel devra être comparé à l'augmentation de complexité qu'entraine l'utilisation d'un tel prétraitement à l'émetteur.

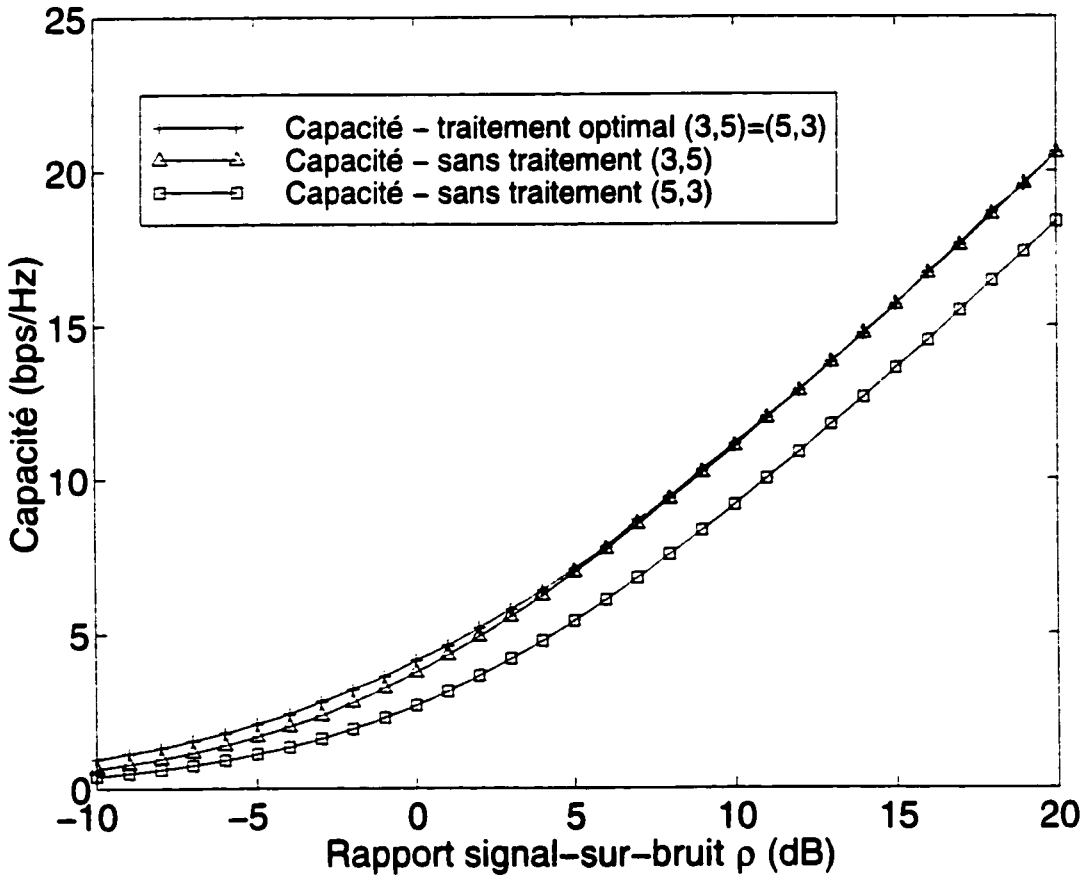


Figure 4.4: Comparaison de la capacité: prétraitement optimal et aucun prétraitement.

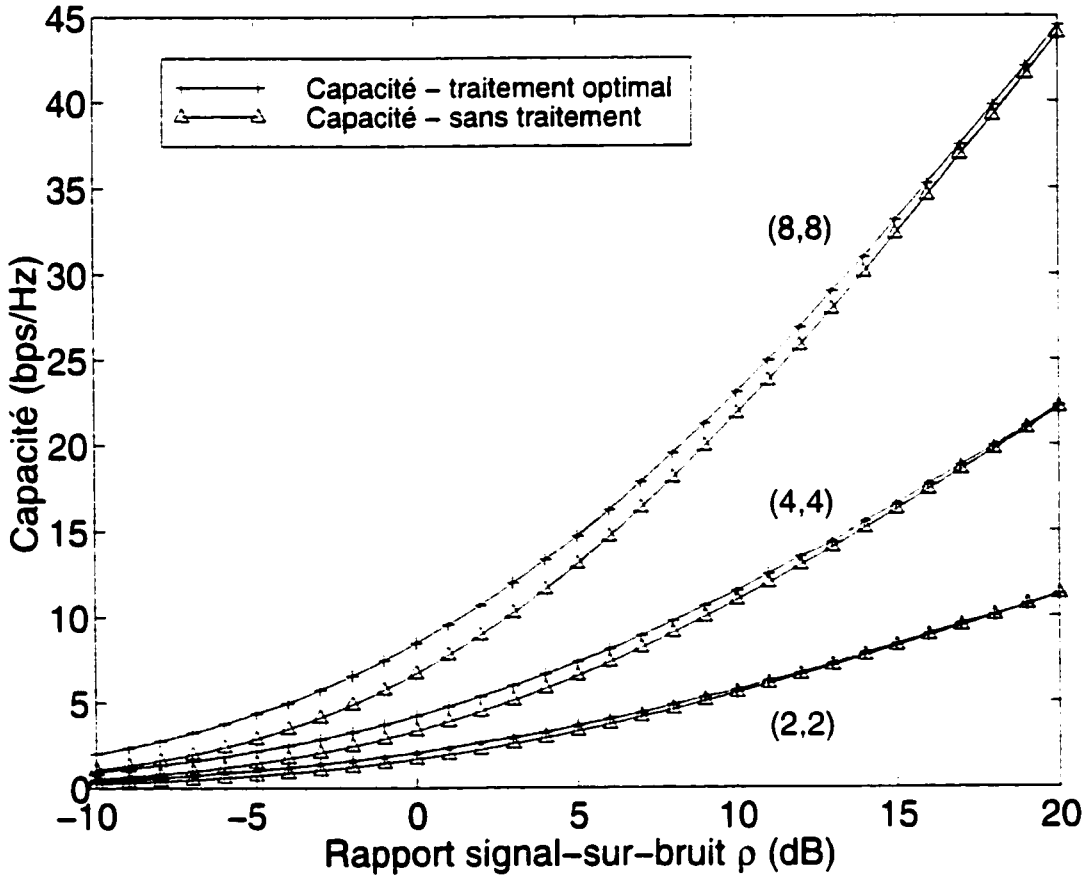


Figure 4.5: Comparaison de la capacité: prétraitement optimal et aucun prétraitement.

4.5 Gain obtenu suite à la transformation linéaire G pour $\rho \gg 1$

Nous comparons maintenant les rapport-sur-bruit nécessaires, avec et sans prétraitement, afin d'obtenir la même valeur de capacité, ce qui nous donne le gain obtenu de la transformation linéaire **G**. Le signal reçu dans le cas avec prétraitement est donné par:

$$\begin{aligned} \mathbf{Y} &= \mathbf{HW} + \mathbf{Z} = \mathbf{HGX} + \mathbf{Z} \\ &= \sqrt{\frac{m}{P}}\mathbf{USD}^{1/2}\mathbf{X} + \mathbf{Z} \end{aligned} \tag{4.52}$$

Il est possible de calculer le gain sur le rapport signal-sur-bruit dû à cette transformation. En effet, le rapport signal-sur-bruit avec prétraitement est donné par:

$$\rho_{opt} = \frac{E\left[(\mathbf{HW})^H \mathbf{HW}\right]}{E\left[\mathbf{Z}^H\mathbf{Z}\right]} \tag{4.53}$$

$$\begin{aligned} &= \frac{E\left[\left(\sqrt{\frac{m}{P}}\mathbf{USD}^{1/2}\mathbf{X}\right)^H \sqrt{\frac{m}{P}}\mathbf{USD}^{1/2}\mathbf{X}\right]}{E\left[\mathbf{Z}^H\mathbf{Z}\right]} \\ &= \frac{E\left[\|\mathbf{SD}^{1/2}\|_f^2\right]}{n\sigma_z^2} \\ &= \frac{\sum_{i=1}^{\min(m,n)} E\left[|s_i|^2 d_i\right]}{n\sigma_z^2} \\ &= \frac{\sum_{i=1}^{\min(m,n)} E\left[\lambda_i d_i\right]}{n\sigma_z^2} \end{aligned} \tag{4.54}$$

Il est difficile de trouver l'espérance pour la solution générale des d_i. Par contre, pour $\rho \gg 1$, la simplicité de la solution permet d'obtenir:

$$\begin{aligned} \rho_{opt} &= \frac{P}{\min(m,n)n\sigma_z^2} \sum_{i=1}^{\min(m,n)} E\left[\lambda_i\right] \\ &= \frac{P}{\sigma_z^2}\frac{\max(m,n)}{n} \\ &= \frac{\max(m,n)}{n}\rho \end{aligned} \tag{4.55}$$

où $\rho = \frac{P}{\sigma_z^2}$ est le rapport signal-sur-bruit sans prétraitement. On remarque que pour $n \geq m$, il n'y a aucun gain sur le rapport signal-sur-bruit, ce qui confirme les résultats précédents. Il est possible d'arriver au même résultat en utilisant les équations de capacité asymptotique avec et sans traitement optimal, C_{asympt}^{opt} et C_{asympt}^{nopt}:

$$\begin{aligned} C_{asympt}^{opt}(\rho_{opt}) - C_{asympt}^{nopt}(\rho_{nopt}) &= \min(m,n)\left[\log_2\left(\frac{\rho_{opt}}{\min(m,n)}\right) - \log_2\left(\frac{\rho_{nopt}}{m}\right)\right] = 0 \\ \Rightarrow \log_2\left(\frac{\rho_{opt}}{\min(m,n)}\right) &= \log_2\left(\frac{\rho_{nopt}}{m}\right) \\ \frac{\rho_{nopt}}{\rho_{opt}} &= \frac{m}{\min(m,n)} = \frac{\max(m,n)}{n} \end{aligned} \tag{4.56}$$

Donc afin d'obtenir la même capacité avec et sans prétraitement, le rapport signal-sur-bruit sans prétraitement doit être $\frac{\max(m,n)}{n}$ supérieur à celui avec prétraitement. Le gain de capacité obtenu du prétraitement à l'émetteur est donné par:

$$\begin{aligned} \Delta C_{asympt} &= C_{opt} - C_{nopt} \\ &= \min(m,n)\left[\log_2\left(\frac{\rho}{\min(m,n)}\right) - \log_2\left(\frac{\rho}{m}\right)\right] \\ &= \min(m,n)\log_2\left(\frac{m}{\min(m,n)}\right) = \min(m,n)\log_2\left(\frac{\max(m,n)}{n}\right) \end{aligned} \tag{4.57}$$

4.6 Symmétrie de la capacité

Tel que mentionné précédemment, les valeurs propres non nulles λ_i de $\mathbf{H}\mathbf{H}^H$ et $\mathbf{H}^T\mathbf{H}^{T^H}$ sont identiques. On remarque donc que le problème de maximisation est identique pour les cas (m,n) et (n,m):

$$C = \max_{\mathbf{D}} \sum_{i=1}^{\min(m,n)} E_{\mathbf{H}}\left[\log_2(1 + \frac{1}{\sigma_z^2}\lambda_i d_i)\right] \tag{4.58}$$

Par conséquent, la capacité dans les deux cas est la même, ce qui nous permet d'affirmer ce qui suit:

Déclaration 1 *La capacité d'un système (m,n) est identique à celle d'un système (n,m) lorsqu'un traitement optimal est appliqué à l'émetteur.*

Cela peut sembler intuitivement incorrect puisque comme discuté auparavant, le cas $(1, n)$ aura l'avantage par rapport au cas $(m, 1)$ de multiplier la puissance par n. Par contre, lorsque le canal est connu à l'émetteur, le prétraitement effectué correspond à une forme de formation de faisceau ("beamforming"), ce qui en fait multiplie la puissance transmise dans la direction des antennes réceptrices, et par conséquent les capacités dans les deux directions deviennent identiques.

Donc, théoriquement, il n'y a aucun avantage à utiliser un système plutôt que l'autre. Par contre, si le rapport signal-sur-bruit est élevé, il a été démontré précédemment qu'aucun pré-traitement n'est nécessaire pour un système (m, n) avec $n \geq m$, ce qui en pratique constitue un avantage important.

4.7 Conclusions

Dans ce chapitre, nous avons d'abord dérivé l'expression de la capacité lorsque l'état du canal est connu de l'émetteur et du récepteur. Dans ce cas, l'émetteur adapte les symboles transmis au canal en appliquant une transformation linéaire $\mathbf{G}$ au vecteur des symboles transmis. Cette expression prends la forme d'une maximisation similaire à l'expression de capacité (2.19) présentée au chapitre 2, avec la différence que la capacité est maintenant aussi fonction de $\mathbf{G}$. La solution de cette maximisation a ensuite été dérivée afin d'obtenir le prétraitement linéaire $\mathbf{G}$ requis à l'émetteur. Aussi, un algorithme permettant de trouver la solution a été présenté. Ensuite, on a remarqué que les capacités asymptotiques avec et sans traitement sont identiques lorsque le nombre d'antennes réceptrices est égal ou supérieur au nombre d'antennes émettrices $(n \geq m)$. Ce phénomène a une conséquence pratique intéressante. En effet, si $n \geq m$, il n'est donc pas très avantageux de concevoir un émetteur complexe avec prétraitement. Par ailleurs, le gain de capacité obtenu du prétraitement devient de plus en plus important à mesure que le rapport m/n augmente et que le rapport signal-sur-bruit diminue. Finalement, il a été démontré que la capacité est symmétrique, c'est-à-dire que la capacité est la même dans les deux directions lorsque l'émetteur connait l'état du canal. En

d'autres mots, la capacité est identique pour les cas $(m, n) = (a, b)$ et $(m, n) = (b, a)$.

Chapitre 5

Erreur d'estimation du canal à l'émetteur

5.1 Introduction

Au chapire précédent. le traitement optimal à l'émetteur a été analysé en supposant que le canal était connu parfaitement à l'émetteur. En pratique. les paramètres du canal doivent être estimés et on peut s'attendre à une réduction de la capacité suite à une estimation imparfaite.

On étudie donc dans ce chapitre l'effet sur la capacité d'une erreur d'estimation à l'émetteur. À cette fin. on modélise à la section 5.2 cette erreur d'estimation par un bruit gaussien. Ce modèle d'erreur nous permet ensuite de concevoir à la section 5.3 un algorithme de calcul de la capacité en fonction du niveau d'erreur d'estimation. et les résultats de simulation utilisant cet algorithme sont présentés en fonction de l'importance de l'erreur d'estimation et du rapport signal-sur-bruit.

5.2 Modélisation de l'erreur d'estimation

La méthode la plus simple pour estimer le canal consiste à émettre une matrice de symboles connus $\mathbf{X}^{m\times m}$, c'est-à-dire un symbole pour chacune des m antennes pendant m intervalles de temps. La matrice des symboles reçus est donc:

$$\mathbf{Y} = \mathbf{HX} + \mathbf{Z} \tag{5.1}$$

L'estimation linéaire de l'état du canal consiste simplement à inverser la matrice des symboles transmis:

$$\hat{\mathbf{H}} = \mathbf{YX}^{-1} = \mathbf{H} + \mathbf{ZX}^{-1} \tag{5.2}$$

On remarque que la variance des éléments de $\hat{\mathbf{H}}$ sera en général différente de celle de $\mathbf{H}$, qui est de valeur unitaire dans notre cas. L'estimation finale consiste donc à normaliser $\hat{\mathbf{H}}$ par rapport à la variance:

$$\tilde{\mathbf{H}} = \frac{\hat{\mathbf{H}}}{\sqrt{\text{var}\left(\hat{h}\right)}} \tag{5.3}$$

où $\text{var}(\hat{h})$ est la variance des éléments de la matrice $\hat{\mathbf{H}}$. La matrice estimée du canal $\tilde{\mathbf{H}}$ est donc de la forme:

$$\tilde{\mathbf{H}} = \frac{\mathbf{H} + \mathbf{E}}{\sqrt{\text{var}\,(h + e)}} \tag{5.4}$$

où $\mathbf{E} = \mathbf{ZX}^{-1}$ est la matrice d'erreur dont les éléments sont modélisés par des variables gaussiennes complexes centrées à zéro avec variance σ_e^2. Le dénominateur $\sqrt{\text{var}\,(h + e)}$ impose aux éléments de la matrice estimée $\tilde{\mathbf{H}}$ cette variance unitaire, tout comme $\mathbf{H}$, et ce peu importe la variance de l'erreur. La variance au dénominateur est:

$$\begin{aligned} \text{var}(h+e) &= E[(h+e)(h+e)^*] && (5.5)\\ &= E[hh^*] + E[ee^*] && (5.6)\\ &= 1 + \sigma_e^2 && (5.7) \end{aligned}$$

L'erreur d'estimation n'étant pas corrélée au gain complexe du canal, la matrice d'estimation du canal est donc:

$$\tilde{\mathbf{H}} = \frac{\mathbf{H} + \mathbf{E}}{\sqrt{1 + \sigma_e^2}} \tag{5.8}$$

La variance de l'erreur d'estimation σ_e^2 peut prendre une valeur dans l'intervalle $[0, \infty[$, où une variance nulle correspond à une estimation parfaite du canal et une variance infinie signifie que l'estimation est une variable complètement aléatoire et indépendante du canal. Il est possible de traduire la variance d'erreur en un coefficient de corrélation entre le canal et son estimation, ce qui rend les comparaisons plus intuitives. Le coefficient de corrélation entre les coefficients de la matrice du canal et de son estimation est:

$$\rho_{h,\tilde{h}} = \frac{E[h\tilde{h}^*]}{\sigma_h \sigma_{\tilde{h}}} = \frac{E[h(h+e)^*]}{\sqrt{1+\sigma_e^2}} = \frac{E[hh^*]}{\sqrt{1+\sigma_e^2}} \tag{5.9}$$

$$= \frac{1}{\sqrt{1+\sigma_e^2}} \tag{5.10}$$

La Figure 5.1 montre la corrélation entre l'état réel du canal $\mathbf{H}$ et son estimation $\tilde{\mathbf{H}}$ en fonction de la variance de l'erreur d'estimation. Le modèle d'erreur consiste donc en l'addition d'une matrice au canal dont les éléments sont des variables gaussiennes complexes.

5.3 Résultats de simulation

L'effet de l'erreur d'estimation de la matrice du canal est étudié par simulation. Pour ce faire, nous avons développpé un algorithme utilisant le modèle d'erreur d'estimation dérivé auparavant afin de calculer la capacité pour un grand nombre de réalisations du canal pour ensuite en faire la moyenne.

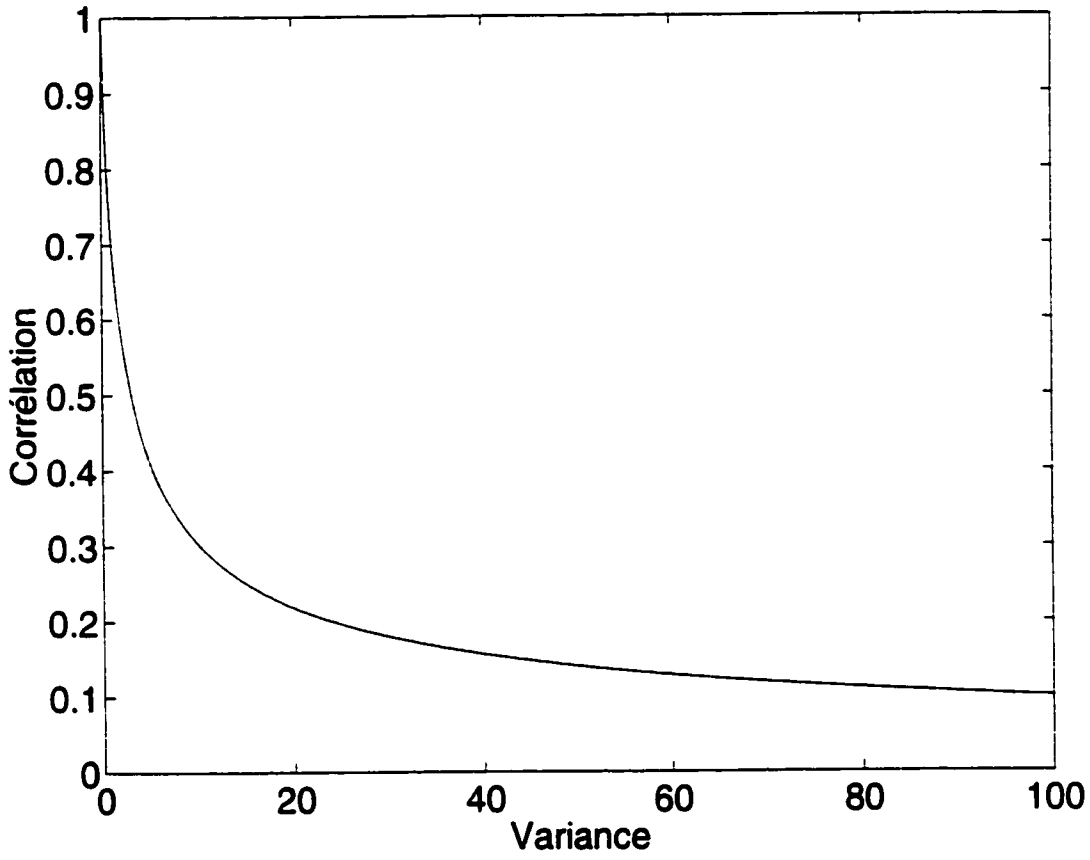


Figure 5.1: Corrélation $\rho_{h\tilde{h}}$ entre $\mathbf{H}$ et $\tilde{\mathbf{H}}$ en fonction de la variance de l'erreur d'estimation σ_e^2.

Algorithme de calcul de la capacité avec erreur d'estimation

- Initialiser le rapport signal-sur-bruit ρ
- Initialiser la variance de l'erreur σ_e^2
- Boucle 1: k réalisations du canal pour faire une moyenne
  - 1.1 Réalisation de $\mathbf{H}$
  - Boucle 2: l matrices d'erreur par réalisation du canal pour faire une moyenne
    - 2.1 Réalisation de la matrice d'erreur $\mathbf{E}$ avec variance σ_e^2
    - 2.2 Simulation de l'estimation du canal:

$$\tilde{\mathbf{H}} = \frac{\mathbf{H} + \mathbf{E}}{\sqrt{\text{var}\,(h + e)}}$$

    - 2.3 Calcul de la matrice de covariance $\tilde{\mathbf{C}}_{\mathbf{w}}$ basé sur l'estimation du canal
    - 2.4 Calcul de la capacité avec erreur d'estimation $\tilde{C}$:

$$\tilde{C} = \log_2 \left| \mathbf{I}_n - \frac{1}{\sigma_z^2} \mathbf{H} \tilde{\mathbf{C}}_{\mathbf{W}} \mathbf{H}^H \right|$$

  - Fin de la boucle 2
  - 1.2 Calcul de la moyenne des l capacités pour une réalisation du canal
- Fin de la boucle 1
- Calcul de la moyenne des k capacités

Cet algorithme est utilisé pour chacune des valeurs de σ_e^2 et de ρ désirées, ce qui permet d'obtenir un graphique de la capacité soit en fonction de la corrélation entre la matrice du canal et son estimation ou soit en fonction du rapport signal-sur-bruit. Dans chacun des cas, la capacité sans traitement à l'émetteur est aussi tracée pour fins de comparaison.

5.3.1 Capacité en fonction de la corrélation

Chacun des graphiques illustrés aux figures 5.2 à 5.10 représente la capacité en fonction de la corrélation entre la matrice du canal $\mathbf{H}$ et son estimation $\tilde{\mathbf{H}}$, pour une valeur du rapport signal-sur-bruit donnée. On remarque d'abord que la capacité est de croissance monotone avec la corrélation. Donc, une meilleure estimation du canal augmente la capacité peu importe la configuration (m, n). Par ailleurs, il est possible, pour de faibles valeurs de corrélation, que la capacité avec prétraitement soit inférieure à la capacité sans prétraitement. Afin de mieux comprendre l'effet de l'estimation du canal sur la capacité avec prétraitement, quelques paramètres sont extraits des graphiques de capacité en fonction de la corrélation.

Pourcentage de réduction de la capacité due à l'estimation du canal

Ce pourcentage, $(C_{max} - C_{min})/C_{max}$, représente la différence de capacité entre les cas où l'estimation du canal est parfaite (corrélation de 1), C_{max}, et où l'état du canal n'est pas connu (corrélation nulle), C_{min}. Le tableau suivant résume les pourcentages de réduction de la capacité tirés des graphiques 5.2 à 5.10.

% de réduction			
(m, n)	-15 dB	3 dB	21 dB
(3,1)	66	49	25
(3,3)	52	23	2
(3,5)	44	11	0.01

Tableau 5.1: Pourcentage de réduction de la capacité avec erreur d'estimation.

On remarque que le pourcentage diminue avec le rapport signal-sur-bruit, c'est-à-dire que plus ρ augmente moins l'effet relatif de l'estimation est important. Aussi, l'effet de l'estimation diminue avec le ratio n/m. Donc, une erreur d'estimation du canal aura un effet plus important lorsque le nombre d'antennes émettrices est supérieur au nombre d'antennes réceptrices et que le rapport signal-sur-bruit est faible.

Corrélation critique

La corrélation $\rho_{h\tilde{h}}$ correspondant à l'intersection de la droite de capacité sans pré-traitement et de la courbe de capacité avec erreur d'estimation est définie comme la corrélation critique $\rho_{critique}$. La corrélation critique permet de comparer différentes configurations (m, n) avec prétraitement au cas sans prétraitement.

	Corrélation critique		
(m, n)	-15 dB	3 dB	21 dB
(3,1)	-	0.3	0.48
(3,3)	0.18	0.67	0.98
(3,5)	0.23	0.78	-

Tableau 5.2: Corrélation critique.

On remarque que la corrélation critique augmente avec le rapport signal-sur-bruit et le rapport n/m. Par conséquent, une estimation de qualité est nécessaire afin d'atteindre une capacité supérieure au cas sans pré-traitement lorsque le nombre d'antennes réceptrices est supérieur au nombre d'antennes émettrices et que le rapport signal-sur-bruit est élevé.

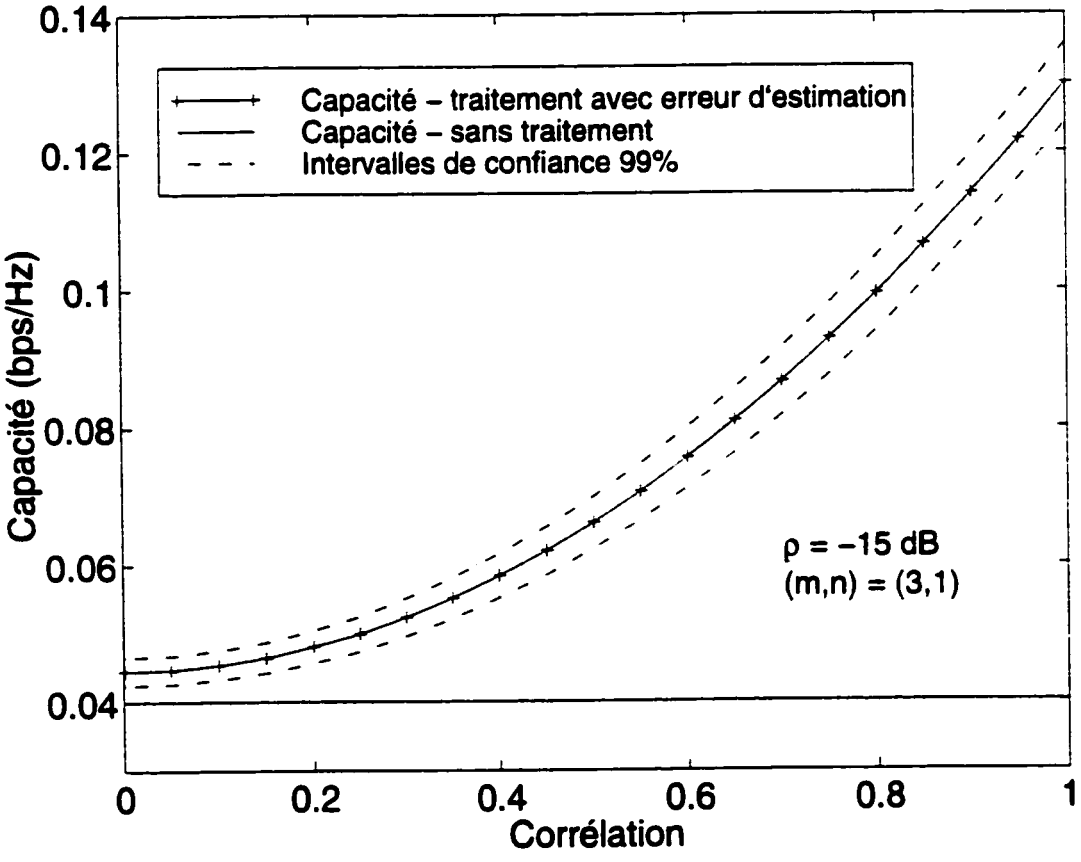


Figure 5.2: Capacité en fonction de la corrélation entre $\mathbf{H}$ et $\tilde{\mathbf{H}}$ pour $\rho = -15$ dB et $(m.n) = (3.1)$.

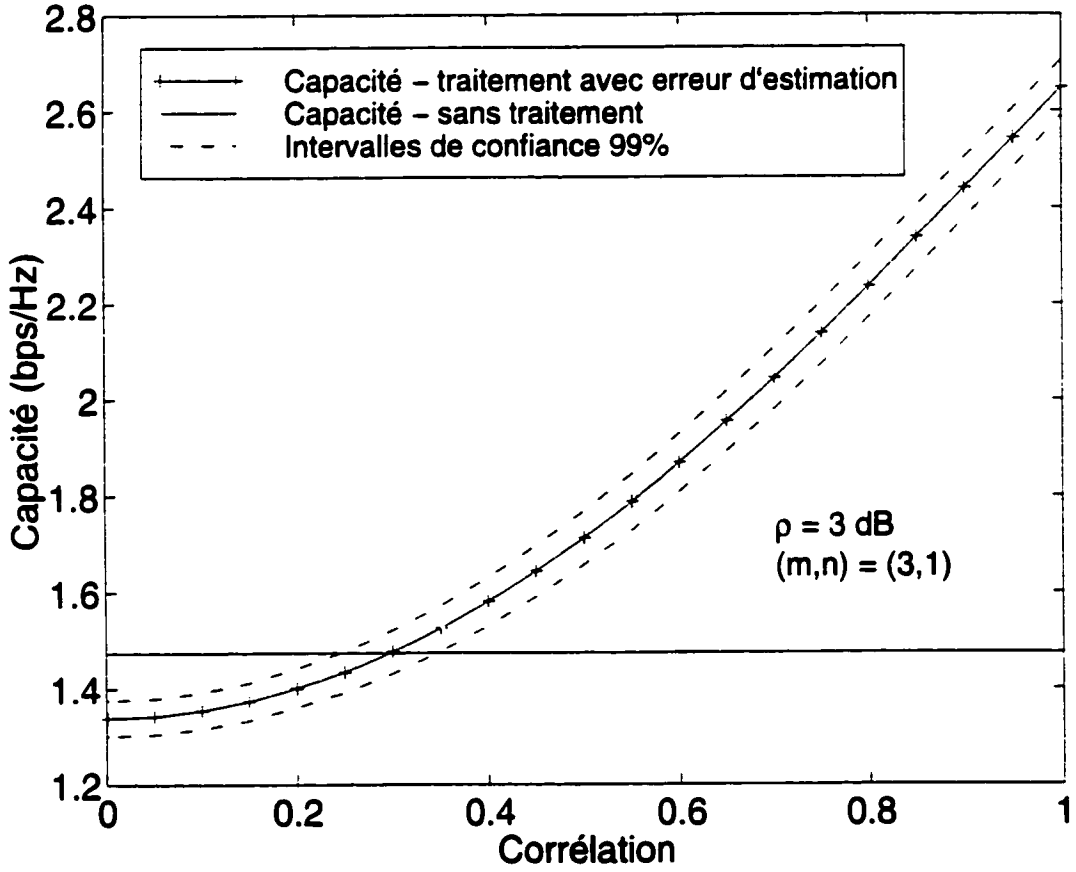


Figure 5.3: Capacité en fonction de la corrélation entre $\mathbf{H}$ et $\tilde{\mathbf{H}}$ pour $\rho = 3$ dB et $(m.n) = (3.1)$.

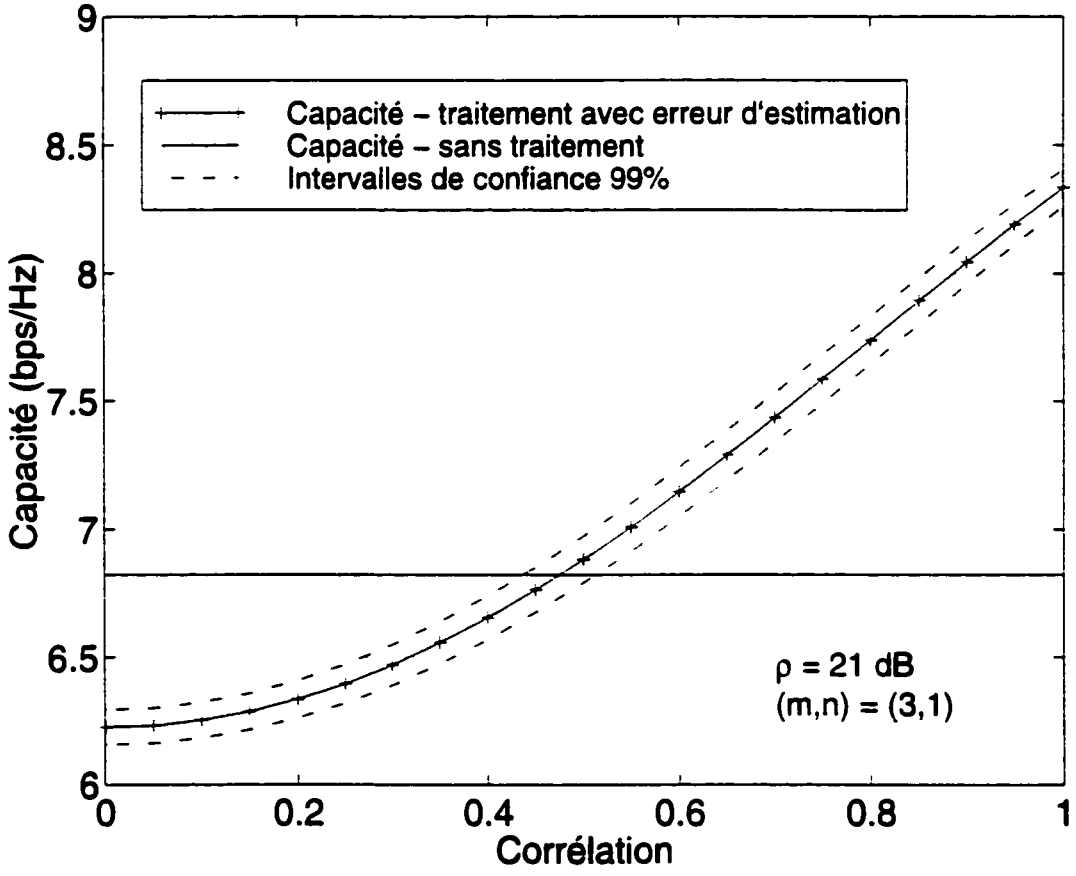


Figure 5.4: Capacité en fonction de la corrélation entre $\mathbf{H}$ et $\tilde{\mathbf{H}}$ pour $\rho = 21$ dB et $(m, n) = (3, 1)$.

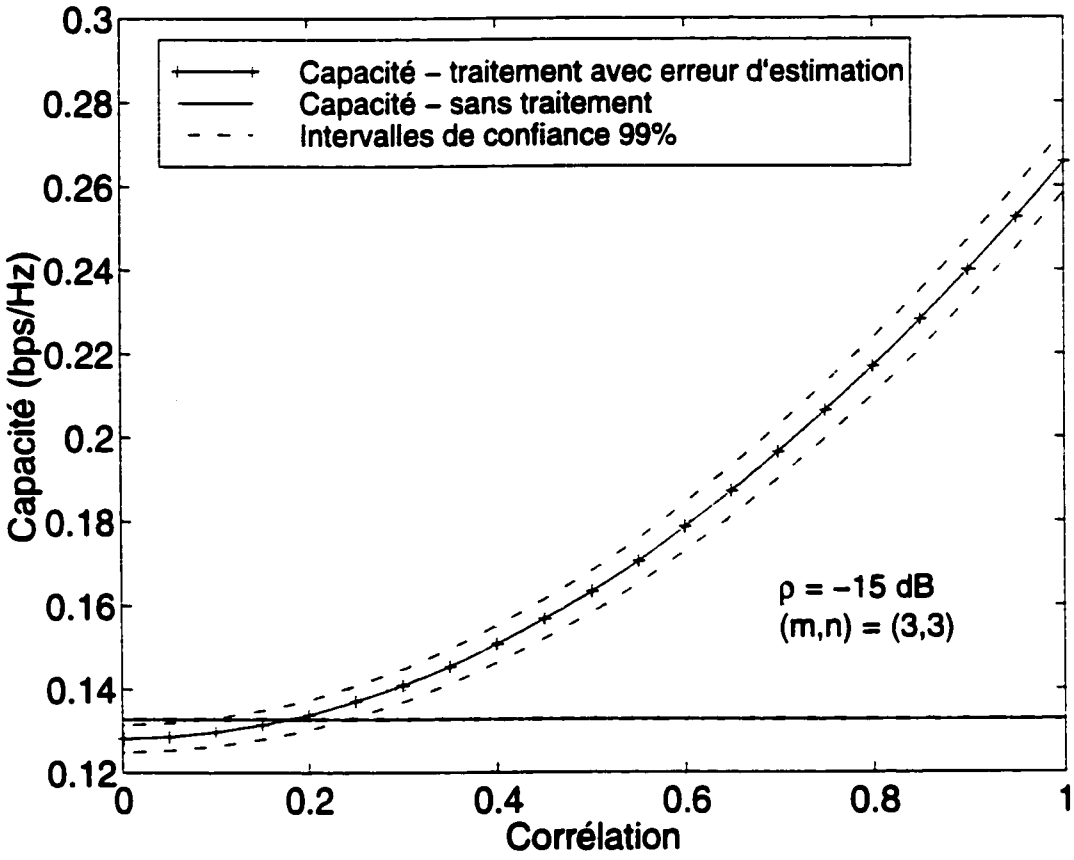


Figure 5.5: Capacité en fonction de la corrélation entre $\mathbf{H}$ et $\tilde{\mathbf{H}}$ pour $\rho = -15$ dB et $(m.n) = (3,3)$.

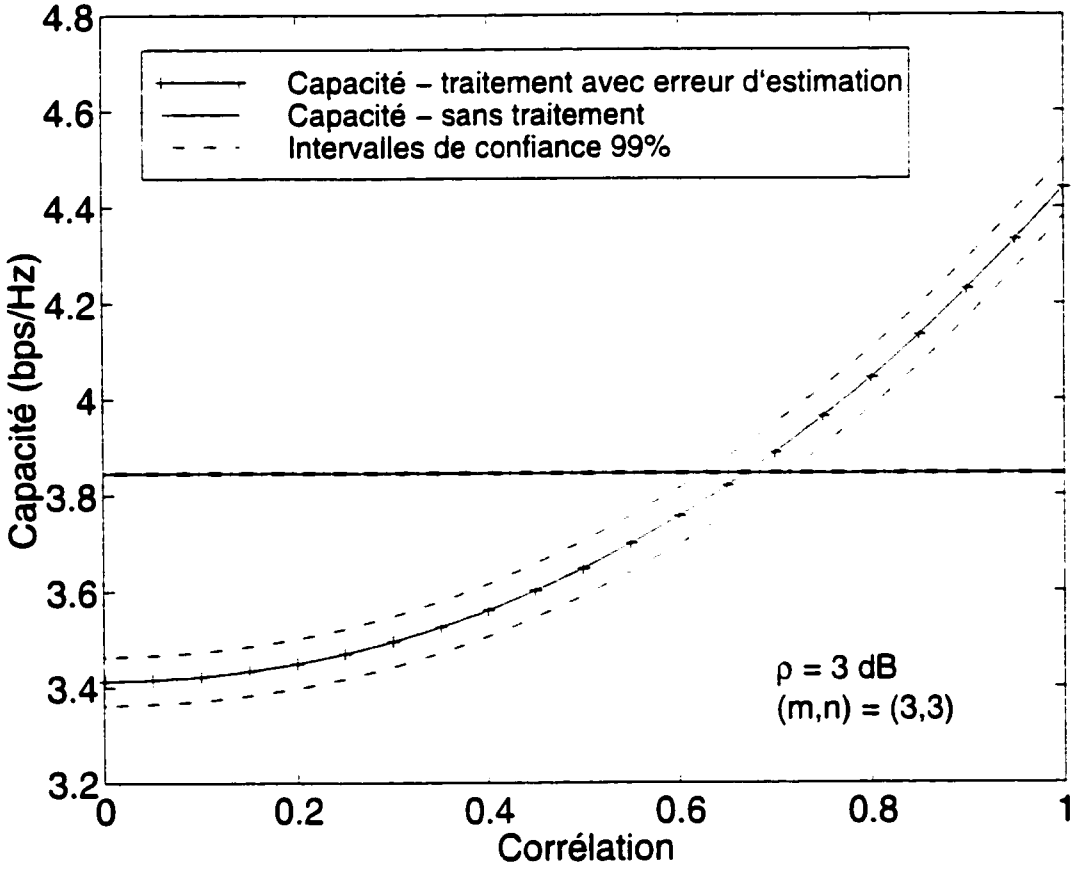


Figure 5.6: Capacité en fonction de la corrélation entre $\mathbf{H}$ et $\hat{\mathbf{H}}$ pour $\rho = 3$ dB et $(m, n) = (3, 3)$.

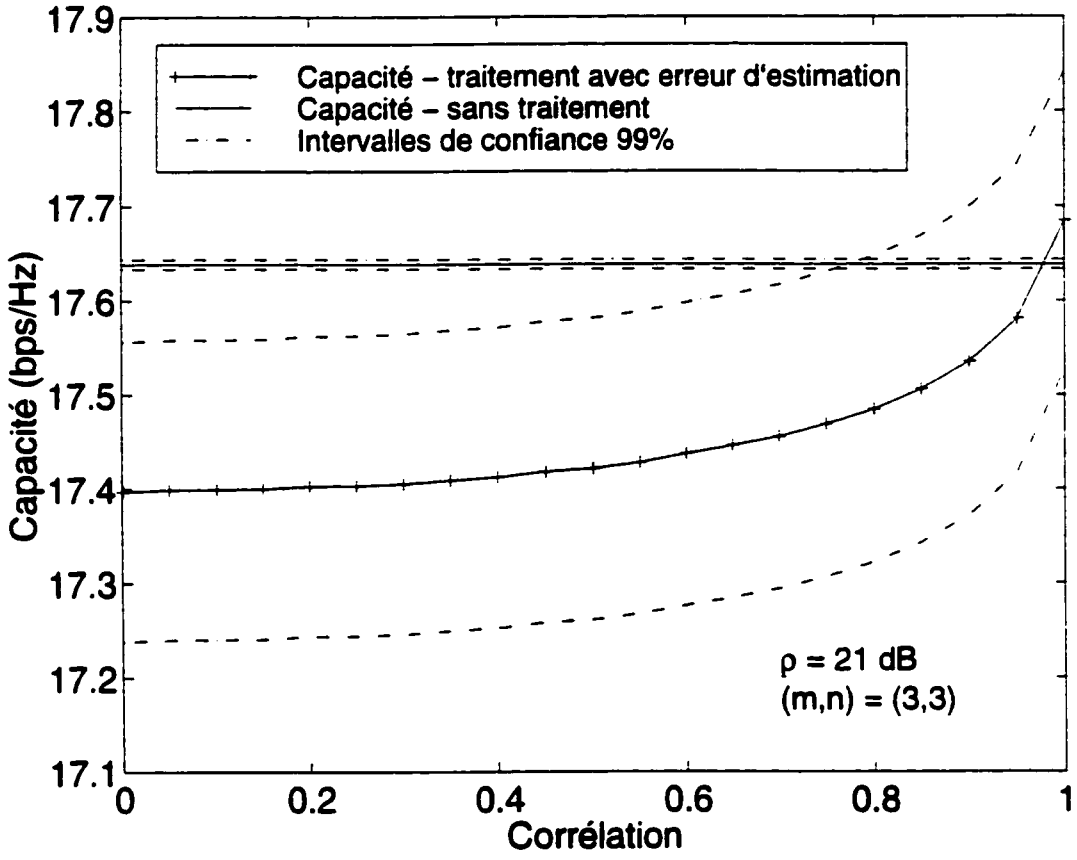


Figure 5.7: Capacité en fonction de la corrélation entre $\mathbf{H}$ et $\tilde{\mathbf{H}}$ pour $\rho = 21$ dB et $(m.n) = (3.3)$.

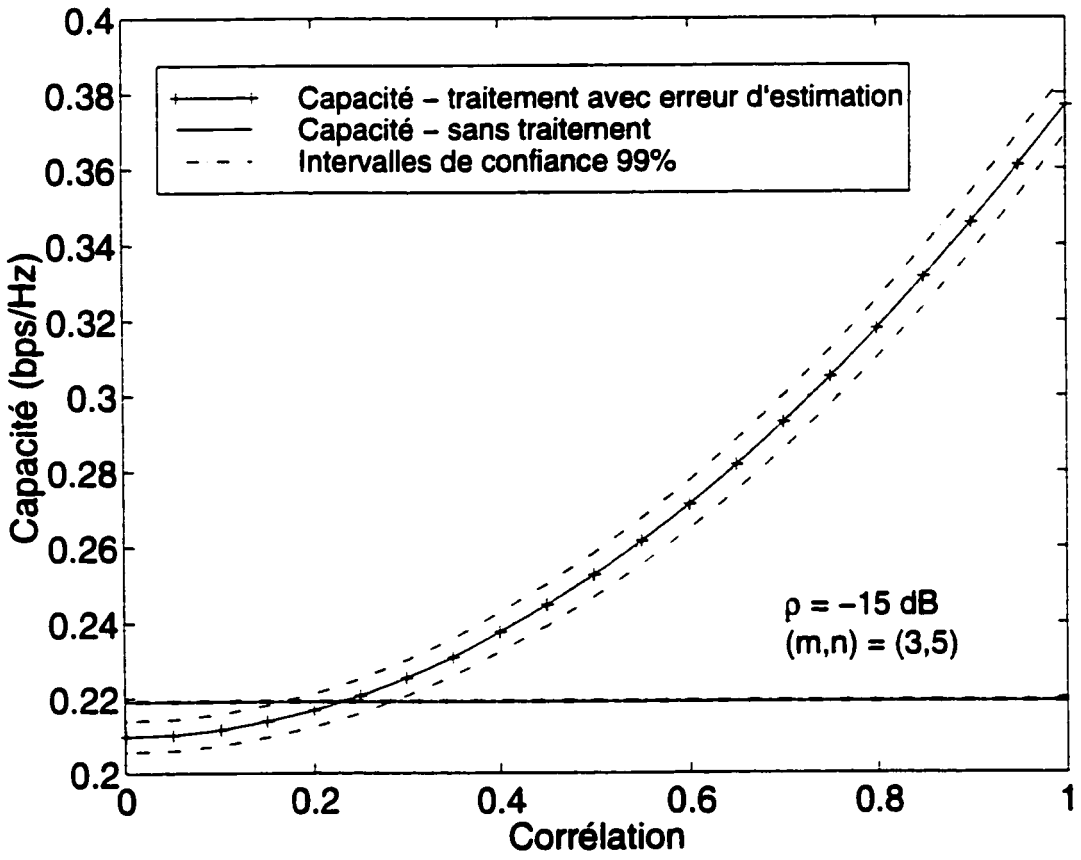


Figure 5.8: Capacité en fonction de la corrélation entre $\mathbf{H}$ et $\tilde{\mathbf{H}}$ pour $\rho = -15$ dB et $(m, n) = (3, 5)$.

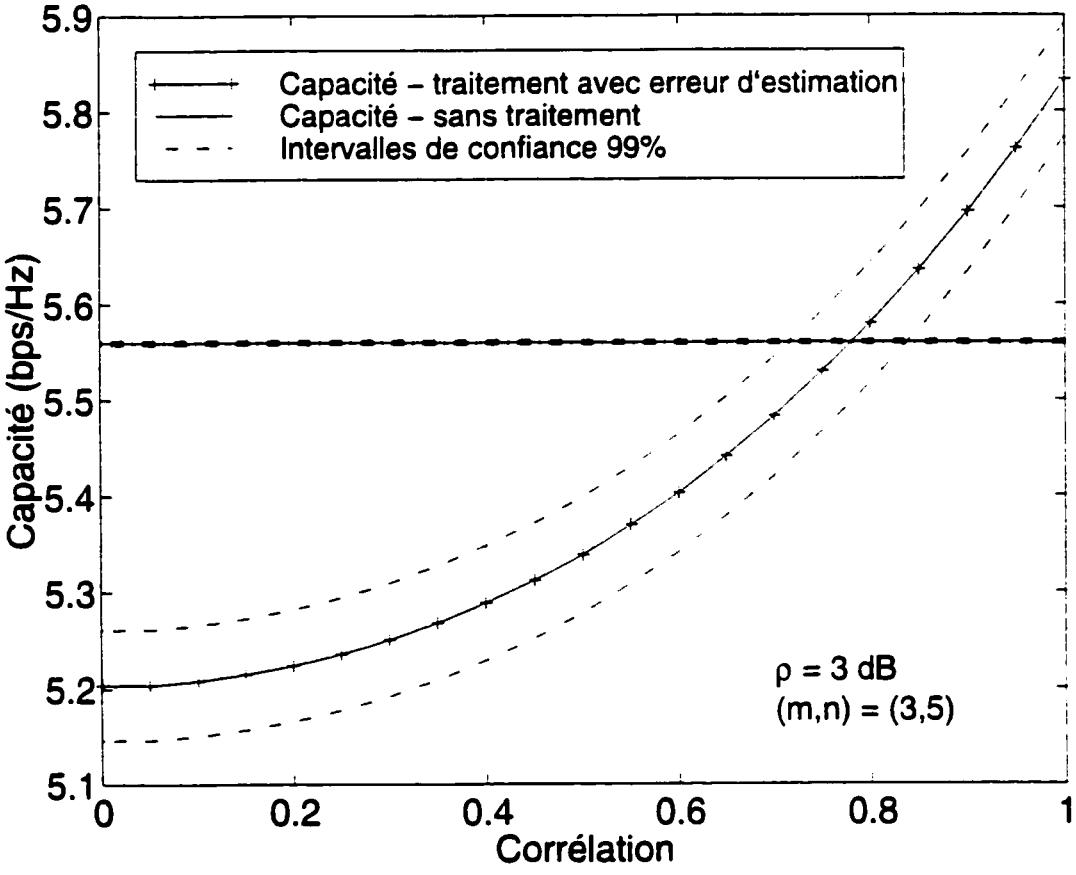


Figure 5.9: Capacité en fonction de la corrélation entre $\mathbf{H}$ et $\tilde{\mathbf{H}}$ pour $\rho = 3$ dB et $(m, n) = (3, 5)$.

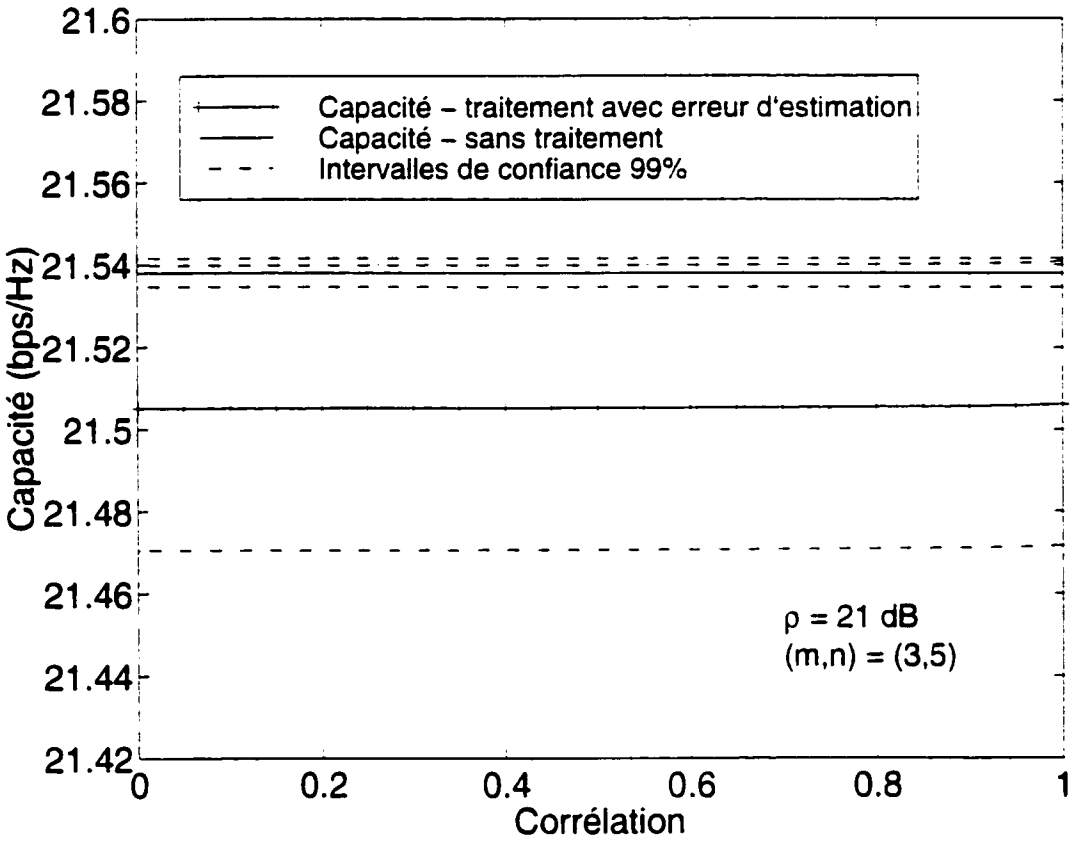


Figure 5.10: Capacité en fonction de la corrélation entre $\mathbf{H}$ et $\tilde{\mathbf{H}}$ pour $\rho = 21$ dB et $(m, n) = (3, 5)$.

5.3.2 Capacité en fonction du rapport signal-sur-bruit

La capacité avec erreur d'estimation est tracée en fonction du rapport signal-sur-bruit pour différentes configurations (m, n) et différentes valeurs de corrélation (0, 0.5, 0.85 et 1) aux figures 5.11 à 5.15.

Quelques remarques sur les graphiques de capacité en fonction du rapport signal-sur-bruit pour différentes valeurs de corrélation suivent. Premièrement, on remarque qu'une mauvaise estimation du canal peut mener à une valeur de capacité inférieure à celle obtenue sans prétraitement. Aussi, comme discuté auparavant, la réduction de capacité due à une estimation imparfaite de l'état du canal se fait sentir surtout lorsque $m > n$. En fait, pour le cas où $n \geq m$ et $\rho \gg 1$, la capacité ne dépend pas de la qualité de l'estimation du canal. Par contre, pour $m > n$, la capacité avec erreur d'estimation est toujours inférieure à la capacité avec estimation parfaite, même pour $\rho \gg 1$.

Ce comportement de la capacité pour une corrélation nulle et $\rho \gg 1$ peut être dérivé. On se base sur l'expression générale de capacité avec prétraitement et erreur d'estimation:

$$C = E_{\mathbf{H}} \left[\log_2 \left| \mathbf{I}_n + \frac{1}{\sigma_{\tilde{z}}^2} \mathbf{H} \tilde{\mathbf{C}}_{\mathbf{W}} \mathbf{H}^H \right| \right] \tag{5.11}$$

où $\tilde{\mathbf{C}}_{\mathbf{W}}$ est la matrice de covariance obtenue après optimisation du prétraitement basée sur la matrice estimée du canal $\tilde{\mathbf{H}}$. La capacité est dérivée pour le cas où la variance de l'erreur tend vers l'infini, c'est-à-dire lorsque la corrélation entre l'estimation du canal et le canal est nulle et le rapport signal-sur-bruit est très grand ($\rho \gg 1$). La matrice de covariance est d'abord décomposée (comme à la section 4.3):

$$\tilde{\mathbf{C}}_{\mathbf{W}} = \tilde{\mathbf{Q}} \tilde{\mathbf{D}} \tilde{\mathbf{Q}}^{-1} \tag{5.12}$$

où $\tilde{\mathbf{D}}$ est une variable déterministe avec $\tilde{d}_{ii} = \frac{P}{\min(m,n)}$ et $\tilde{\mathbf{Q}}$ est une matrice aléatoire indépendante de $\mathbf{H}$ due à une corrélation nulle entre $\mathbf{H}$ et $\tilde{\mathbf{H}}$.

Pour le cas $n \geq m$, en utilisant l'identité $\left|\mathbf{I}_n + \mathbf{A}\mathbf{A}^H\right| = \left|\mathbf{I}_m + \mathbf{A}^H\mathbf{A}\right|$ et $\tilde{\mathbf{D}} =$

$\tilde{\mathbf{D}}^{1/2}\tilde{\mathbf{D}}^{1/2}$, on obtient:

$$C = E_{\mathbf{H},\tilde{\mathbf{Q}}}\left[\log_2\left|\mathbf{I}_n - \frac{1}{\sigma_\xi^2}\mathbf{H}\tilde{\mathbf{Q}}\tilde{\mathbf{D}}\tilde{\mathbf{Q}}^{-1}\mathbf{H}^H\right|\right] \tag{5.13}$$

$$= E_{\mathbf{H},\tilde{\mathbf{Q}}}\left[\log_2\left|\mathbf{I}_m + \frac{1}{\sigma_\xi^2}\tilde{\mathbf{D}}^{1/2}\tilde{\mathbf{Q}}^{-1}\mathbf{H}^H\mathbf{H}\tilde{\mathbf{Q}}\tilde{\mathbf{D}}^{1/2}\right|\right] \tag{5.14}$$

Il est à noter que l'espérance se fait sur $\mathbf{H}$ et $\tilde{\mathbf{Q}}$ puisqu'ils sont indépendants. Pour $\rho \gg 1$, on a:

$$C_{asympt} = E_{\mathbf{H},\tilde{\mathbf{Q}}}\left[\log_2\left|\frac{1}{\sigma_\xi^2}\tilde{\mathbf{D}}^{1/2}\tilde{\mathbf{Q}}^{-1}\mathbf{H}^H\mathbf{H}\tilde{\mathbf{Q}}\tilde{\mathbf{D}}^{1/2}\right|\right] \tag{5.15}$$

$$= m\log_2\left(\frac{1}{\sigma_\xi^2}\right) + \log_2\left|\tilde{\mathbf{D}}\right| - E_{\mathbf{H}}\left[\log_2\left|\mathbf{H}^H\mathbf{H}\right|\right] \tag{5.16}$$

$$= m\log_2\left(\frac{\rho}{m}\right) - \sum_{i=1}^{min(m,n)} E_{\mathbf{H}}\left[\log_2\lambda_i\right] \tag{5.17}$$

On obtient donc le même résultat que pour l'équation (4.47) avec $n \geq m$, c'est-à-dire que la capacité pour $\rho \gg 1$ est la même peu importe la qualité de l'estimation du canal lorsque $n \geq m$.

Pour le cas $m > n$, on a pour $\rho \gg 1$:

$$C_{asympt} = E_{\mathbf{H},\tilde{\mathbf{Q}}}\left[\log_2\left|\frac{1}{\sigma_\xi^2}\mathbf{H}\tilde{\mathbf{Q}}\tilde{\mathbf{D}}\tilde{\mathbf{Q}}^{-1}\mathbf{H}^H\right|\right] \tag{5.18}$$

$$= n\log_2\left(\frac{1}{\sigma_\xi^2}\right) - E_{\mathbf{H},\tilde{\mathbf{Q}}}\left[\log_2\left|\mathbf{H}\tilde{\mathbf{Q}}\tilde{\mathbf{D}}\tilde{\mathbf{Q}}^{-1}\mathbf{H}^H\right|\right] \tag{5.19}$$

En utilisant la décomposition singulière $\mathbf{H} = \mathbf{USV}$:

$$C_{asympt} = n\log_2\left(\frac{1}{\sigma_\xi^2}\right) - E_{\mathbf{H},\tilde{\mathbf{Q}}}\left[\log_2\left|\mathbf{H}\tilde{\mathbf{Q}}\tilde{\mathbf{D}}\tilde{\mathbf{Q}}^{-1}\mathbf{H}^H\right|\right] \tag{5.20}$$

$$= n\log_2\left(\frac{1}{\sigma_\xi^2}\right) - E_{\mathbf{H},\tilde{\mathbf{Q}}}\left[\log_2\left|\mathbf{SV}\tilde{\mathbf{Q}}\tilde{\mathbf{D}}\tilde{\mathbf{Q}}^{-1}\mathbf{V}^{-1}\mathbf{S}^H\right|\right] \tag{5.21}$$

Puis, avec l'inégalité d'Hadammard, on a:

$$\left|\mathbf{SV}\tilde{\mathbf{Q}}\tilde{\mathbf{D}}\tilde{\mathbf{Q}}^{-1}\mathbf{V}^{-1}\mathbf{S}^H\right| \leq \left|\mathbf{S}\tilde{\mathbf{D}}\mathbf{S}^H\right| \tag{5.22}$$

$$\leq \prod_{i=1}^{n}\lambda_i d_i \tag{5.23}$$

$$\leq \prod_{i=1}^{n}\frac{P}{n}\lambda_i \tag{5.24}$$

avec égalité si et seulement si $\tilde{\mathbf{Q}} = \mathbf{V}^{-1}$. Or, comme $\tilde{\mathbf{Q}}$ est indépendant de $\mathbf{H}$, on a $\tilde{\mathbf{Q}} \neq \mathbf{V}^{-1}$ avec une probabilité de 1 et on a donc:

$$E_{\mathbf{H},\tilde{\mathbf{Q}}}\left[\log_2\left|\mathbf{S}\mathbf{V}\tilde{\mathbf{Q}}\tilde{\mathbf{D}}\tilde{\mathbf{Q}}^{-1}\mathbf{V}^{-1}\mathbf{S}^H\right|\right] \;<\; E_{\mathbf{H},\tilde{\mathbf{Q}}}\left[\log_2\left|\mathbf{S}\tilde{\mathbf{D}}\mathbf{S}^H\right|\right] \tag{5.25}$$

Finalement, on obtient l'inégalité suivante:

$$C < n\log_2\left(\frac{\rho}{n}\right) + \sum_{i=1}^{min(m,n)} E_{\mathbf{H}}\left[\log_2\lambda_i\right] \tag{5.26}$$

En comparant l'inégalité (5.26) et l'équation (4.47) pour $m > n$, on conclue que la capacité avec erreur d'estimation infinie (corrélation nulle) sera toujours inférieure à la capacité avec estimation parfaite lorsque $m > n$ et $\rho \gg 1$.

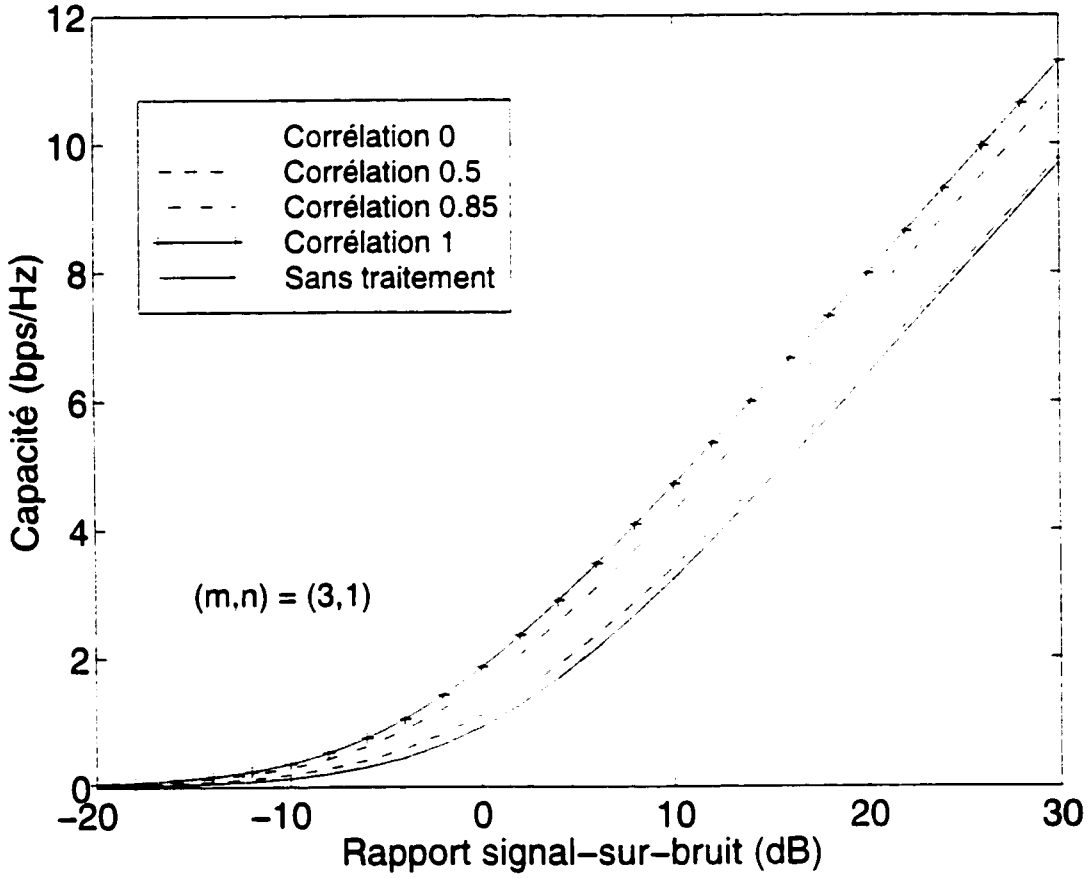


Figure 5.11: Capacité pour différentes valeurs de corrélation entre $\mathbf{H}$ et $\tilde{\mathbf{H}}$. $(m, n) = (3, 1)$.

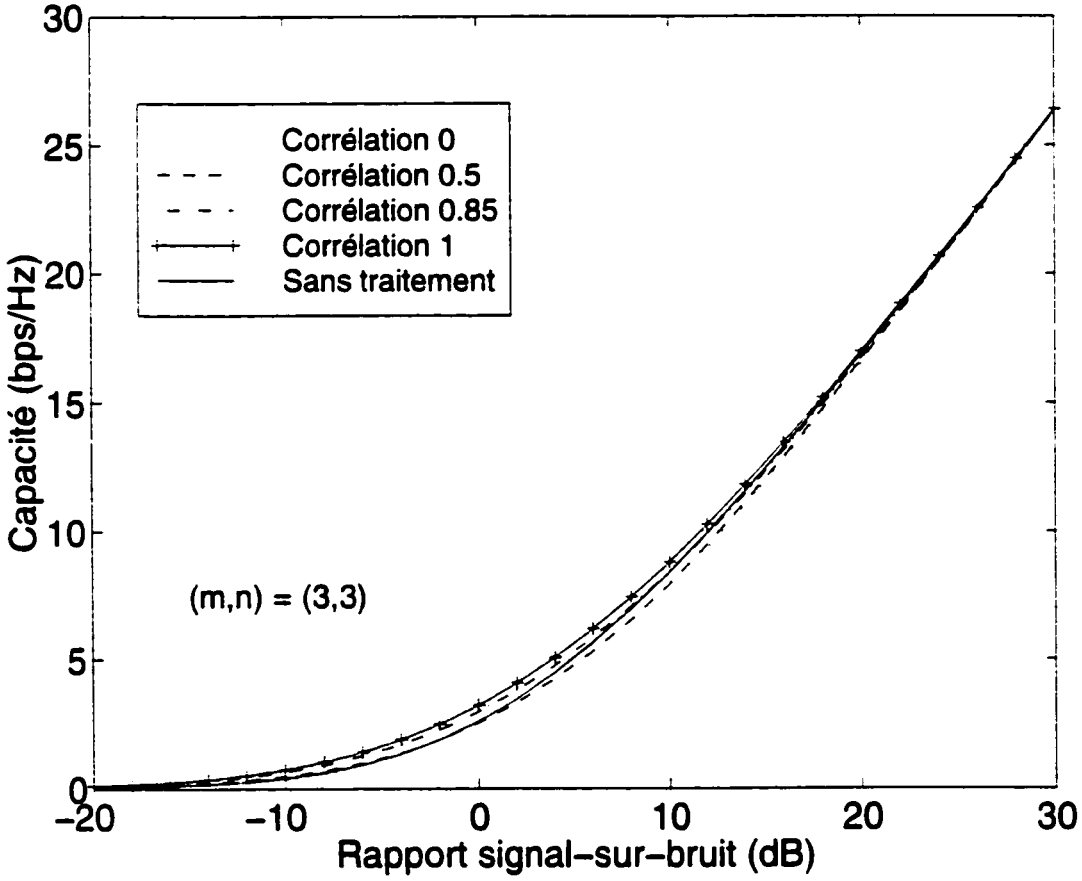


Figure 5.12: Capacité pour différentes valeurs de corrélation entre $\mathbf{H}$ et $\tilde{\mathbf{H}}$. $(m, n) = (3, 3)$.

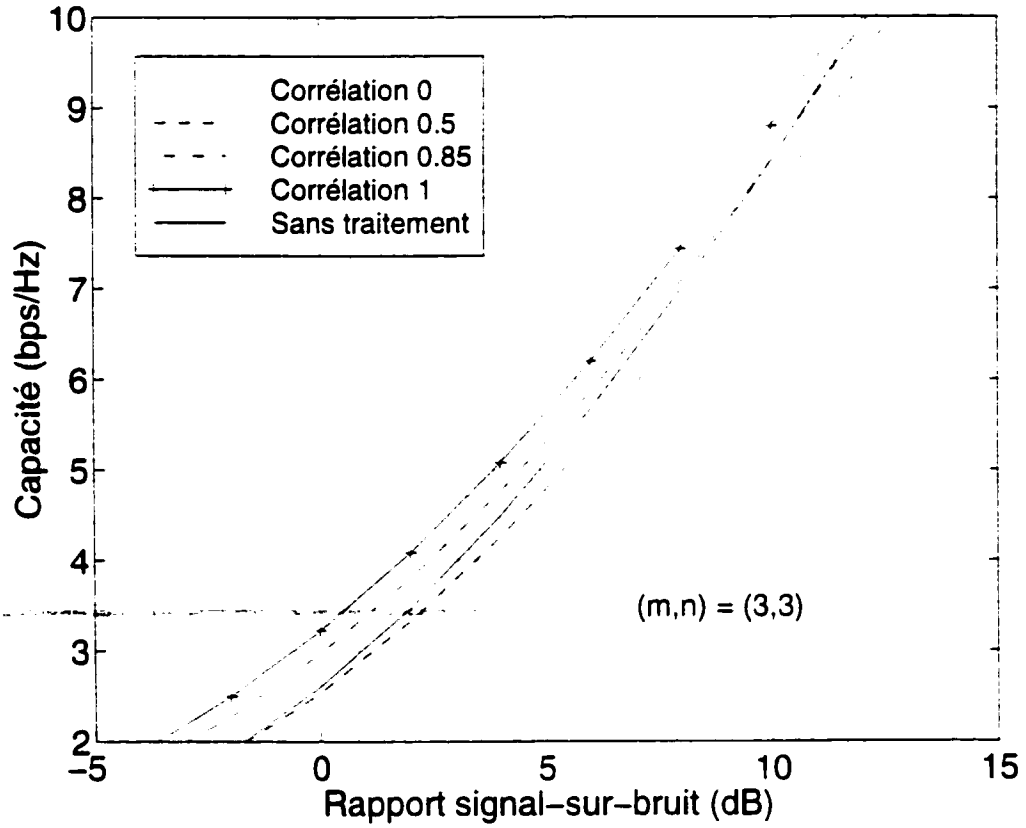


Figure 5.13: Capacité pour différentes valeurs de corrélation entre $\mathbf{H}$ et $\tilde{\mathbf{H}}$. $(m, n) = (3, 3)$ (agrandissement).

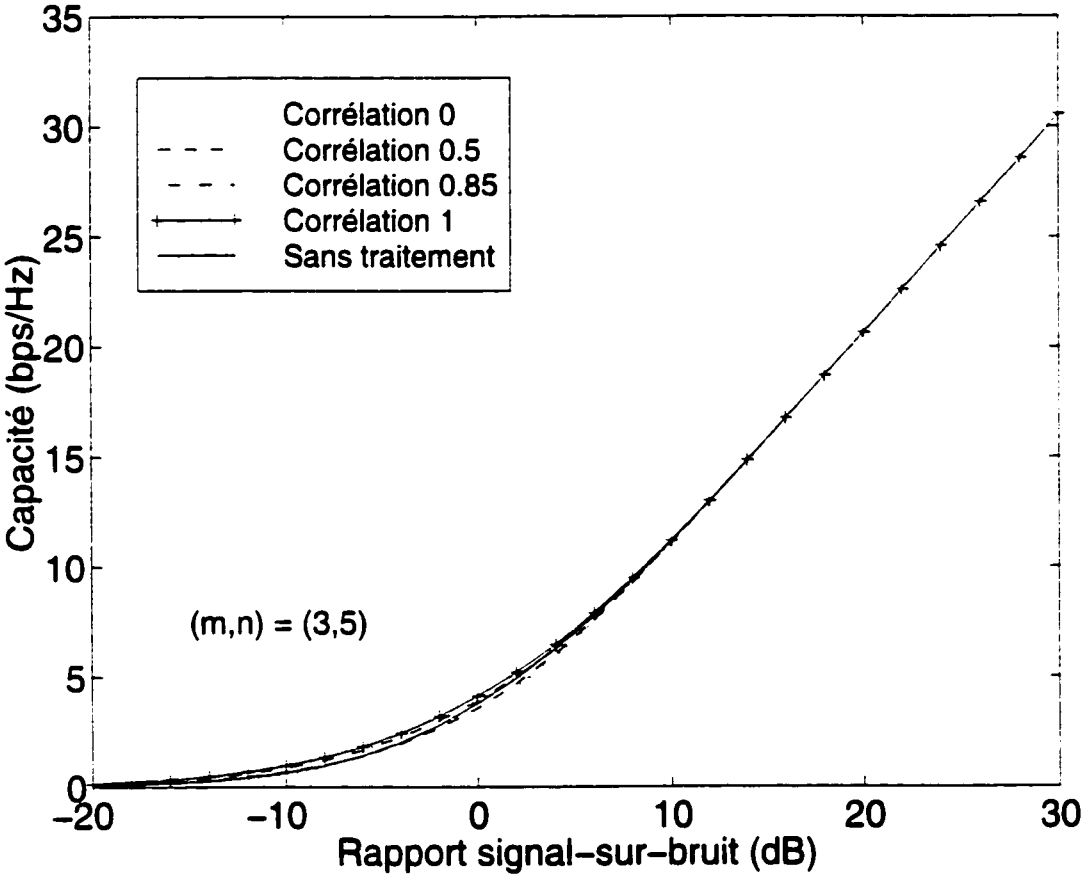


Figure 5.14: Capacité pour différentes valeurs de corrélation entre $\mathbf{H}$ et $\tilde{\mathbf{H}}$. $(m, n) \doteq (3, 5)$.

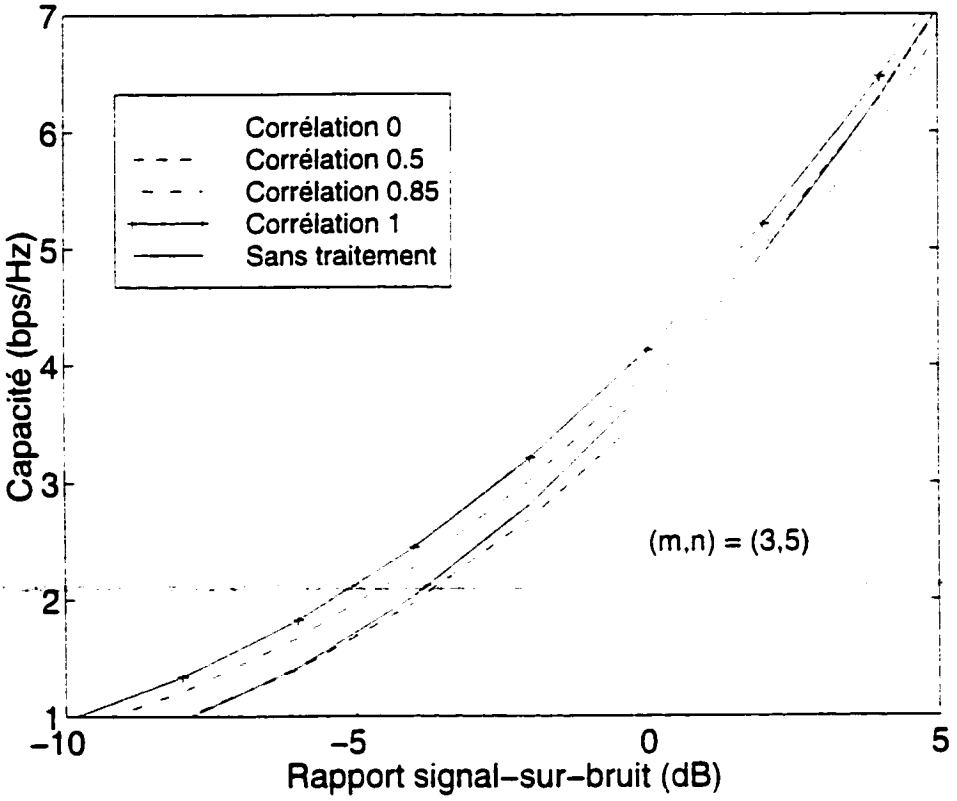


Figure 5.15: Capacité pour différentes valeurs de corrélation entre $\mathbf{H}$ et $\tilde{\mathbf{H}}$. $(m, n) = (3, 5)$ (agrandissement).

5.4 Conclusion

Dans ce chapitre, nous avons étudié l'effet de l'erreur d'estimation du canal sur la capacité de systèmes avec prétraitement à l'émetteur. Cette étude s'est faite par simulation et les résultats ont été présentés en fonction de la corrélation entre la matrice du canal et son estimation, et en fonction du rapport signal-sur-bruit. Nous avons obtenu des conclusions pour deux cas oppposés. Dans le cas où le nombre d'antennes émettrices est supérieur au nombre d'antennes réceptrices et que le rapport signal-sur-bruit est faible, nous avons conclu qu'une erreur d'estimation du canal a un effet important sur la capacité: la différence de capacité avec et sans erreur d'estimation est la plus élevée dans ce cas. Dans le cas où le nombre d'antennes réceptrices est supérieur au nombre d'antennes émettrices et que le rapport signal-sur-bruit est élevé, nous avons conclu qu'une estimation de qualité est nécessaire afin d'atteindre une capacité supérieure au cas sans prétraitement. Il est donc possible dans ce cas qu'un système sans prétraitement surpasse un système avec prétraitement si l'estimation du canal de ce dernier est erronée. Finalement, le comportement de la capacité avec erreur d'estimation infinie pour un grand rapport signal-sur-bruit ($\rho \gg 1$) a été étudié. Dans le cas $m > n$, la capacité avec erreur d'estimation infinie (corrélation nulle) sera toujours inférieure à la capacité avec estimation parfaite, alors qu'elle est la même peu importe la qualité de l'estimation du canal lorsque $n \geq m$. Nous concluons donc que le prétraitement est un avantage pour les systèmes où $m > n$ et le rapport signal-sur-bruit est faible, à la condition que l'estimation du canal par l'émetteur soit adéquate. Par contre, c'est justement dans une telle situation, c'est-à-dire lorsque le rapport signal-sur-bruit est faible, que les erreurs d'estimation du canal sont le plus probables, ce qui potentiellement pourrait limiter l'avantage du prétraitement au point où le gain obtenu ne justifie pas l'ajout de complexité au système.

Chapitre 6

Conclusions et suggestions de travaux de recherche

6.1 Sommaire

La diversité spatiale à l'émetteur et au récepteur est une technique qui semble très prometteuse pour les systèmes de communications dans un proche futur. Afin de connaître les avantages possibles de cette diversité en ce qui a trait à la capacité, nous avons étudié la capacité dans trois situations: le canal n'est connu qu'au récepteur, le canal est aussi connu à l'émetteur et finalement, l'émetteur connait le canal mais avec une erreur d'estimation.

Au chapitre trois, nous avons dérivé des expressions d'asymptotes pour la capacité. Nous avons conclu que la capacité est proportionnelle à $\min(m, n)$. Aussi, nous avons montré que lorsque l'émetteur ne connait pas le canal, un plus grand gain sur la capacité est obtenu en augmentant le nombre d'antennes réceptrices comparativement à une augmentation du nombre d'antennes émettrices. Nous avons aussi montré que la diversité soit à l'émetteur ou au récepteur ($\min(m, n) = 1$) est très peu sensible à la corrélation entre les coefficients d'atténuation des différentes antennes. Par contre, la corrélation entre les coefficients d'atténuation a un effet marqué sur un système avec diversité à l'émetteur et au récepteur. Finalement, la fonction de densité de probabilité de la capacité prise comme variable aléatoire

pour $\min(m, n) = 1$ a été dérivée.

Au chapitre quatre nous avons étudié le cas intéressant où l'émetteur possède une connaissance parfaite du canal, ce qui permet alors de modifier la distribution des symboles transmis afin de l'adapter au canal et ainsi augmenter la capacité du système. Un algorithme permettant le calcul de la matrice de covariance des symboles transmis et de la capacité résultante a été dérivé. Comme au chapitre deux, les asymptotes de capacité ont été dérivées. Nous avons déterminé que la capacité pour ces systèmes est la même dans les deux directions, contraitrement au cas où l'émetteur ne connaît pas le canal. Ceci était attendu, étant donné que la capacité du système avec connaissance parfaite du canal correspond à la définition fondamentale de la capacité et que celle-ci est symmétrique. Nous avons donc conclu qu'une augmentation du nombre d'antennes émettrices était aussi efficace au point de vue de la capacité qu'une augmentation du nombre d'antennes réceptrices. Finalement, nous avons remarqué que les asymptotes de la capacité avec, et sans, la connaissance du canal à l'émetteur pour $n \geq m$ étaient identiques. Il semble donc que la connaissance du canal à l'émetteur ait un effet limité dans ce cas. L'augmentation de complexité dû au traitement des symboles transmis et l'effet limité sur la capacité du traitement devront être considérés lors de la conception d'un système avec prétraitement à l'émetteur.

Au chapitre cinq, une erreur d'estimation du canal a été introduite à l'émetteur afin de déterminer son effet sur la capacité. Il a été montré que l'erreur d'estimation a un effet plus prononcé pour le cas $m > n$ et lorsque le rapport signal-sur-bruit est faible. En résumé, la connaissance de l'état du canal à l'émetteur peut mener à de plus grandes capacités, surtout lorsque le rapport signal-sur-bruit est faible et qu'il y a plus d'émetteurs que de récepteurs. Par contre, son avantage diminue grandement lorsque le nombre d'antennes réceptrices est plus élevé que le nombre d'antennes émettrices.

6.2 Suggestions de travaux de recherche

Au chapitre quatre, il a été démontré qu'un prétraitement avec matrice unitaire ne modifie pas la capacité par rapport à un système sans prétraitement. Il serait intéressant de vérifier si ce type de transformation pourrait avoir un avantage au niveau de la réception du signal. Par exemple, est-il possible d'appliquer une transformation unitaire à l'émetteur qui pourrait faciliter la tâche du récepteur, sans toutefois modifier la capacité? Aussi, la corrélation entre les antennes à été étudiée brièvement pour les deux cas extrêmes: sans corrélation et avec corrélation parfaite. Il serait possible d'étudier l'effet d'une corrélation quelconque dans l'intervalle [0, 1]. D'un point de vue pratique, on pourrait déterminer s'il est possible d'obtenir un algorithme plus rapide afin de déterminer la solution optimale de prétraitement. Il serait aussi intéressant d'obtenir des algorithmes non optimaux mais plus rapides et facilement réalisables. Enfin, l'étude de l'effet de l'estimation de l'état du canal sur la capacité est en grande partie obtenue par simulation. Une étude plus approfondie pourrait permettre d'obtenir les expressions des courbes de capacité en fonction de la corrélation de l'estimation, ou des approximations de ces courbes.

Annexe A

Calculs alternatifs des asymptotes pour $\min(m, n) = 1$

Nous présentons ici des dérivations alternatives à celles présentées au chapitre 3 afin d'obtenir les asymptotes de capacité lorsque le canal n'est connu qu'au récepteur pour le cas $\min(m, n) = 1$. Ces dérivations utilisent les coefficients de la matrice du canal $\mathbf{H}$ et non ses valeurs propres.

Le modèle du canal utilisé est le canal à évanouissement Rayleigh où les coefficients h_{ij} du canal sont de la forme $h_{ij} = \mathcal{N}(0, 1/2) + j\mathcal{N}(0, 1/2)$. La norme $X = |H_{ij}|$ de chaque coefficient est donc une variable Rayleigh normalisé avec $E\left[|H_{ij}|^2\right] = 2\sigma^2 = 1$ et la fonction de densité de probabilité de X est:

$$\begin{aligned} f_{\mathbf{X}}(x) &= \frac{x}{\sigma^2} e^{-x^2/2\sigma^2} \\ &= 2xe^{-x^2} \qquad \text{pour} \quad x \geq 0 \end{aligned} \tag{A.1}$$

La variable aléatoire $Y = X^2 = |H_{ij}|^2$ est donc une variable exponentielle avec une fonction de densité de probabilité donnée par:

$$\begin{aligned} f_{\mathbf{Y}}(y) &= \frac{f_{\mathbf{X}}(\sqrt{y})}{2\sqrt{y}} \\ &= e^{-y} \qquad \text{pour} \quad y \geq 0 \end{aligned} \tag{A.2}$$

A.1 Cas à antennes uniques de transmission et de réception $(m = n = 1)$

Dans le cas (1, 1), l'équation de capacité est donnée par:

$$\begin{aligned} C &= E_{\mathbf{H}}\left[\log_2\left(1+\rho|h|^2\right)\right] \\ &= E_{\mathbf{H}}\left[\log_2\left(1+\rho Y\right)\right] \end{aligned} \tag{A.3}$$

En appliquant l'opérateur espérance mathématique et en utilisant la fonction de densité de probabilité exponentielle de (A.2), l'équation devient:

$$C = \frac{1}{\ln(2)}\int_0^\infty \ln(1+\rho y)e^{-y}dy \tag{A.4}$$

En posant $u = 1 + \rho y$, l'équation (A.4) devient:

$$C = \frac{e^{1/\rho}}{\ln(2)}\int_1^\infty \ln(u)e^{-u/\rho}du \tag{A.5}$$

En utilisant l'identité:

$$\mathrm{Ei}(-x) \equiv -x\int_1^\infty e^{-tx}\ln(t)dt \qquad x > 0 \tag{A.6}$$

$$C = -\frac{e^{1/\rho}}{\ln(2)}\mathrm{Ei}(-1/\rho) \tag{A.7}$$

De même, en utilisant l'identité:

$$\mathrm{Ei}(x) \equiv \gamma - \ln(-x) - \sum_{k=1}^{\infty}\frac{x^k}{k\cdot k!} \qquad x < 0 \tag{A.8}$$

la capacité devient:

$$C = \frac{e^{1/\rho}}{\ln(2)}\left[\ln(\rho) - \gamma - \sum_{k=1}^{\infty}\frac{(-1/\rho)^k}{k\cdot k!}\right] \tag{A.9}$$

En ne gardant que les deux premiers termes de la série infinie, la limite inférieure de la capacité est donnée par:

$$C < \frac{e^{1/\rho}}{\ln(2)}\left[\ln(\rho) - \frac{1}{\rho} - \frac{1}{4\rho^2} - \gamma\right] \tag{A.10}$$

où $\gamma = 0.5772\ldots$ est la constante d'Euler. Ce résultat a été obtenu auparavant par Alouini et Goldsmith [AG99]. L'asymptote de la capacité est donnée par:

$$C_{asympt} = \log_2(\rho) - \frac{\gamma}{\ln(2)} \tag{A.11}$$

La démarche utilisée permet d'obtenir une limite inférieure aussi serrée que l'on désire en utilisant plus ou moins de termes de la série mais l'asymptote de la capacité aurait pu être obtenue plus rapidement de cette façon:

$$\begin{aligned} C_{asympt} &= \frac{1}{\ln(2)} \int_0^\infty \ln(\rho y) e^{-y} dy \\ &= \frac{1}{\ln(2)} \left[\ln(\rho) + \int_0^\infty \ln(y) e^{-y} dy \right] \\ &= \frac{1}{\ln(2)} [\ln(\rho) - \gamma] \end{aligned} \tag{A.12}$$

A.2 Cas avec une antenne émettrice et n antennes réceptrices $(1, n)$

Si les éléments de $\mathbf{H}$ sont indépendants. on trouve que l'argument du logarithme est:

$$\begin{vmatrix} 1+\rho|h_1|^2 & \rho h_1 h_2^* & \ldots & \rho h_1 h_n^* \\ \rho h_2 h_1^* & 1+\rho|h_2|^2 & \ldots & \rho h_2 h_n^* \\ \vdots & & \ddots & \vdots \\ \rho h_n h_1^* & \rho h_n h_2^* & \ldots & 1+\rho|h_n|^2 \end{vmatrix} = 1 + \rho(|h_1|^2 + |h_2|^2 + \ldots + |h_n|^2) \tag{A.13}$$

La capacité est alors donnée par:

$$\begin{aligned} C &= E_{\mathbf{H}} \left[\log_2 \left(1 + \rho \sum_{i=1}^{n} |h_i|^2 \right) \right] \\ &= E_{\mathbf{H}} \left[\log_2 \left(1 + \rho \sum_{i=1}^{n} Y \right) \right] \\ &= E_{\mathbf{H}} \left[\log_2 (1 + \rho Z_n) \right] \end{aligned} \tag{A.14}$$

Comme la variable Y a une distribution exponentielle, la variable $Z_n = \sum_{i=1}^{n} Y$ a une distribution d'Erlang:

$$f_{\mathbf{Z_n}}(z) = \frac{z^{n-1}e^{-z}}{(n-1)!} \tag{A.15}$$

En appliquant l'espérance mathématique et en utilisant la fonction de densité de probabilité d'Erlang d'ordre n de (A.15), l'équation devient:

$$C = \frac{1}{\ln(2)} \int_0^{\infty} \ln(1+\rho z) \frac{z^{n-1}e^{-z}}{(n-1)!} dz \tag{A.16}$$

et donc:

$$C_{asympt} = \frac{1}{\ln(2)} \int_0^{\infty} [\ln(\rho) + \ln(z)] \frac{z^{n-1}e^{-z}}{(n-1)!} dz \tag{A.17}$$

En utilisant l'identité:

$$\int_0^{\infty} z^{n-1}e^{-z} dz \equiv \Gamma(n) = (n-1)! \tag{A.18}$$

et l'identité suivante:

$$\int_0^{\infty} \ln(z) z^{n-1}e^{-z} dz = (n-1)! \left[1 - \frac{1}{2} + \ldots - \frac{1}{n-1} - \gamma\right] \tag{A.19}$$

l'asymptote de la capacité pour le cas $(1, n)$ s'exprime alors par:

$$C_{asympt} = \frac{1}{\ln(2)} \left[\sum_{i=1}^{n-1} \frac{1}{i} - \ln(\rho) - \gamma\right] \tag{A.20}$$

A.3 Cas à m antennes émettrices et une antenne réceptrice $(m, 1)$

Si les éléments de $\mathbf{H}$ sont indépendants, on a:

$$\begin{aligned} C &= E_{\mathbf{H}} \left[\log_2 \left(1 - \frac{\rho}{m} \sum_{i=1}^{m} |h_i|^2\right)\right] \\ &= E_{\mathbf{H}} \left[\log_2 \left(1 - \frac{\rho}{m} Z_m\right)\right] \end{aligned} \tag{A.21}$$

où Z_m est une variable aléatoire décrite par la distribution d'Erlang d'ordre m telle que définie dans la section précédente. En remplaçant ρ par ρ/m dans les dérivations pour le cas $(1, n)$, on trouve l'asymptote suivante:

$$C_{asympt} = \frac{1}{\ln(2)} \left[\sum_{i=1}^{m-1} \frac{1}{i} - \ln(m) + \ln(\rho) - \gamma \right] \tag{A.22}$$

Annexe B

Les valeurs propres non nulles des matrices $\mathbf{HH}^H$ et $\mathbf{H}^T\mathbf{H}^{T^H}$ sont identiques (preuves)

Nous dérivons la preuve que les valeurs propres non nulles des matrices $\mathbf{HH}^H$ et $\mathbf{H}^T\mathbf{H}^{T^H}$ sont identiques. La décomposition de $\mathbf{H}$ en valeur singulière donne:

$$\mathbf{H}^{n\times m} = \mathbf{U}^{n\times n}\mathbf{S}^{n\times m}\mathbf{V}^{m\times m} \tag{B.1}$$

où $\mathbf{U}$ et $\mathbf{V}$ sont unitaires et $\mathbf{S}$ est diagonale. En utilisant ($\mathbf{VV}^H = \mathbf{I}_m$) et ($\mathbf{U}^H = \mathbf{U}^{-1}$), on obtient:

$$\begin{aligned} \mathbf{HH}^H &= \mathbf{USV}(\mathbf{USV})^H \\ &= \mathbf{USVV}^H\mathbf{S}^H\mathbf{U}^H \\ &= \mathbf{USS}^H\mathbf{U}^{-1} \end{aligned} \tag{B.2}$$

où les éléments de $\mathbf{SS}^H$ sont les valeurs propres de $\mathbf{HH}^H$. La matrice du canal inverse est $\mathbf{H}^T$ et on a:

$$\begin{aligned}
\mathbf{H}^T\mathbf{H}^{TH} &= (\mathbf{USV})^T(\mathbf{USV})^{TH} \\
&= \mathbf{V}^T\mathbf{S}^T\mathbf{U}^T\mathbf{U}^{TH}\mathbf{S}^{TH}\mathbf{V}^{TH} \qquad \text{où} \quad \mathbf{U}^T\mathbf{U}^{TH} = (\mathbf{U}^{-1}\mathbf{U})^T = \mathbf{I}_n \\
&= \mathbf{V}^T\mathbf{S}^T\mathbf{S}^{TH}\mathbf{V}^{T-1} \qquad \text{(B.3)}
\end{aligned}$$

où les éléments de $\mathbf{S}^T\mathbf{S}^{TH}$ sont les valeurs propres de $\mathbf{H}^T\mathbf{H}^{TH}$. Finalement, comme $\mathbf{S}$ est diagonale, les éléments non nuls de $\mathbf{SS}^H$ et $\mathbf{S}^T\mathbf{S}^{TH}$ sont identiques et les valeurs propres non nulles des matrices $\mathbf{HH}^H$ et $\mathbf{H}^T\mathbf{H}^{TH}$ sont identiques.

Annexe C

Calcul de $\sum_{i=1}^{\min(m,n)} E_{\mathbf{H}}\left[\log_2 \lambda_i\right]$

Ce résultat est utilisé au chapitre trois afin d'obtenir les expressions de capacité asymptotique. L'espérance mathématique que l'on désire calculer est donnée par:

$$\sum_{i=1}^{\min(m,n)} E_{\mathbf{H}}\left[\log_2 \lambda_i\right] = \sum_{i=1}^{\min(m,n)} \int_0^{\infty} \cdots \int_0^{\infty} \log_2 \lambda_i \frac{f_{\Lambda}(\lambda_1, \ldots, \lambda_{\min(m,n)})}{\min(m,n)!} d\lambda_1 \ldots d\lambda_{\min(m,n)} \tag{C.1}$$

où la distribution des valeurs propres est obtenue en modifiant la distribution (voir Edelman [Ede89]). pour $\lambda_1 > \lambda_2 \ldots > \lambda_{\min(m,n)} > 0$:

$$f_{\Lambda}(\lambda_1, \lambda_2, \ldots, \lambda_{\min(m,n)}) = \frac{\exp(-\sum_{i=1}^{\min(m,n)} \lambda_i) \prod_{i=1}^{\min(m,n)} \lambda_i^{m-n} \prod_{i=1, i<j}^{\min(m,n)} (\lambda_i - \lambda_j)^2}{\prod_{i=1}^{\min(m,n)-1} i! \prod_{i=1}^{\min(m,n)} (\max(m,n) - i)!} \tag{C.2}$$

Les modifications apportées à la distribution d'Edelman consistent en deux changements simples:

- 1. Utiliser des variables gaussiennes avec une variance égale à $1/2$ pour la matrice du canal dans les dérivations: Edelman utilise une variance unitaire.
- 2. Modifier l'expression de la distribution afin de tenir compte des deux cas. $n > m$ et $m > n$: Edelman ne l'exprime que pour le cas où $n > m$.

Notez que la distribution des valeurs propres $f_\Lambda(\lambda)$ est divisée par $\min(m,n)!$ afin d'éliminer l'ordre abitraire $\lambda_1 > \lambda_2 > \ldots > \lambda_{\min(m,n)}$. Aussi, étant donnée la symmétrie de $f_\Lambda(\lambda)$, on remarque que chacune des $\min(m,n)$ intégrales multiples donne le même résultat. On obtient donc:

$$\begin{aligned}\sum_{i=1}^{\min(m,n)} E_{\mathbf{H}}[\log_2 \lambda_i] &= \frac{\min(m,n)}{\min(m,n)!} E_{\mathbf{H}}[\log_2 \lambda_1] \\ &= \frac{1}{(\min(m,n)-1)!} E_{\mathbf{H}}[\log_2 \lambda_1] \end{aligned} \tag{C.3}$$

Pour $\min(m,n) = 1$, en posant $\max(m,n) = t$, le calcul a déjà été effectué:

$$E_{\mathbf{H}}[\log_2 \lambda_1] = \frac{1}{\ln(2)} \sum_{i=1}^{t-1} \frac{1}{i} - \gamma \tag{C.4}$$

Pour $\min(m,n) = 2$, on a:

$$\begin{aligned} E_{\mathbf{H}}[\log_2 \lambda_1] &= \frac{1}{\ln(2)} \int_0^\infty \int_0^\infty \ln(\lambda_1) \frac{e^{-\lambda_1-\lambda_2}(\lambda_1\lambda_2)^{t-2}(\lambda_1-\lambda_2)^2}{(t-1)!(t-2)!} d\lambda_1 d\lambda_2 \\ &= \frac{1}{(t-1)!(t-2)!} \left[t! \left(\sum_{i=1}^{t} \frac{1}{i} - \gamma \right) \Gamma(t-1) \right. \\ &\quad \left. -2(t-1)! \left(\sum_{i=1}^{t-1} \frac{1}{i} - \gamma \right) \Gamma(t) + (t-2)! \left(\sum_{i=1}^{t-2} \frac{1}{i} - \gamma \right) \Gamma(t) \right] \end{aligned} \tag{C.5}$$

où $\Gamma(x+1) = x!$. On obtient donc pour $\min(m,n) = 2$:

$$E_{\mathbf{H}}[\log_2 \lambda_1] = \frac{1}{\ln(2)} \left[2 \sum_{i=1}^{t-2} \frac{1}{i} + \frac{1}{t-1} - 2\gamma \right] \tag{C.6}$$

De la même façon, on obtient pour $\min(m,n) = 3$:

$$\frac{1}{2} E_{\mathbf{H}}[\log_2 \lambda_1] = \frac{1}{\ln(2)} \left[3 \sum_{i=1}^{t-3} \frac{1}{i} + \sum_{i=1}^{2} \frac{1}{t-i} - 3\gamma \right] \tag{C.7}$$

L'équation générale peut s'obtenir par extrapolation:

$$\begin{aligned} \sum_{i=1}^{\min(m,n)} E_{\mathbf{H}}[\log_2 \lambda_i] &= \frac{1}{\ln(2)} \left[\min(m,n) \sum_{i=1}^{|m-n|} \frac{1}{i} \right. \\ &\quad \left. + \sum_{i=1}^{\min(m,n)-1} \frac{i}{\max(m,n)-i} - \min(m,n)\gamma \right] \end{aligned} \tag{C.8}$$

L'équation (C.8) a été vérifiée numériquement à l'aide du logiciel mathématique Maple.

Annexe D

Calcul de $\sum_{i=1}^{\min(m,n)} E_{\mathbf{H}}[\lambda_i]$

Ce résultat est utilisé au chapitre quatre afin de calculer le gain obtenu suite au prétraitement.

Par symmétrie (voir l'annexe C), on a:

$$\begin{aligned}\sum_{i=1}^{\min(m,n)} E_{\mathbf{H}}[\lambda_i] &= \frac{\min(m,n)}{\min(m,n)!} E_{\mathbf{H}}[\lambda_1] \\ &= \frac{1}{(\min(m,n)-1)!} E_{\mathbf{H}}[\lambda_1] \end{aligned} \tag{D.1}$$

Pour $\min(m,n) = 1$, on a:

$$\begin{aligned} E_{\mathbf{H}}[\lambda_1] &= \int_0^\infty \lambda_1 e^{-\lambda_1} \lambda_1^{t-1} d\lambda_1 \\ &= \frac{\Gamma(t+1)}{(t-1)!} = t \end{aligned} \tag{D.2}$$

Pour $\min(m,n) = 2$, on a:

$$\begin{aligned} E_{\mathbf{H}}[\lambda_1] &= \int_0^\infty \int_0^\infty \lambda_1 \frac{e^{-\lambda_1-\lambda_2}(\lambda_1\lambda_2)^{t-2}(\lambda_1-\lambda_2)^2}{(t-1)!(t-2)!} d\lambda_1 d\lambda_2 \\ &= \frac{1}{(t-1)!(t-2)!}\left[\Gamma(t+2)\Gamma(t-1) - 2\Gamma(t+1)\Gamma(t) - \Gamma(t)\Gamma(t+1)\right] \\ &= 2t \end{aligned} \tag{D.3}$$

De la même façon, pour $\min(m,n) = 3$:

$$\frac{1}{2} E_{\mathbf{H}}[\lambda_1] = 3t \tag{D.4}$$

L'extrapolation donne:

$$\sum_{i=1}^{\min(m,n)} E_{\mathbf{H}}[\lambda_i] = \min(m,n)\max(m,n) = mn \tag{D.5}$$

L'équation (D.5) a été vérifiée également numériquement à l'aide du logiciel mathématique Maple.

Bibliographie

[AG99] M. Alouini and A. J. Goldsmith. Capacity of rayleigh fading channels under different adaptive transmission and diversity-combining techniques. *IEEE Transactions on Vehicular Technology*. pages 1165 1181. juillet 1999.

[And58] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley and Sons. New York. 1958.

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience. New York. 1991.

[Ede89] A. Edelman. *Eigenvalues and Condition Numbers of Random Matrices*. Thèse de doctorat. MIT. Cambridge. 1989.

[FG98] G. J. Foschini and M. J. Gans. On limits of wireless communications in a fading environment when using multiple antennas. *Wireless Personal Communications*. pages 311 335. volume 6 1998.

[Kul68] S. Kullback. *Information Theory and Statistics*. Dover. New York. 1968.

[Lan69] P. Lancaster. *Theory of Matrices*. Academic Press. New York. 1969.

[LGon] A. Leon-Garcia. *Probability and Random Processes for Electrical Engineering*. Addison-Wesley. New York. 1994. 2ième édition.

[MF70] R. K. Mueller and G. J. Foschini. The capacity of linear channels with additive gaussian noise. *The Bell System Technical Journal*. pages 81 94. janvier 1970.

[NLTW98] A. Narula, M. J. Lopez, M. D. Trott, and G. W. Wornell. Efficient use of side information in multiple-antenna data transmission over fading channels. *IEEE Journal on Selected Areas in Communications*, pages 1423–1436, octobre 1998.

[NTW99] A. Narula, M. D. Trott, and G. W. Wornell. Performance limits of coded diversity methods for transmitter antenna arrays. *IEEE Transactions on Information Theory*, pages 2418–2433, novembre 1999.

[Pro95] J. G. Proakis. *Digital Communications*. McGraw Hill, New York, 1995.

[Rap96] T. S. Rappaport. *Wireless Communications, Principles and Practice*. Prentice Hall, New Jersey, 1996.

[Tel99] E. Telatar. Capacity of multi-antenna gaussian channels. *European Transactions on Telecommunications*, 10(6): 585–595, novembre-décembre 1999.

[Win87] J. H. Winters. On the capacity of radio communications systems with diversity in a rayleigh fading environment. *IEEE Journal on Selected Areas in Communications*, pages 871–878, juin 1987.

MQ

66043

UMI

MICROFILMED 2002

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

PERMISSION DE REPRODUIRE ET DE DISTRIBUER LA THÈSE

PERMISSION TO REPRODUCE AND DISTRIBUTE THE THESIS

NOM DE L'AUTEUR / NAME OF AUTHOR:	GEORGEFF, Jennifer
ADRESSE POSTALE / MAILING ADDRESS:	1812 - 1541 River side Dr. Ottawa, Ontario K1G 4E2
GRADE / DEGREE: M.SC. (BIOLOGY)	**ANNÉE D'OBTENTION / YEAR GRANTED** 2001
TITRE DE LA THÈSE / TITLE OF THESIS: MYCORRHIZAL RESPONSIVENESS OF CULTIVARS AND WILD VARIETIES OF SWITCHGRASS, PANICUM VIRGATUM L.	

L'auteur permet, par la présente, la consultation et le prêt de cette thèse en conformité avec les règlements établis par le bibliothécaire en chef de l'Université d'Ottawa. L'auteur autorise aussi l'Université d'Ottawa, ses sumccesseurs et cessionnaires, à reproduire cet exemplaire par photograhie ou photocopie pour fins de prêt ou de vente au prix coûtant aux bibliothèques ou aux chercheurs qui en feront la demande.

Les droits de publication par tout autre moyen et pour vente au public demeureront la propriété de l'auteur de la thèse sous réserve des règlements de l'Université d'Ottawa en matière de publication de thèses.

The author hereby permits the consultation and the lending of this thesis pursuant to the regulations established by the Chief Librarian of the University of Ottawa. The author also authorizes the University of Ottawa, its successors and assignees, to make reproductions of this copy by photographic means or by photocopying and to lend or sell such reproductions at cost to libraries and to scholars requesting them.

The right to publish the thesis by other means and to sell it to the public is reserved to the author, subject to the regulations of the University of Ottawa governing the publication of theses.

N.B. LE MASCULIN COMPREND ÉGALEMENT LE FÉMININ

Nov 9 2001

DATE

(AUTEUR) SIGNATURE (AUTHOR)



Université d'Ottawa · University of Ottawa

Université d'Ottawa · University of Ottawa

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

GEORGEFF, Jennifer

AUTEUR DE LA THÈSE - AUTHOR OF THESIS

M.Sc. (Biology)

GRADE - DEGREE

Department of Biology

FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

Mycorrhizal Responsiveness of Cultivars and Wild Varieties of Switchgrass, *Panicum Virgatum* L.

Christiane Charest

DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

J. Arnason

N. Cappuccino

B. Philogène

J.-M. De Koninck, Ph.D.

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

SIGNATURE

DEAN OF THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

MYCORRHIZAL RESPONSIVENESS OF CULTIVARS AND WILD VARIETIES OF SWITCHGRASS, *PANICUM VIRGATUM* L.

JENNIFER GEORGEFF

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
University of Ottawa
in partial fulfillment of the requirements for the
MSc degree in the

Ottawa-Carleton Institute of Biology

August 2001

© **Jennifer Georgeff, Ottawa, Canada, 2001**

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-66043-5

Canada

Abstract

This study focuses on switchgrass, *Panicum virgatum* L., a North American tallgrass prairie species that has been used in agriculture for only 50 years. We hypothesized that wild variants of this species are more mycorrhizal dependent than the cultivars due to selection against mycorrhizae in agricultural conditions. Mycorrhizae, naturally occurring associations between roots and symbiotic fungi, are known to benefit plants in most environments. These associations have been proposed as a means for crop improvement in agriculture. However, current agricultural practices tend to diminish the potential for crops to benefit from mycorrhizae. Two greenhouse experiments with a factorial design were performed using switchgrass inoculated or not with *Glomus intraradices* Schenck and Smith. In the first experiment, four wild varieties (PH, ONP, Ojibway and Pterophylla) and six cultivars (Forestburg, Summer, Shelter, Caddo, NU and Trailblazer) were grown for 12 weeks. From those, 3 wild types (PH, ONP and Ojibway) and 3 cultivars (Forestburg, Caddo and NU) were chosen for the second experiment. Switchgrass plants were analyzed for various physiological, root and mineral parameters in order to assess their overall mycorrhizal dependency (MD). The average mycorrhizal colonization of 37% was not different between the wild and cultivated varieties. Variety was the most significant source of variation for most of the measured parameters. It was found that each variety responded to mycorrhizal colonization by modifying different parameters. The responses of the wild group were more variable than those of the cultivated group. Forestburg, a cultivar which has not been agriculturally selected, and wild variety ONP, responded positively to mycorrhizal colonization by modifying their physiological and root parameters while mineral levels were increased in the cultivar Caddo. A cluster analysis was performed on the effect size of the mycorrhizal treatment to establish the groupings among varieties. Cluster analysis results from the second experiment

show one group with Forestburg (non-selected cultivar), ONP (wild), two varieties which improved their condition with mycorrhizal colonization, and PH (wild), which had a neutral response. The second group included NU (cultivated) and Caddo (cultivated), which had neutral responses and Ojibway (wild), which responded negatively to mycorrhizae. The MD of wild varieties in this study was lower than expected, since they showed inconsistent responses to mycorrhizae. Because of this, our hypothesis that mycorrhizal dependency of wild varieties is higher than that of cultivars cannot be strictly retained. However, the two cultivars that had undergone many rounds of agricultural selection (NU and Caddo) did not show positive growth responses to mycorrhizal colonization. This suggests that agricultural practices may have some detrimental effect on the natural ability of switchgrass to benefit from mycorrhizae. Since some switchgrass varieties clearly have positive responses to mycorrhizae, these types could be used in selective breeding programs to enhance the beneficial impact of mycorrhizae and be included in sustainable agriculture practices.

Résumé

Cette étude s'intéresse au panicum, *Panicum virgatum* L., une graminée indigène des prairies de l'Amérique du Nord qui n'est cultivée que depuis les 50 dernières années. Nous avons postulé que les variétés sauvages du panicum sont plus dépendantes des mycorhizes que les cultivars, en raison de la sélection agricole. Les mycorhizes, associations naturelles entre racines et champignons symbiotiques, sont reconnues pour bénéficier aux plantes dans la plupart des conditions environmentales. Eu égard à ces bénéfices, il a été proposé de les utiliser pour l'amélioration des plantes agricoles. Les pratiques usuelles en agriculture ne prennent pas en compte les bénéfices potentiels de la symbiose mycorhizienne, et même les contrecarrent. Deux expériences en serre ont été réalisées selon un plan factoriel avec le panicum inoculé ou non avec *Glomus intraradices* Schenck et Smith. Dans une première expérience, quatre variétés sauvages (PH, ONP, Ojibway et Pterophylla) et six sélectionnées (Forestburg, Summer, Shelter, Caddo, NU et Trailblazer) ont cru pendant douze semaines. Parmi celles-ci, 3 variétés sauvages (PH, ONP et Ojibway) et 3 sélectionnées (Forestburg, Caddo et NU) ont été choisies pour une deuxième expérience. Les plantes ont été analysées à partir de certains paramètres physiologiques, racinaires et de minéraux afin d'estimer leur dépendance mycorhizienne (DM). La colonisation mycorhizienne moyenne de 37% était semblable parmi les variétés sauvages et sélectionnées. La variété s'est révélé être le facteur le plus significatif de la réponse mycorhizienne. Il a été trouvé que chaque variété répond différemment à la mycorhization en fonction des paramètres physiologiques. Lorsque mycorhizés, Forestburg (cultivar non-sélectionné) et la variété sauvage ONP avaient des réponses physiologiques et racinaires positives, alors que Caddo (cultivar) avait une teneur accrue en minéraux. En général, les variétés sauvages avaient des réponses plus variables. Une analyse de groupement a été réalisée

sur l'ampleur de l'effet du traitement mycorhizien afin de répartir les variétés en groupes. Le premier groupe inclut Forestburg (cultivar) et ONP (sauvage) qui répondaient positivement aux mycorhizes, ensemble avec PH (sauvage) qui était plus ou moins neutre. Le deuxième groupe se compose de NU et Caddo (cultivars) dont les réponses semblaient plutôt neutres, avec Ojibway (sauvage) qui répondait de façon négative à la colonisation mycorhizienne. La dépendance mycorhizienne trouvée dans cette étude est moins élevée que celle anticipée, en raison de réponses inconstantes des variétés sauvages. L'hypothèse que les variétés sauvages ont une DM plus élevée que celle des cultivars ne peut être strictement retenue. Les deux cultivars les plus sélectionnés (NU et Caddo) se retrouvent ensemble dans le groupe qui répondait le moins positivement à la mycorhization. Ceci suggère que les pratiques agricoles ne favorisent pas et même contrecarrent la capacité du panicum à bénéficier des associations mycorhiziennes. Puisque certaines variétés de panicum ont clairement répondu de façon positive à la mycorhization, celles-ci pourraient être utilisées dans des programmes de sélection agricole pour accroître les bénéfices de la mycorhization et ainsi être incluses dans les pratiques d'agriculture durable.

Acknowledgements

I am very grateful to my supervisor Dr. Christiane Charest as well as the members of my advisory commitee, Dr. J.T. Arnason, Dr. N. Cappuccino, and Dr. A. McElroy. The unwavering support that I received from all of you was very encouraging.

I would also like to express my appreciation to all of the people who aided in this study, Drs. D. Johnson and S. Findlay at U of O, Drs. A. McElroy, Y. Dalpé, G. Fedak, H. Dinel, M. Morrison, Ms. S. Séguin, and S. Kritenko, at ECORC, and Ms. W. Abdi at the J.J. Hatch Laboratory. The friendly help and access to research equipment were encouraging and essential to this accomplishment.

Switchgrass seed was donated from Dr. J. Dickerson (USDA), Dr. C. Taliaferro, Dr. A. Boe, Dr. A. McElroy, the Ojibway Nature Centre, and the Ontario Native Plants Co.

During the course of my studies at the University of Ottawa I have had the pleasure of getting to know many people without whom this project could not have been completed. I would like to thank greenhouse technician Ms. Huguette Allard, sympathetic colleagues Rob Beiko, Candace Webb, Jean-Patrick Toussaint, Steve Stockton, Shannon Binns as well as everyone who helped to harvest the plants and to make the last two years so enjoyable.

Finally, I would like to thank my parents for their support over my many years of academic studies, and Rob for his personal and professional encouragement.

Table of Contents

Abstract ii
Résumé iv
Acknowledgements vi
Table of Contents vii
List of Figures x
List of Tables xi
List of Abbreviations xii

Introduction

1.0- Mycorrhizal Symbiosis 1
1.1- Symbiosis 1
1.2- Mycorrhizal Fungi 2
1.2.1 Vesicular Arbuscular Mycorrhiza Anatomy 3
1.2.2 Mode of Colonization 7
1.2.3 Benefits of Mycorrhizal Symbiosis to Mineral Nutrition and Stress 8
1.3- Plant Response to Mycorrhizal Colonization 10
1.3.1 Carbon Use by Mycorrhizae 11
1.3.2 Cost/Benefit of Mycorrhizal Symbiosis 12
1.3.3 Symbiotic Continuum of Mycorrhizae 14
1.3.4 Mycorrhizal Dependency 14
2.0- Agriculture and Mycorrhizae 18
2.1- Effects of Agriculture on Mycorrhizae 19
2.2- Cultivars and Mycorrhizae 20
2.3- Selection in Agricultural Environments: Wild Varieties vs. Cultivars 21
3.0- Switchgrass 23
3.1- Genetics 26
3.2- Agricultural Selection 27
3.3- Mycorrhizae and Switchgrass 28
4.0- Hypothesis and Objectives 30

Materials and Methods

1.0- Switchgrass Varieties 31
2.0- Seeds 33
3.0- Preparation for Germination 35
4.0- Planting and Soil Preparation 36
5.0- Soil Medium 37
6.0- Growth Conditions 38
6.1- Growth Measurements 39
6.2- Chlorophyll Determination 40

7.0- Harvest 41
8.0- Acid Phosphatase Determination 42
9.0- Root Staining and Percent Colonization Determination 42
10.0- Root Morphology Determination 43
11.0- Analysis of Carbon and Nitrogen 44
12.0- Analysis of Other Macronutrients 44
13.0- Chromosome Count 45
14.0- Statistical Analyses 45

Results

1.0- Germination 49
2.0- Chromosome Numbers 49
3.0- Growth Rate 53
4.0- Chlorophyll Concentration 53
5.0- Mycorrhizal Colonization 55
6.0- Physiological Data 57
6.1- Mass 57
6.2- Root to Shoot Ratio 58
6.3- Plant Height 59
6.4- Number of Tillers 60
6.5- Summary of Physiological Parameters 61
7.0- Acid Phosphatase 69
8.0- Root Architecture 73
8.1- Total Root Length 73
8.2- Average Root Diameter 74
8.3- Number of Root Tips 75
8.4- Percentage of Small Roots 76
8.5- Percentage of Large Roots 77
8.6- Summary of Root Parameters 78
9.0- Mineral Content 87
9.1- Carbon 87
9.2- Nitrogen 88
9.3- Phosphorus 89
9.4- Calcium 89
9.5- Potassium 90
9.6- Magnesium 90
9.7- Overall Mineral Effects 91
10.0- Statistical Analysis 101
10.1- Cluster Analysis 101
10.2- Canonical Discriminant Analysis 105

Discussion

1.0- Switchgrass Varieties..... 110
1.1- Physiological Differences Among Varieties..... 110
1.2- Chromosome Numbers..... 112
2.0- Mycorrhizal Response..... 113
2.1- Colonization..... 113
2.2- Physiological Responses to Mycorrhizae..... 114
2.3- Root Architecture Responses to Mycorrhizae..... 118
2.4- Mineral Responses to Mycorrhizae..... 121
2.5- Cost/Benefit Analysis..... 124
2.6- Mycorrhizal Responsiveness of Switchgrass Varieties..... 127
2.7- Wild vs. Cultivated..... 129
2.8- Implications to Agriculture..... 132
3.0- Conclusions..... 135

Appendix..... 137

References..... 138

List of Figures

Introduction

Figure 1.1	Anatomy of vesicular abuscular mycorrhizae	6
Figure 1.2	Model of plant 'benefit' and % colonization.....	13
Figure 1.3	Symbiotic continuum for mycorrhizae.....	14
Figure 1.4	Diagram of switchgrass, including panicle and rhizomes..	23

Materials and Methods

Figure 2.1	Switchgrass plants.....	47
Figure 2.2	*G. intraradices* colonization.....	48

Results

Figures 3.1 -3.10	Growth rates of each variety/treatment as measured by plant height.....	52
Figure 3.11	The percentage of mycorrhizal colonization of switchgrass with *G. intraradices* from the first experiment.....	56
Figure 3.12	The percentage of mycorrhizal colonization of switchgrass with *G. intraradices* from the second experiment.....	56
Figures 3.13 -3.16	Dry mass, root/shoot ratio, height, and number of tillers from experiment ! switchgrass plants.....	64
Figures 3.17 -3.20	Dry mass, root/shoot ratio, height, and number of tillers from experiment 2 switchgrass plants	68
Figure 3.21	Acid phosphatase levels for experiment 1 switchgrass plants	71
Figure 3.22	Acid phosphatase levels for experiment 2 switchgrass plants	71
Figures 3.23 -3.27	Root length, average root diameter, number of root tips, % small roots, % of large roots from experiment 1 plants...	82
Figures 3.28 -3.32	Root length, average root diameter, number of root tips, % small roots, % of large roots from experiment 2 plants...	86
Figures 3.33 - 3.36	Carbon and nitrogen content of shoots and roots from experiment 1 switchgrass plants	93
Figures 3.37 - 3.40	Carbon and nitrogen content of shoots and roots from experiment 2 switchgrass plants	95
Figures 3.41 - 3.44	Phosphorus and calcium contents of shoots and roots from experiment 2 switchgrass plants	98
Figures 3.45 - 3.38	Potassium and magnesium contents of shoots and roots from experiment 2 switchgrass plants	100
Figure 3.49	Cluster analysis of mycorrhizal treatment effect size for the first experiment using ward joining.....	103
Figure 3.50	Cluster analysis of mycorrhizal treatment effect size for the second experiment using ward joining.....	104
Figure 3.51	Canonical scores of cluster groups as determined by canonical discriminant analysis in experiment 1	108
Figure 3.52	Canonical scores of cluster groups as determined by canonical discriminant analysis in experiment 2.....	109

List of Tables

Introduction

Table1.1	Mycorrhizal dependency of wild varieties vs. cultivars.....	21

Materials and Methods

Table 2.1	Origin of switchgrass varieties.....	34

Results

Table 3.1	Germination percentage of switchgrass seeds.....	50
Table 3.2	Chromosome numbers in switchgrass.....	51
Table 3.3	Chlorophyll concentration of switchgrass varieties.....	54
Table 3.4	Physiological data for experiment 1 switchgrass plants....	62
Table 3.4.1	Results of two-way ANOVA for physiological parameters	65
Table 3.5	Physiological data for experiment 2 switchgrass plants.....	66
Table 3.6	Acid phosphatase levels for experiment 1 switchgrass plant	70
Table 3.7	Acid phosphatase levels for experiment switchgrass 2 plants	72
Table 3.8	Root anatomy data for experiment 1 plants.....	80
Table 3.8.1	Results of two-way ANOVA for root parameters	83
Table 3.9	Root anatomy data for experiment 2 plants.....	84
Table 3.10	Total % of C and N for experiment 1 plants.....	92
Table 3.11	Mineral analysis of C and N for experiment 2 plants.....	94
Table 3.11.1	Results of two-way ANOVA for C &N analysis.....	96
Table 3.11.2	Results of two-way ANOVA for P, Ca, K, Mg analysis ...	96
Table 3.12	Mineral analysis of P and Ca for experiment 2 plants.....	97
Table 3.13	Mineral analysis of K and Mg for experiment 2 plants.....	99
Table 3.14	Mean effect size values for each cluster in experiment 1....	107
Table 3.15	Statistics from CDA of experiment 1 clusters.....	107
Table 3.16	Mean effect size values for each cluster in experiment 2....	107
Table 3.17	Statistics from CDA of experiment 2 clusters.....	107

List of Abbreviations

AAC	Agriculture and Agrifood Canada, ECORC, Ottawa, ON
AM	Arbuscular Mycorrhizae (Endomycorrhizae)
ANOVA	Analysis of Variance
CDA	Canonical Discriminant Analysis
EM	Ectomycorrhizae
FM	Fresh Mass
IVDMD	*In Vitro* Dry Matter Digestibility
M+	Treatment with mycorrhizae
M-	Treatment without mycorrhizae
MD	Mycorrhizal Dependency
NU	Northern Upland
OD	Optical Density
ONP	Ontario Native Plants
PH	Prairie Habitat
PNPP	p-Nitrophenyl Phosphate
PUE	Phosphorus Use Efficiency
R/S	Root to Shoot Ratio
USDA	United States Department of Agriculture

Introduction

The agricultural sciences attempt to improve plants in order to make them more useful. Often, humans are unaware of the important processes that occur below ground and of the essential associations crop plants have with microorganisms in the natural environment. A preferable method of improving plants would be to increase plants' natural capacities to thrive in their environment, instead of altering the environment with fertilizers and tillage. Mycorrhizal associations, naturally occurring symbioses, have long been proposed as a feasible means to improve crops in agricultural systems. Unfortunately, current agricultural practices tend to discourage maximal benefits from mycorrhizal associations. It is important to assess the effect that cultivation has had on the ability of plant species to form mycorrhizal associations.

1.0 - Mycorrhizal Symbiosis

1.1– Symbiosis

Symbiosis is defined as 'the long term association of organisms of two different species'. A. De Bary, a German mycologist, was one of the first to investigate the phenomenon of symbiosis, which he defined in 1879 as 'the living together of dissimilar or differently named organisms'. These relationships span a continuum from antagonistic to mutualistic, facultative to obligate associations. Commonly, the term symbiosis is meant to imply a mutualistic relationship, but close associations can move along the symbiotic continuum depending upon the environmental conditions. Symbiotic partnerships can occur between members of all the kingdoms.

Some particularly interesting and important symbioses have occurred between members of the plant and fungal kingdoms. Because of the heterotrophic lifestyle of fungi, symbiosis as a means to acquire carbon has become a common strategy. It is estimated that a third of all fungi are involved in mutualistic symbioses (Kendrick, 1991).

1.2– Mycorrhizal Fungi

Mycorrhiza, meaning 'fungus root', is the term for associations that symbiotic fungi form with the root systems of terrestrial plants. Over 80% of plant species examined worldwide have associations with mycorrhizal fungi (Malloch et al., 1980). It has been suggested that this symbiosis has an ancient origin and that mycorrhizae have been instrumental in the colonization of terrestrial environments by plants (Pirozynski, 1980; Atsatt, 1988; Simon et al., 1993).

There exist two common types of mycorrhizae, Ectomycorrhizae and Endomycorrhizae. The ectomycorrhizae (EM) are associated with certain groups of trees and shrubs where they form a sheath around the roots. The fungal symbionts of EM, mostly basidiomycetes or ascomycetes, are very numerous, diverse and often host specific. Some EM also have the ability to live saprotrophically (Deacon, 1997), therefore are not obligate symbionts.

Endomycorrhizae are more common. They are associated with herbaceous vascular plants, some trees and non-vascular plants, the hyphae penetrate the root cortex. Some endomycorrhizal fungi are named Arbuscular Mycorrhizae (AM), because of the structures they form when in association with a host plant. Fungi of this group are

zygomycetes, order Glomales or Endogonales. There are fewer fungal species involved in endomycorrhizal associations, ~150 species are known in 7 genera, with many more hosts. The AM fungi are not host specific, although some symbiotic partnerships may be more effective than others (O'Bannon et al., 1980; Ollivier et al., 1982; Anderson et al., 1994; Graham and Abbott, 2000; Smith et al., 2000). All AM symbioses involve obligate biotrophism. Specialized types of AM, the vesicular arbuscular mycorrhizae (VAM) are restricted to the order Glomales and form distinctive vesicles inside the host root.

1.2.1– Vesicular Arbuscular Mycorrhiza Anatomy

Endomycorrhizae form structures to connect with the host plant which are primarily inside the cortex of the root. Characteristic structures of VAM are vesicles, arbuscules, large intercellular hyphae and the extraracinary hyphal network (Fig. 1.1).

The arbuscules are fungal structures that penetrate the cell wall. These finely branched structures are contained inside the cell within a periarbuscular membrane, and so never make direct contact with the plant cytoplasm (Smith and Read, 1997). The repeated branching pattern of the arbuscules is what gives them their name. They are an important site of nutrient transfer between the two partners. The large surface area created by branching makes nutrient transfer efficient. Evidence for this was presented by Rosewarne et al. (1999), who showed that in mycorrhizal plants, phosphorus transporters are preferentially located in cortical cells that contain arbuscules. Arbuscules have a short lifespan, 7-14 days. They are continually being formed in new cells and disintegrated by autolysis or digestion by host cells during the colonization (Deacon, 1997).

The other fungal structures that are found internally in the plant are the intercellular hyphae and vesicles. The more common type of AM fungi have an anatomy named 'Arum' type where the hyphae grow in the apoplastic areas. The hyphae use the space between cell walls of adjacent cortical cells to grow and expand their colonization. Some AM fungi expand by growing through the cells whey they form coils inside the cell wall; these are named 'Paris' types. The hyphae do not generally grow very far along the root and have sparse branching. The diameter of hyphae is approximately 3-4 μm (Clark and Zeto, 2000). Internal hyphae preferentially grow near the vascular bundle where phloem exudates are plentiful.

Vesicles are large ovate structures that lie intercellularly, but can equal or exceed the cell in size. These hyphal swellings are usually terminal on a hyphal branch. Vesicles are generally thought to be the energy storing structures of the fungi. They have a high lipid content, many nuclei and thick walls. It has been suggested, but not conclusively proven, that they may be able to colonize new roots (Smith and Read, 1997). A large number of vesicles would indicate a healthy plant and an efficient symbiosis.

Outside the host plant, the hyphae form a large network in the soil. This extraradical hyphal network performs various functions in the soil. The hyphae have differing diameters (2-27 μm) and wall thicknesses. They can form bridges between plants, which could be of ecological importance (Francis and Read, 1984). Hyphal networks are necessary for the survival of the mycorrhizae and are a primary source for root colonization. The small diameter of the hyphae allows access to a soil volume not available to roots alone. The hyphae can acquire minerals and water in the soil more efficiently than do roots. It has also been suggested that hyphae are able to access different sources of minerals in the soil since the biochemistry of the fungi is different from that of the plant (Tarafdar and Marschner, 1994).

Large spores are formed outside the root in the soil. Spores are lipid rich, multinucleate, and can support a germinating hypha for a week while it searches for a live root to colonize. Spores are characteristic of the species of mycorrhizal fungus. All are very large in size in comparison to hyphae or vesicles, and have thick walls, which protect against adverse conditions such as drought and cold (Smith and Read, 1997).

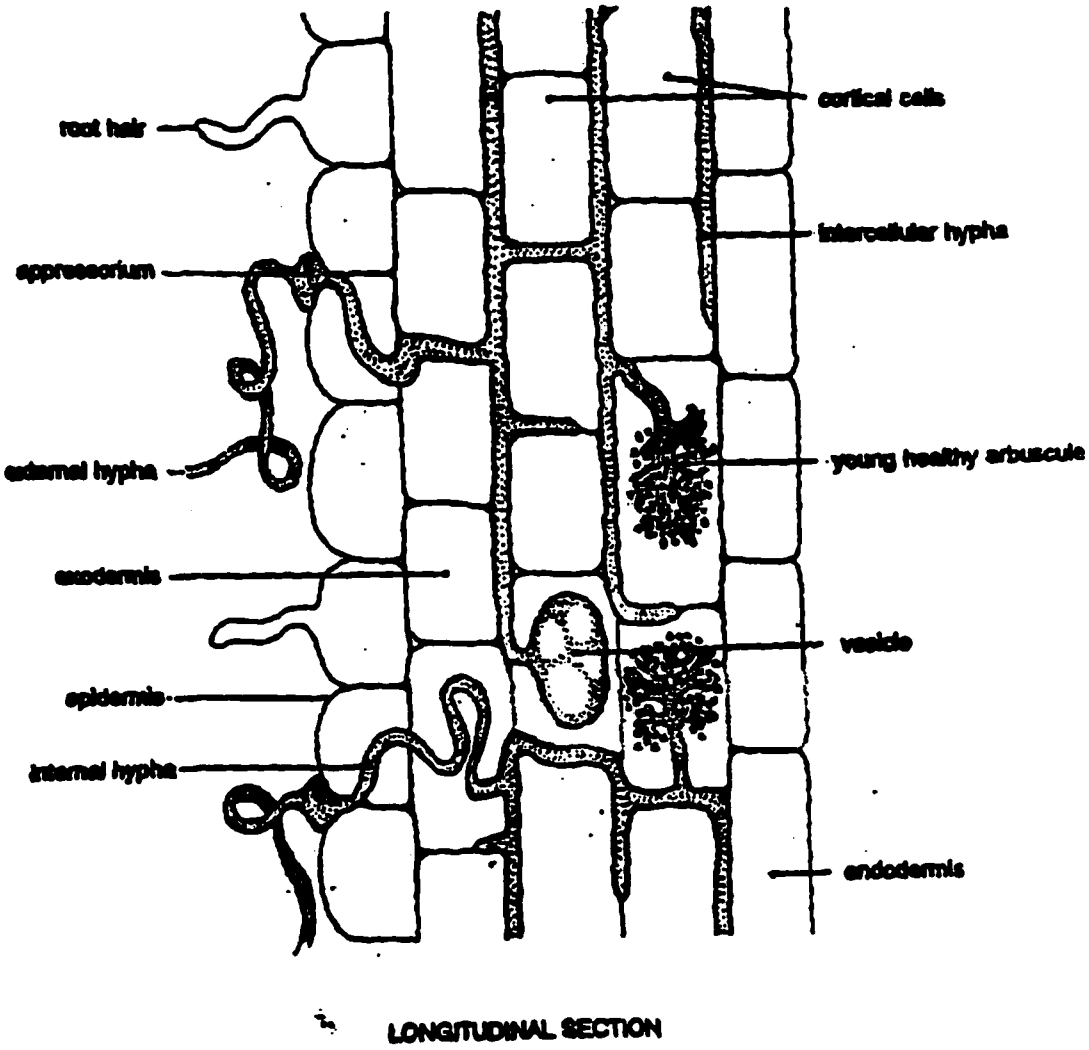


Figure 1.1. Anatomy of VAM, from Brundett et al., 1994.

1.2.2– Mode of Colonization

There are several ways for mycorrhizae to colonize roots. Spores in the soil may germinate and find their way to a live root starting a completely new colonization. Most commonly, hyphae in the existing soil network that come in contact with a root start new colonization sites. Mycorrhizal structures from previously colonized roots may colonize new roots in secondary colonization events.

It is not uncommon for a root to host several different species of mycorrhizal fungi simultaneously. Plants are able to 'control' the colonization level the roots will support and exclude new invading mycorrhizal fungi after a threshold level of colonization is reached (Vierheilig et al., 2000). Functional complementarity suggests that mycorrhizal fungal species having a structure complementary to that of the root are preferred (Koide, 2000). For example, plants with short roots prefer AM fungi which have long hyphae and can reach far into the soil while tap roots would favour finely branched fungal species.

Germinated spores or sections of the extraradical hyphal network can find their way to a root using chemical signals such as flavonoids, sugars or acids (Azcon and Ocampo, 1981; Ocampo and Azcon, 1985; Anderson, 1988; Smith and Read, 1997). Once a live hypha comes in contact with a live root, it initiates entry in the root by forming a hyphal swelling called an appressorium, which attaches to the root surface and penetrates the epidermis using hydrolytic enzymes. The fungus forms a penetration peg that forces its way into the cortex using pressure (Smith and Read, 1997).

1.2.3 –Benefits of Mycorrhizal Symbiosis to Mineral Nutrition and Stress

Growth benefits observed in mycorrhizal plants have been attributed, in a large part, to their greater mineral nutrition (Smith and Read, 1997). The most prominently researched of the elements enhanced by mycorrhizae is phosphorus (P). Many other macro and micronutrients (N, K, Ca, Mg, S, Zn, and Cu) in the soil have been shown to be absorbed at a higher rate with mycorrhizae. Mycorrhizae enhance the absorption of ions that are immobile in the soil to a greater extent than mobile ions.

A macronutrient often limiting in natural soil, phosphorus is a relatively immobile ion. Phosphates are more easily absorbed by the hyphal network, which mines the soil volume more efficiently than do roots alone because of its small diameter. Greater P content has repeatedly been found in mycorrhizal plants under both greenhouse and field conditions (Dickson et al., 1999; Schweiger and Jakobsen, 1999). Phosphates from both inorganic and organic sources were shown to be absorbed directly by hyphae in axenic cultures (Joner et al., 2000). The genes for inorganic phosphorus transport have been putatively discovered, and it seems that fungi and roots have separate P uptake transporters (Rosewarne et al., 1999).

Increased P uptake in mycorrhizal plants may be due to phosphatase enzymes. The production of acid phosphatase, which hydrolyzes the phosphate ester bond in organic sources to create more easily absorbed inorganic phosphates, may provide an additional source of P available to mycorrhizal roots (Joner et al., 2000). Roots naturally produce their own acid and alkaline phosphatases on their epidermal surface; this is supplemented

with the fungal production of phosphatases. The amount of phosphatase increases with root density and with mycorrhizae (Tarafdar and Marschner, 1994). A high proportion of the acid phosphatase produced comes from portions of the root colonized by AM fungi (Grierson and Comerford, 2000). A higher phosphatase activity has been associated with greater P efficiency in the plant (McLachlan, 1980).

Other macronutrients, (N, K, Ca and Mg) are also needed in large quantities by plants and can be limiting in the soil. Mycorrhizae have been implicated in improved absorption of N in all soil types, and of K, Mg and Ca in acid soils (Clark and Zeto, 2000). Levels of micronutrients also increase during mycorrhizal colonization. The most important uptake occurs with copper (Cu) and zinc (Zn), but iron (Fe) and sulphur (S) are also increased (Clark and Zeto, 2000). Increased absorption of Cu, Zn, Mn and Fe were shown to be dependent on soil mineral levels and P nutrition; a mycorrhizal advantage was seen at low P levels (Liu et al., 2000). Some metals, such as aluminium (Al), are taken up in lower quantities with AM, thus preventing mineral toxicity (Clark and Zeto, 2000). Heavy metals such as caesium (Cs) and strontium (Sr) can accumulate in higher concentrations in mycorrhizal plants making them useful in phytoremediation (Entry et al., 1999). Some plant/mycorrhizal fungi species are able to thrive in polluted sites; therefore mycorrhizae have been suggested as a means to aid clean up of heavy metal contamination at former mine sites (Pleger et al., 1994; Noyd et al., 1995).

Increased resistance to environmental stress factors has often been reported in mycorrhizal plants; this may contribute to observed growth or fitness gains. Resistance to

abiotic stress factors such as drought and cold, as well as biotic stress from pathogens and herbivores has been observed in mycorrhizal plants. A greater tolerance to stress has led to mycorrhizae being used in revegetation of marginal sites (Reeves et al., 1979; Noyd et al., 1995; Smith et al., 1998). Mycorrhizae can aid with water acquisition through their hyphal network. Colonized plants were able to maintain higher water potentials longer under drought conditions in greenhouse and field trials (Subramanian and Charest, 1995 & 1998; Subramanian et al., 1996; Gemma et al., 1997). Improved condition during cold acclimation was seen in mycorrhizal wheat (Paradis et al., 1995). Many studies have shown AM fungi have a deleterious effect upon pathogenic soil microorganisms (St-Arnaud et al., 1994; Filion et al., 1999; Ravnskov et al., 1999). An increase in the secondary metabolites of mycorrhizal roots has been observed (Maier et al., 1997; Peipp et al., 1997), suggesting this may be the source of biotic protection.

1.3 - Plant Response to Mycorrhizal Colonization

Environmental factors as well as genetic ones mediate efficient symbiosis, where the plant and fungus both gain the maximum increase in fitness. Soil mineral content, pH, temperature, as well as plant species, all affect the ability of mycorrhizal fungi to colonize and thrive. AM fungi are very sensitive to soil P content. In high P environments, mycorrhizae have reduced colonization rates and have even been found to be detrimental to the host plant (Manjunath and Habte, 1992; Dekkers and van der Werff, 2001).

Although AM fungi are not host specific, certain host plant- fungus combinations are functionally better than others. Symbiotic compatibility can be demonstrated by testing

many species or strains of AM fungi with one species of plant. This method has repeatedly shown that some species pairs are more advantageous than others to either the plant or the fungus (O'Bannon et al., 1980; Ollivier et al., 1982; Anderson et al., 1994; Graham and Abbott, 2000; Smith et al., 2000; Zhu et al., 2000). Some plant species are non-mycotrophic and do not allow colonization (Newman and Reddell, 1987; Tester et al., 1987). Myc- mutants, that are not able to support mycorrhizal colonization, invoke the salicylic acid defence pathway when colonized by AM fungi (Duc et al., 1989; Blilou et al., 1999).

1.3.1 – Carbon Use by Mycorrhizae

As obligate biotrophs, arbuscular mycorrhizae depend upon their host plant for all of their carbon (C) energy. It has been estimated that the fungi consume 10% of the photosynthate produced by the plant (Rygiewicz and Andersen, 1994; Tinker et al., 1994). Colonization may alter the physiology of the plant so that it diverts more of its C to the roots in order to feed the fungus, which is acting as a C sink (Rygiewicz and Andersen, 1994; Graham et al., 1997). Mycorrhizal roots receive 4-20% more photosynthate than non-mycorrhizal roots (Smith and Read, 1997). The type of carbohydrates transferred is altered so that reducing sugars such as glucose are available in the roots (Graham et al., 1997). The fungus uses glucose as its primary energy source (Solaiman and Saito, 1997), but other carbohydrates such as trehalose are also transferred to the mycelium (Bago et al., 2000). A higher rate of below ground respiration has been observed in mycorrhizal roots (Nielsen et al., 1998). However, no differences in total root C levels were noted. Carbon use efficiency has been calculated to as high as 96% in

mycorrhizal plants, due to an increase in the photosynthetic rate (Tinker et al., 1994). The amount of sugar in shoots has been found to increase in mycorrhizal maize plants (Boucher et al., 1999). The amount of C in roots is associated both with mycorrhizal dependency of the plants (Graham et al., 1997) and the level of root colonization (Schwab et al., 1991).

1.3.2 - Cost/Benefit of Mycorrhizal Symbiosis

When assessing the cost/benefit relationships of a mycorrhizal symbiosis, cost to the plant is commonly measured in carbon loss and benefit in terms of mineral acquisition. The plant also benefits by increased stress tolerance and competitive advantages in some ecosystems. These have not been included in traditional cost/benefit calculations because of the difficulty in building equations that encompass complex biotic relationships. Relative costs and benefits to the AM fungi cannot be assessed because they are obligate biotrophs and do not grow if not involved in a symbiosis.

The amount of C allocated to the fungus depends both on the host plant species or variety and the AM fungal symbiont. Whether the carbon lost represents a significant cost to the plant depends on the photosynthate available, increases in photosynthesis that may arise from increased mineral availability, and increased stress resistance conferred to the plant by the symbiosis. Monz et al. (1994) found increased atmospheric CO_2 concentration increases AM colonization in C4 grasses. This supports the hypothesis that enhanced C availability increases the amount of fungus supported by the plant. Plants are able to some extent to regulate the carbon costs that they expend on mycorrhizae (Vierheilig et

al., 2000). A model of carbon use efficiency predicts that mycorrhizal plants should have higher carbon efficiency than non-mycorrhizal plants (Tinker et al., 1994). This would suggest that the benefits of mycorrhization outweigh the costs. Douds et al. (1988) showed that there might be an optimal level of mycorrhizal colonization, since plants continue to transport photosynthate to mycorrhizal roots at a higher rate even when transport of P is not increased by the colonization. A model of cost/benefit (Fitter, 1991) proposed that the symbiosis is beneficial to plants only up to a threshold level of colonization where C costs equal P uptake by mycorrhizae. A more encompassing model of plant benefit based on mineral acquisition and % colonization was proposed by Gange (1999) (Fig 1.2). This model allows for negative responses to colonization at high soil mineral levels and predicts an optimal % colonization. This optimum depends on the environment and the genetics of the symbiotic pair. A more recent model of cost/benefit in mycorrhizae (Tuomi et al., 2001) suggests that the symbiosis is inherently inefficient in terms of C cost and is detrimental in the case of high soil mineral levels, low photosynthetic efficiency, and stress conditions.

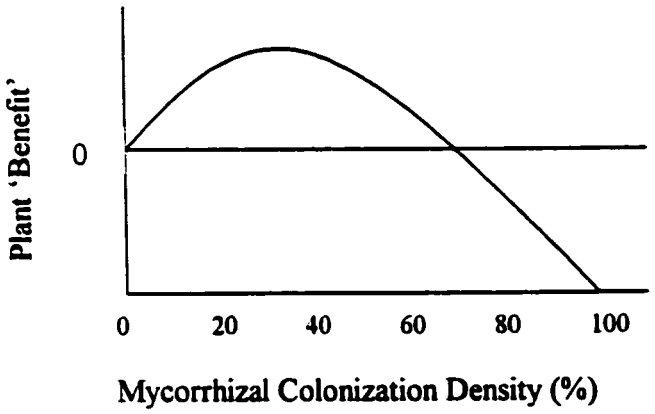


Figure 1.2. Model of plant 'benefit' and % colonization by Gange (1999).

1.3.3 – Symbiotic Continuum of Mycorrhizae

Early studies on mycorrhizae treated the symbionts exclusively as beneficial mutualists. We know that this is not always the case; host plant response to colonization can range from beneficial to detrimental. A symbiotic continuum for mycorrhizae that includes commensalism and antagonism (Fig 1.3) is included in recent studies of cost/benefit (Francis and Read, 1995; Johnson et al., 1997). These studies propose that the antagonistic condition is induced by high soil mineral abundance, carbon limitation, genotypic pairs that are not well suited, and non-native soil type.

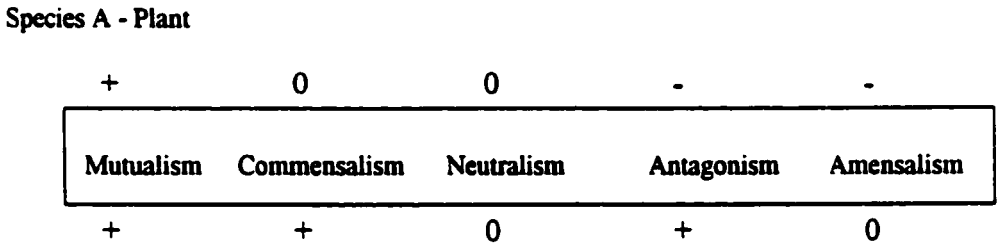


Figure 1.3. Symbiotic continuum for mycorrhizae modified from Francis and Read, (1995). +, beneficial effect; 0, no effect; - detrimental effect.

1.3.4 – Mycorrhizal Dependency

Mycorrhizal dependency (MD) refers to the amount a species or variety benefits from the symbiosis. Since 'dependency' implies a necessity and a positive response from the plant, researchers have recently proposed using the term 'mycorrhizal responsiveness' (Smith, 2000). Dependency of a species upon mycorrhizal symbiosis has traditionally been measured by the biomass increase in the host plant. The equation most often used to calculate mycorrhizal dependency is:

$$MD = [(M- NM) / M] \times 100\%$$

Where M= mycorrhizal dry mass, NM= non-mycorrhizal dry mass

Plant biomass is most commonly used to measure MD because plant mass is a good indicator of fitness, and also a parameter of much value to cultivators. It has become clear in the past three decades that plant response to mycorrhizae goes far beyond simple growth increases. Mycorrhizal dependency should take into account the directional response of the plant in factors such as root morphology, production of enzymes and hormones, increases in mineral levels and stress resistance. These factors have been shown to be influenced by mycorrhizae and could impact the plant's fitness. Interestingly, MD does not seem to be related to the % colonization of the roots (Manjunath and Habte, 1991), which was an assumption made by many researchers.

Root morphology has been used as an indicator of the MD of plants since Baylis (1970) proposed that plants with large root diameters, few root hairs and little branching would be more dependent upon mycorrhizae. If plants rely on mycorrhizae to exploit the soil area with their fine hyphae, the plant would be wasteful in producing many fine roots. Plants confined to small soil volumes, which can be adequately exploited by the root system alone, are less MD than plants in large soil volumes (Baath and Hayman, 1984). Further studies have shown that plants with large root mass, long root hairs, and small root diameters tend to be less MD (Manjunath and Habte, 1991) while increased root branching decreases MD (Hetrick et al., 1992b). Field studies of tallgrass prairie species

have shown that root architecture is a good predictor of MD, even in Gramineae which are known to have adventitious root systems (Hetrick et al., 1991; Hetrick, 1991b).

The link between root morphology and P efficiency has been found in non-mycorrhizal plants. Gahoonia et al. (1999) suggest that increased root hairs and more efficient P uptake are alternative mechanisms to dependence upon mycorrhizae. Plants that are less dependent upon mycorrhizae have higher root P absorption rates (Manjunath and Habte, 1991b & 1992; Schweiger et al., 1995). Phosphorus efficiency is not always correlated with increased growth, normally associated with MD (Smith et al., 2000). Azcon and Ocampo (1981) did not find any relationship between MD and leaf mineral content (P, N, K, Ca, Mg) as might be expected if mineral concentration and MD were directly related. Although mineral increases are seemingly a good indicator of MD, this has not yet been reliably proven. Overall, an increase in available P in the soil tends to reduce MD values (Anderson et al., 1994). Enhanced production of acid phosphatase by mycorrhizal plants may also serve as an indicator of MD (Tarafdar and Marschner, 1994).

Mycorrhizal dependency has been shown to vary between plant species and varieties (Bentivenga and Hetrick, 1992; Hetrick et al., 1992; Anderson et al., 1994). Cultivar differences have been shown in many plant species: citrus (Menge et al., 1978), wheat (Bertheau et al., 1980; Azcon and Ocampo, 1981), maize (Toth et al., 1984; Kaeppler et al., 2000), soybean (Heckman and Angle, 1987), pea (Estaun et al., 1987), cowpea (Ollivier et al., 1983), alfalfa (Lackie et al., 1988), pearl millet (Krishna et al., 1985), rye and barley (Baon et al., 1992). Many of these species have been cultivated for an

extended period of time, giving time for genetic divergence. Ecotype specificity has also been shown, MD being highest in the native soils of a particular plant (Anderson et al., 1994). Soil type specificity was shown in citrus cultivars whose MD changed depending upon the fertilization regime used (Menge et al., 1978).

Differences in mycorrhizal response among cultivars have been attributed to many factors. Mineral responsiveness has been assessed as: response to fertilizer under non-mycorrhizal conditions (Menge et al., 1978; Kaeppler et al., 2000), P uptake (Baon et al., 1992), P efficiency (Baon et al., 1993), P leaf content (Toth et al., 1984), and phosphatase synthesis (Ollivier et al., 1983). Other research examined cultivar differences in root parameters such as root length (Krishna et al., 1985), root diameter (Khalil et al., 1999), root content of reducing sugars (Ocampo and Azcon, 1985), and root starch content (Graham et al., 1997). Cultivar responsiveness may also be explained by inadvertent selection against mycorrhizae in some strains. Host plant resistance to pathogenic fungi decreased the ability of some corn or barley cultivars to be colonized by mycorrhizal fungi (Toth et al., 1990; Ruiz-Lozano et al., 1999). It is probable that all of these factors, since they are related to MD, contribute to mycorrhizal effectiveness.

The variation seen among cultivars was found to be heritable, suggesting genetic control of MD (Krishna et al., 1985; Lackie et al.,1988; Manske, 1989; Mercy et al., 1990). Recently, a quantitative trait locus that controls mycorrhizal responsiveness was found in maize (Kaeppler et al., 2000). Genetic control of MD makes it interesting from a crop breeding point of view. The concept of selection for MD has been frequently proposed

(Smith et al., 1992; Smith and Read, 1997; Manske et al., 2000). The benefit of selection for MD would be to increase plant yield and health with minimum input to the system.

2.0- Agriculture and Mycorrhizae

As the benefits of mycorrhizae became elucidated, researchers were interested in their practical applications. Use of mycorrhizae in agricultural systems was promoted as an alternative to commercial fertilization. Commercial application of AM fungi was investigated, the early focus being on finding strains of fungi that produced the greatest benefit in crop plants (Menge, 1983). Jeffries called for 'routine commercial propagative practices' in his 1987 review of mycorrhizae in agriculture. Since AM fungi are ubiquitous in soil, it was necessary to find the right type of soil in which inoculation would be useful. Tropical soils, which are often poor in mineral content, and soils damaged by mining or other contamination were cited as places where large scale inoculation might be beneficial. The new emphasis for mycorrhizae in agriculture is on selectively breeding mycorrhizal efficient cultivars and the introduction of agricultural practices, such as no-till, low-input, and crop rotation, that maintain the diversity of natural fungal populations (Hamel, 1995; Dodd, 2000).

In agricultural systems, responses to mycorrhizae have been inconsistent. Some crops show great improvement with mycorrhizae in the greenhouse (Kaeppler et al., 2000; Lambert et al., 1980) and in sterilized soil in the field (Azcon and Ocampo, 1981; Menge et al., 1978). However, in crop fields, mycorrhizal fungi can sometimes act as pathogens (Hendrix et al., 1992).

2.1- Effects of Agriculture on Mycorrhizae

Modern agricultural practices are now known to potentially have serious effects on the natural populations of AM fungi in the soil. Disruption of the extraracinary hyphal network, loss of colonization potential in fertilized soils, and host plant compatibility with fungal species all limit mycorrhizae in traditional agricultural environments. It has been shown that agricultural practices have made AM parasitic in citrus cultivation (Graham, 2000).

Tillage of the soil has been found to have negative effects on AM fungi and soil properties in general. Breaking the soil during tilling destroys the extraracinary hyphal network that is essential for colonizing new roots. Undisturbed soil maintains high mineralizeable N and acid phosphatase activity, while colonization potential, hyphal length, number of AM spores and AM species richness are reduced in soil where the hyphal network has been broken (Jasper et al., 1989; Kabir et al., 1998; Boddington and Dodd, 2000; Drijber et al., 2000). Plants in a tilled soil have been found to have lower AM colonization and lower P uptake (McGonigle and Miller, 2000; Galvez et al., 2001). No-till management has been suggested as the best method to preserve soil resources and the benefits of no-till have been linked to the presence of AM fungi (McGonigle and Miller, 2000; Mozafar et al., 2000).

Fertilized soils have less AM colonization and extraradical hyphae (Smith and Read, 1997; Liu et al., 2000). These negative effects can remain long after the applications have stopped (Dekkers and van der Werff, 2001). Nutrients supplied in organic form do not

have a negative effect on mycorrhizae and can maintain the same level of internal mineral concentration as in conventionally fertilized plants (Mader et al., 2000; Franke-Snyder et al., 2001). Organic farming, without pesticides, fungicides particularly, increases the number of AM fungi (Scott et al., 1994; M.D Smith et al., 2000). As discussed earlier, mycorrhizae can reduce fungicide and pesticide use in agriculture because of their ability to deter pathogens.

The crop species may affect the mycorrhizal populations because of host preference. AM species composition differs depending on crop species and between agricultural and natural sites (Schenck and Kinloch, 1980; Guo et al., 1993; Talukdar and Germida, 1993; Hendrix et al., 1995; Boddington and Dodd, 2000). Plant species diversity was shown to be dependent on AM fungal species diversity (van der Heijden et al., 1998a & b). Evidence suggests that fungal species diversity is also dependent on plant species despite the fact that AM fungi are considered generalists (Dhillion, 1992; Bever et al., 1996).

2.2- Cultivars and Mycorrhizae

As discussed earlier, mycorrhizal responsiveness is dependent upon genetics. In an agricultural environment, this means that some cultivars should benefit more from mycorrhizae than others. Crop breeders could benefit from the exploitation of genetic differences.

To our knowledge, no cultivar has yet been developed to maximize mycorrhizal benefit. Under intense agricultural practices, plants with high MD may have a lower yield

because the energy costs to the fungus are not balanced by increased mineral nutrition or stress tolerance. This situation would create an inadvertent selection against mycorrhizal variants. Inadvertent selection against varieties easily colonized by AM fungi has already been shown in maize by Toth et al. (1990), and this may be happening in other intensely cultivated species. It is possible that over the long term, varieties that were most responsive to mycorrhizae were discarded under high input systems and the remaining gene stock is not suitable for MD selection.

2.3- Selection in Agricultural Environments; Wild Varieties vs. Cultivars

One method of assessing the impact agriculture has had on the ability of crops to respond to mycorrhizae is to compare modern cultivars with wild varieties of the same species. Selection pressure on plants in natural environments are quite different from those imposed by human crop breeding processes, therefore wild plants should be different from their agricultural progeny. Thus far, determinations of MD in wild and cultivated variants have been done for only a few plant species (Table 1.1).

Table 1.1. Mycorrhizal dependency of wild varieties vs. cultivars for some plant species

Species	MD	References
Wheat *Triticum aestivum* L.	Wild>Cultivated	Kapulnik and Kushnir, 1991; Hetrick et al., 1992,1993
Soybean *Glycine max* L.	Wild>Cultivated	Khalil et al., 1994
Corn *Zea mays* L.	Wild>Cultivated	Khalil et al., 1994
Oats *Avena spp.*	Wild<Cultivated	Koide et al, 1988; Haynes et al., 1991
Tomato *Lycopersicon esculentum* Mill.	Wild<Cultivated	Bryla and Koide, 1990a & b

In wheat, which has been cultivated for millennia, Hetrick et al. (1992) found that wild landraces were extremely mycorrhizal dependent (55-169%) while cultivars were dependent to a lesser extent (29-100%). A genetic relationship was not found in this study as in a previous one (Kapulnik and Kushnir, 1991) where certain genomes in the ancestors of wheat were related to MD. While the first studies depended solely upon biomass to asses MD, the same cultivars were found to have different P uptake responses when mycorrhizal which were not necessarily correlated with biomass increases (Hetrick et al., 1996). So far, the evidence suggests that wheat has decreased its MD during its cultivation (Hetrick et al., 1993).

Assessments of MD, which included parameters affected by mycorrhizae such as root architecture and acid phosphatase production, were undertaken for corn and soybean (Khalil et al., 1994). This study concluded that both corn and soybean cultivars are less MD than their wild kin. Decreasing MD within an agricultural environment was not seen for all species. In oat (Koide et al., 1988; Haynes et al., 1991) and tomato (Bryla and Koide 1990a & b), cultivars were found to be more MD than the wild plants. Whether a species increases or decreases MD in an agricultural setting may depend on growth practices. Crops that are less managed, such as oats, may be at an advantage when mycorrhizal, while highly fertilized crops, such as maize, are not. According to the model of carbon efficiency of Tuomi et al. (2001), wild plants should be more MD since selective advantage disappears in commercial agriculture.

3.0- Switchgrass

Switchgrass, *Panicum virgatum* L. (Graminae; Panicoidae) (Fig. 1.4), has not been cultivated for a very long time, but its popularity is increasing, as it has many uses. Switchgrass was first used as fodder for cows and sheep and now is being developed as a potential biofuel. As a dominant grass of the native tallgrass prairie habitat, it is in demand for prairie restoration efforts. Other uses include planting for erosion control and as an ornamental plant.



Figure 1.4. Diagram of switchgrass including panicle and rhizomes.

Panicoid grasses, which are of tropical origin, use the C4 photosynthetic pathway, thus are classified as 'warm season' grasses. A native North American species, switchgrass is normally associated with the tallgrass prairie habitat and can be found across North America in a range that spans from Manitoba to Texas and from Nova Scotia to Colorado (Hitchcock, 1950). In Canada, it is principally found in Manitoba and southeastern Saskatchewan, though it can be found in patches of southern Ontario and Québec. Switchgrass usually grows to 1-2 m in height, but the native Ontarian varieties have been noted to be considerably shorter (0.5-1m) (Dore and McNeill, 1980).

Switchgrass is a perennial plant that spreads vegetatively via tillers and scaly rhizomes (a unique characteristic among panicoid grasses). Switchgrass is photoperiod sensitive and requires short days to flower. In order to produce seeds, the plants must be outcrossed, they do not self-pollinate. Switchgrass has large feathery panicles that produce small, smooth, hard seeds with poor germination rates (Looman, 1983). Seeds are often dormant and require stratification and scarification to produce good germination rates in cultivation. The shiny seeds may have had aesthetic value for the native people of North America, who reportedly used them as beads and as a component of pemmican (Dore and McNeill, 1980).

Warm season grasses have been used as summer forage because they continue to produce good fodder after cool season grasses have peaked. Switchgrass is successful in many environments and popular as an agricultural species because of its general stress

tolerance. Able to grow in acid soils, ranging from pH 4-7, switchgrass is also drought tolerant and thus has the ability to grow on marginal land (Jung et al., 1988). In Canada, cold tolerance is an issue for this grass of tropical origin. Hope and McElroy (1990) found that switchgrass originating in the Great Plains overwintered successfully due to its tough rhizomes.

One of the most desirable qualities of switchgrass to cultivators is its ability to survive in infertile soil without the need for large fertilizer inputs. Switchgrass is very P efficient; P concentrations as low as 5 mg/kg support this grass while cool-season grasses could not thrive (Moser and Vogel, 1994). In field trials, switchgrass did not increase its biomass when fertilized with P (Jung et al., 1988; Sanderson et al., 1996). Phosphorus fertilizer, if applied, is maximal at 50kg/ha (Jung et al., 1988; Staley et al., 1991; Moser and Vogel, 1994; Samson et al., 1999). Nitrogen is considered the limiting nutrient for switchgrass, and N fertilizer is often applied to switchgrass fields. Rates of N application in the field tend to be 40 kg/ha (normal), 100 kg/ha (high) and up to 180 kg/ha during experimental trials.

Recently, interest in switchgrass has focused on its possible use as a biofuel. It is proposed that fuel ethanol could be produced from cellulose derivates by enzymatic conversion to sugars and subsequent fermentation (Hopkins et al., 1995). Biofuel researchers chose switchgrass for its ability to grow in diverse environments with little input, its contribution to soil conservation and revegetation of the prairies, and because of the abundance of cellulose in the stem (Sanderson et al., 1996; Madakadze et al., 1999).

A highly desirable characteristic for cultivators is the apparent tolerance of switchgrass seedlings to herbicides (Vogel, 1987), which can reduce weed competition. Resistance to toxins seems to extend to heavy metals as well. Switchgrass is able to grow normally on mine tailings that have elevated concentrations of lead and zinc without any signs of phytotoxicity (Levy et al., 1999). Some negative aspects include possible toxicity to lambs (Puoli et al., 1992), susceptibility to *Puccinia* spp and the *Panicum* mosaic virus (Moser and Vogel, 1994).

3.1- Genetics

Switchgrass has a base chromosome number of n=9. Varieties of native populations have been found that carry chromosome numbers from diploid (2n=18) to dodecaploid (2n=108) (Nielsen, 1944). In subsequent examinations of wild populations, most native accessions in the northeast prairies were found to be tetraploid (2n=36), although hexaploid (2n = 54), and octoploid (2n= 72) varieties were reported in central North America (McMillan and Weiler, 1959). Commercial cultivars are either 2n= 36, 54 or 72 (Riley and Vogel, 1982; Hultquist et al., 1996). Conflicting reports of chromosome numbers in the same variety of switchgrass (Hopkins et al., 1996) have appeared, making the issue of ploidy still relevant in this species. No morphological or physiological characteristics have been attributed solely to chromosome number.

Switchgrass has been separated into two groups based on morphological characteristics of native populations (Eberhart and Newell, 1959). The two types are now referred to as 'Upland' and 'Lowland' types and are differentiated by morphology and geography.

Upland types are found in more northern regions, tend to be fine stemmed, semi-decumbant, have a shorter growing season, and are broad based. Lowland types are southern, coarse stemmed, erect, have a longer growing season and bunched tillering. Evidence for the groups being genetically distinct as well as morphologically so, was found by polymorphisms of the chloroplast DNA (Hultquist et al., 1996).

3.2- Agricultural Selection

In order to conduct an effective selective breeding program, there must be heritable differences among populations. Variation among wild populations was established at the beginning of switchgrass cultivation (Nielsen, 1944). Estimates of the heritability of phenotypic traits were soon launched (Eberhart and Newell, 1959; McMillan and Weiler, 1959). As research continued, heritable varietal differences in switchgrass have been found for physiological parameters such as: crown node placement (Elbersen et al., 1999), leaf area index (Madakadze et al., 1998), tiller numbers and height difference (Madakadze et al., 1998b), acid soil tolerance (Hopkins and Taliaferro, 1997), response to environmental variation (Hopkins et al., 1995), % N (Talbert et al., 1983), and *in vitro* dry matter digestibility (IVDMD) (Vogel et al., 1981). Many of these traits have been selected for and incorporated into elite breeding populations. Several commercial cultivars of switchgrass have been released. Some are products of selective breeding programs (Trailblazer, Caddo, Shelter, Summer, NU) and some are native populations from which the hardiest individuals were selected (Forestburg, Cave-in-Rock).

Basis for genetic variation among cultivars being greater than variation within cultivars was recently found using molecular methods. Cultivars of switchgrass have a within population similarity of 81% while the similarity among the cultivars was 65.2% (Gunter et al., 1996). It is generally accepted among switchgrass breeders that there is high genetic variation in the species for potential crop improvements using traditional selective breeding methods (Godshalk et al., 1988).

3.3- Mycorrhizae and Switchgrass

As a C4 warm season grass, switchgrass forms mycorrhizal associations and is highly mycorrhizal dependent (Hetrick et al., 1988 and 1990). Hetrick et al. (1991) showed that C4 grasses had a much higher MD than C3 grasses in natural tallgrass prairie. Estimates based on dry biomass of wild populations from Kansas tallgrass prairie have found MD to be up to 98% (Wilson and Hartnett, 1998). Wild populations can be found with 12-50% mycorrhizal colonization (Hetrick and Bloom, 1983). Switchgrass usually shows positive growth responses to mycorrhizal inoculation, but neutral or negative responses were also found, depending on environmental conditions (Bredja et al., 1993; Entry et al., 1999).

Mineral concentrations in switchgrass were shown to increase with mycorrhizae. The minerals N and P, generally limiting in tallgrass prairie soils, showed the greatest increase in mycorrhizal switchgrass plants (Bredja et al., 1998; Johnson, 1998).

In the tallgrass prairie, mycorrhizal fungi are important determinants of the plant community composition. Competitive relationships between plants are mediated by mycorrhizal colonization. Dominant prairie grasses such as switchgrass are reduced in % cover when AM fungi are excluded (Wilson and Hartnett, 1997). Plant-fungal species preference has been noted in switchgrass. Different species of AM fungi were found to produce varying colonization levels and growth responses in switchgrass (Clark et al., 1999b). In the prairie, it was found that mycorrhizal diversity is lowest in switchgrass stands when compared with five other prairie grasses (Eom et al., 2000).

4.0- Hypothesis and Objectives

Hypothesis- Wild populations of switchgrass have a higher mycorrhizal dependency than cultivars.

This hypothesis is based on cost-benefit models and the observation that modern agricultural practices are detrimental to mycorrhizal fungi. If populations of switchgrass are in a high-input management system during selective breeding, mycorrhizae will tend towards the detrimental end of the symbiotic continuum. Therefore, plants which are less susceptible to mycorrhizal colonization will be selected and, as a result, cultivars will have a lower MD than wild plants. If conditions are more natural during the selection process, then mycorrhizae will tend towards the beneficial end of the symbiotic continuum, the largest plants will be the mycorrhizal ones, and cultivars will have a higher MD than wild plants.

The objectives of this study are:

1) To assess whether MD of wild varieties of switchgrass is different than that of cultivars.
2) To determine if there are differences in MD among switchgrass varieties.
3) To determine the impact of mycorrhizal associations on several physiological parameters in wild and cultivated varieties of switchgrass.

Materials and Methods

1.0- Switchgrass Varieties

Switchgrass was selected for this study because of its reported mycorrhizal dependency as a C4 grass species, its many agricultural uses. Because switchgrass has not been under agricultural usage for very long, cultivars may not be very genetically divergent from the wild varieties. All the varieties used in this study are of the 'upland' switchgrass type.

Genotypes of switchgrass were chosen for this experiment to represent several native populations as well as a gradient of selection pressure among the cultivars. Wild varieties were difficult to obtain since there are few harvesters of wild switchgrass seeds. It was not possible to obtain native populations from the U.S because of importation regulations. The wild varieties in this study were selected based on availability. Cultivars of switchgrass were selected based on availability from breeders, and to reflect different levels of selective breeding and ploidies.

Forestburg originates from native accessions that performed well in comparison with other switchgrass varieties. It has not undergone any selective breeding for a specific trait (Barker et al., 1988). In this way, Forestburg is similar to the wild varieties despite being a commercial cultivar. Caddo has been through selection rounds for biomass production and seed size, Shelter was selected for stalk hardiness, Summer had been selected for fungal resistance and Trailblazer had been through selection rounds for IVDMD (Alderson and Sharp, 1994). Northern Upland (NU) is comprised of a mix of many cultivars, Blackwell, Cave-in-Rock, Pathfinder and other breeder stock. Along with the

selection that was done on the cultivars that comprise NU, recurrent selection for biomass was undergone in the formation of this cultivar (C. Taliaferro, pers. comm.). A gradient of agricultural selection pressure is found in the cultivars of this experiment with Forestburg having the least and NU the most.

Of the wild varieties, Prairie Habitat (PH) is likely to be the most similar to original native populations. The seeds of this ecotype came from populations that live in the tallgrass prairie of Manitoba. This is the natural environment of switchgrass, and the population size in the prairie would be large enough to ensure good genetic diversity. Ontario populations come from very small isolated patches and they are likely to be inbred and strongly ecotypic. Ontario Native Plants (ONP) seeds come from the Toronto area, specifically High Park in the city's downtown, and the few fields still existing that contain switchgrass. Ojibway seed comes from the Ojibway Nature Centre located within the city of Windsor, Ontario. Pterophylla seeds are native to the Long Point area in Southwestern Ontario. Environmental adaptations to the particular site of these plants are likely to be emphasized in small populations such as ONP and Ojibway.

Two greenhouse experiments were performed in this study. The first experiment involved all ten varieties and had more replicates. From the results of the first experiment, six varieties were chosen (Forestburg, Caddo, NU, ONP, Ojibway and PH) for the second experiment. They were selected because they included a representative range of mycorrhizal responses, were easy to grow, represented all ploidy levels, and included

equal amount of wild and cultivated types. Changes were made to the experimental design in the second experiment, most noticeably increasing the pot size.

2.0 - Seeds

Seeds of the different cultivars used in this study were obtained from plant breeders (Table 2.1). They are either breeding stock from Universities, stock from the U.S Department of Agriculture (USDA) or from AAC. Wild seeds were obtained from native seed distributors who collect seeds from non-cultivated sources. Prairie Habitat was collected from native tall grass prairie in the Winnipeg area, Ontario Native Plants seeds were collected from the Toronto Region and Ojibway Nature Centre is a natural prairie park in the Windsor area. Pterophylla seeds are native from the Long Point region of Ontario, but the switchgrass is grown in monoculture in order to harvest the seed more efficiently.

Table 2.1. Origin of switchgrass seeds used in both greenhouse experiments.

Cultivation Level	Variety	Year of seed collection	Reported Germ %	Area of Origin	Coordinates of Origin	Distributors	Contact Person
Wild	Prairie Habitat *	1999	unknown	Winnipeg, Manitoba, Canada	49°53 N 97°09 W	Prairie Habitiats Inc., Argyle, MB	John Morgan
	ONP *	1999	unknown	Toronto , Ontario, Canada	43°39 N 79° 23 W	Ontario Native Plant Co. Downsview, ON	Charles Kinsley
	Ojibway *	1999	unknown	Windsor, Ontario, Canada	42°18 N 83°01 W	Ojibway Nature Centre Windsor, ON	Karen Cedar
	Pterophylla	1999	unknown	Long Point, Ontario, Canada	42°26 N 81°54 W	Pterophylla, Long Point, ON	Mary Gartshore
Cultivated	Shelter	1993	68%	St Marys, West Virginia, U.S	39°24 N 81°12 W	USDA New York Plant Materials Center	John Dickerson
	Caddo *	unknown	76%	Central Oklahoma, US	36° 07 N 97°03 W	Oklahoma State University	Charles Taliaferro
	Northern Upland NU 94 –1 *	1999	42%	Nebraska/ Kansas/Illinois/ Oklahoma	36°- 39° N 104° - 85°W	Oklahoma State University	Charles Taliaferro
	Forestburg *	<1998	91%	Forestburg,South Dakota, U.S	44°10 N 98°06 W	USDA North Dakota Plant Materials Center	John Dickerson
	Trailblazer	1994	67%	Nebraska and Kansas, U.S	~40° N ~ 95° W	AAC, Ottawa, ON.	Art Mc Elroy
	Summer	unknown	63%	Nebraska City, Nebraska, U.S	40°41 N 95°52 W	South Dakota State University	Arvid Boe

* - used in the second experiment

3.0- Preparation for Germination

Seeds were prepared based on the procedure detailed by Haynes et al. (1997) for germination of switchgrass. Seeds were scarified in 8 M H_2S0_4 for 5 min., rinsed with distilled water (dH_2O), surface sterilized in 5.25 % NaOCl (Clorox bleach) for 15 min. and rinsed with water. Seeds were placed one layer thick in new covered Petri dishes lined with Whatman #1 filter paper. For the first experiment plants, approximately 3 mL of 0.2% KNO_3 were added to moisten the seeds, dH_2O was used in the second experiment. Sealed Petri dishes were kept at 4 °C in the refrigerator for 14 days to stratify the seeds before removing the plates. Seeds were then transferred onto new Petri dishes with new filter papers and moistened with dH_2O. Seeds were well separated in the Petri dishes and germination was allowed to take place in the greenhouse.

Percentage germination was determined before the first experiment. Five plates of each variety received ten seeds in order to determine the % germination. After seven days, the number of seeds per plate that had emerged coleoptiles was recorded. Petri dishes were moistened daily with dH_2O. For plates that developed contamination, the germinated seeds were transferred to a clean plate. The seedlings were between 0.5 and 2 cm in height when they were planted seven days after germination began.

4.0 - Planting and Soil Preparation

First Experiment

Seedlings were planted seven days after being placed in Petri dishes for germination. The plates were randomly selected for either mycorrhizal (M+) or non-mycorrhizal (M-) treatments. If possible, plates with no contamination were used first. Each container received one germinated seed. During the first seven days of growth, if the plant died, a healthy one from the Petri plates replaced it. After seven days from the planting date, no more replacements were done and the date of plant death was recorded. Forty-nine plants per variety/treatment were planted, for a total of 998 plants.

A week after the first planting, seven more plants of each variety/treatment were planted under the same conditions as the preceding plants. These plants were kept separate and used for the root architecture analysis.

All plants were planted in SC-10 Super Cell Conetainers (Stuewe and Sons, Oregon, USA) that allowed for 164 mL of soil in 21cm depth, 3.8 cm diameter cones. Each cone was surface sterilized with Liquinox before planting. Cones were placed in RL 98 trays that allow for 14 rows of 7 cones, 98 plants in a 61cm x 30 cm tray for a total density of 528 cells/m^2 (Fig 2.1-A). Trays were rotated weekly.

Second Experiment

Seedlings were planted nine days after being placed in the Petri dishes for germination. No seedling deaths occurred. Larger pots were used in the second experiment, as the Conetainers seemed to be limiting in the first experiment. The 500 mL pots were surface sterilized with Liquinox prior to planting. Ten pots of the same variety/treatment were contained in each tray. Trays were rotated as a unit in the greenhouse once a week. Twenty plants per variety/treatment were planted (with the exception of 10 for Ojibway) for a total of 220 plants (Fig 2.1-B,C,D).

5.0 - Soil Medium

Medium consisted of a 1:1:1 ratio of potting soil: vermiculite: moistened sand (Ritchie's Feed and Seed, Ottawa, ON). This mixture was found to have a pH of ~ 4.5 before planting. All substrates were autoclaved separately for 45 min. at 120 °C and allowed to cool for 2 weeks before planting.

In the first experiment, cones were filled with soil mixture to 2 cm from the top, then approximately 0.5 cm (5-10 mL) of mycorrhizal inoculum or non-mycorrhizal control substrate was added to each cone and 0.5-1cm of soil mixture was added to top up the cone. Non-mycorrhizal treatments were prepared first in order to prevent contamination.

In the second experiment, pots were filled 1/3 with soil mixture, then 200 mL of inoculum or non-mycorrhizal control substrate was added and an additional 1 cm of soil topped the pot. Non-mycorrhizal treatments were again prepared first in order to prevent contamination.

The mycorrhizal inoculum consisted of a commercial strain of *Glomus intraradices* Schenck & Smith. (DAOM 181602) (Mycorhize Pro, Premier Tech, Rivière-du-Loup, QC). Premier Tech also supplied the non-mycorrhizal control substrate that contained filtrate of the rhizosphere microflora while excluding fungal spores. This was added to the non-mycorrhizal cones in the same manner as the mycorrhizal inoculum. Mycorhize Pro was obtained 2 weeks before planting and stored at 4 °C.

The fertilizer consisted of a low phosphorus, modified ¼ Hoagland solution adjusted to pH 6 ($NaH_2PO_4 \cdot 2H_2O$, 0.25 mM; KNO_3, 1.5mM; $Ca(NO_3) \cdot 4H_20$, 1.0 mM; $MgSO_4 \cdot 7H_2O$, 0.5mM; FeEDTA, 62.5 µM; $MnSO_4$, 2.29 µM; H_3BO_3, 11.56 µM; $ZnSO_4 \cdot 7H_2O$, 0.19µM; $CuSO_4 \cdot 5H_2O$, 0.08 µM; $H_2MoO_3 \cdot H_2O$, 0.028 µM).

6.0 - Growth Conditions

First Experiment

Seedlings were planted on April 1st (M-) or 2nd (M+) 2000 and harvested June 26 and 27, respectively, after 12 weeks of growth. All plants were grown in the greenhouse under natural light conditions. Light measurements were taken with a LiCor L.I. 185B light

meter unit on May 24 at 1 PM. Measurements were on average 1750 micromoles $M^{-2}sec^{-1}$ in the light and 600 micromoles $M^{-2}sec^{-1}$ in the shade. Plants were rotated clockwise in the greenhouse every week. Each cone was watered daily with distilled water. The ¼ Hoagland solution (20 mL/ plant) was added on April 5, May 3, 17, 30, and June 10.

Second Experiment

Seedlings were planted on August 25, 2000 and harvested on October 24 (M-) and October 25 (M+) after 10 weeks of growth. Plants were harvested early to prevent them from entering the flowering stage. Plants received natural light and 12h of light from sodium lamps. Light measurements were taken with a LiCor L.I. 185B light meter unit September 6th at noon. Average measurements were 1330 micromoles $M^{-2}sec^{-1}$ in the light and 400 micromoles $M^{-2}sec^{-1}$ in the shade. Each pot was watered daily with dH_2O. Trays were rotated clockwise in the greenhouse weekly. Pots were fertilized on August 30, September 13, 27, and October 11, with 20 mL of the ¼ Hoagland solution.

6.1- Growth Measurements

Plant height was measured weekly during the first experiment. The same 14 plants of each variety/treatment were measured. Height from the basal knot to the tip of the tallest leaf was recorded, including only live tissue. Height was measured only at harvest in the second experiment.

6.2- Chlorophyll Determination

Chlorophyll content was determined for the first experiment plants only, as chlorophyll differences were deemed not to be important. Greenness numbers were taken on June 14th2000, using a Minolta SPAD 502 handheld chlorophyll meter (provided by Dr. Malcolm Morrison, AAC, Ottawa, ON). Seven plants were chosen from each variety/treatment for chlorophyll determination. The top three leaves that were completely unfurled were used for chlorophyll determination. A minimum of five points were taken on each leaf, and these were averaged to give a leaf greenness number. Because traditional chlorophyll determination is destructive, samples were taken from Cave-in-Rock switchgrass not used in the experiment in order to create a standard curve for chlorophyll/greenness number. Samples of leaves representing various greenness number readings were assayed for chlorophyll content. Weighed leaf pieces were submerged in 1 mL of 95% ethanol and refrigerated for 14 hrs before the absorbancies at 663 and 645 nm were read on a NovaSpec 2 spectrophotometer. Chlorophyll content was determined using the following equations (Bruinsma, 1963):

$$\text{Chl a} = (\text{OD}_{663} \times 12.7) - (\text{OD}_{645} \times 2.7)$$

$$\text{Chl b} = (\text{OD}_{645} \times 22.9) - (\text{OD}_{663} \times 4.7)$$

$$\text{Total Chl (a+b)} = \text{Chl a} + \text{Chl b}$$

The graph of Chl a+b/ mg tissue vs Greenness number was used to determine the chlorophyll content associated with the chlorophyll meter readings. (See Appendix).

7.0 - Harvest

Plants were harvested over 3 days June 19 to 21, 2000 for the first experiment. For the second experiment plants were harvested on Oct 24 (M-) and Oct 25 (M+) 2000. Plants were harvested during the vegetative stage before entering reproductive stage as described by Moore et al. (1991). Roots of the plants were shaken free of soil substrate and rinsed with water in order to remove any remaining substrate. Height was measured from the basal knot to the tip of the tallest leaf. Tiller numbers were recorded. A tiller was considered to be a photosynthetic vegetative shoot arising from either an underground rhizome or a distinct branching of the stalk at the base. Fresh weights of the plants were taken first as a whole, then roots and shoots were weighed separately. The root sections of the first 10 plants of each variety/treatment were used for determination of acid phosphatase level and mycorrhizal colonization %.

Fresh plants were put in a –80 °C freezer within 2 hours of harvest. Frozen material was lyophilized for 24 hrs (Virtis UNI-TRAP freeze dryer model 10-100, Virtis Co, Gardiner N.Y, US). Dry weight was subsequently measured for root and shoot sections. Dry material was ground to powder in a modified coffee bean grinder (CBM100, Black and Decker Canada Inc, Brockville, ON) with the root and shoot sections remaining separate.

8.0 - Acid Phosphatase Determination

The method used for acid phosphatase determination was derived from McLaughlan (1980). Immediately after being removed from the soil and washed, 0.5g of fresh roots were immersed in 25 mL of phosphate solution [acetic acid buffer (0.2 M sodium acetate-acetic acid made to pH 5) containing 12.5 mg p-nitrophenyl phosphate (PNPP)]. The bottles were kept in the dark for 60 min. at 20 °C. After incubation, 5 mL from the buffered PNPP solution was removed, then titrated to pH 11 with 2N NaOH and brought to a 50 mL volume. A yellow colour appears when the solution becomes basic and the optical density of the solution was read at 400 nm. The colour is stable for 24 hrs. The blank consisted of acetic acid buffer and PNPP without root material. Results are reported in OD 400 nm/g fresh root weight/hr. Ten root samples per variety/ treatment were assayed for acid phosphatase.

9.0 - Root Staining and Percent Colonization Determination

Fresh roots were washed and patted dry. Approximately 0.5g of root material were covered with 2.5% KOH and heated in a hot water bath at 90°C for ~ 10 min. until bleached. Roots were drained, washed with dH_2O and allowed to soak in 1% HCl overnight. The acid was drained and replaced by Aniline Blue staining solution (50 mL 1% HCl, 450 mL dH_2O, 500 mL glycerol, 0.5g aniline blue). Roots were stained by heating in a hot water bath for 3-4 min. and then placed in destaining solution (50 mL 1%HCl, 450 mL dH_2O, 500 mL glycerol). Ten 1 cm root segments/plant were lined up per slide. A polyvinyl alcohol-lactic acid-glycerol medium, (PVLG- Polyvinyl alcohol

8.33g, dH_2O 50 mL, Lactic acid 50mL, Glycerol 50 mL) was used as a preservative and the slide was sealed with clear nail polish.

Percentage colonization was determined by counting mycorrhizal structures on the root segments with a light microscope at 100X and 400X magnification. A root section was considered mycorrhizal if it contained any of the structures associated with *Glomus intraradices*, intercellular hyphae, vesicles or arbuscules (Fig 2.2). Ten plants per variety/ treatment were used to determine colonization. Twenty root sections were observed for each plant (2 slides). A representative number of non-mycorrhizal plants from each of the treatments were observed to verify that no contamination occurred.

10.0 - Root Architecture Determination

The Winrhizo™ Image Analysis Program (Régent Instruments, Québec, QC) was used to determine root architecture. This method has been assessed as reliable to 200 cm root length (Bauhus and Messier, 1999). Fresh roots were scanned (HP DeskScanII, maximum dpi 2400) and then the image analyzed by the Winrhizo™ program. If the root mass was very large, it was broken into 2 parts from the tip, each were scanned separately and the results combined on a spreadsheet. The program was provided for use in the lab of Dr. Yolande Dalpé, AAC, Ottawa, ON.

11.0 – Analysis of Carbon and Nitrogen

Combustion and gas chromatography were used to determine the % of carbon and nitrogen in the dried plant samples. This was done with an Elemental Analyser (Perkin Elmer Series II 2400, Foster City, CA, US) at the JJ Hatch Isotope Laboratory, University of Ottawa. Approximately 4 mg of dried tissue was used for each sample. Three samples of root and shoot portions were measured per variety/treatment. Samples each contained the combined dried ground material of 3 plants.

12.0 - Analysis of Other Macronutrients

Macronutrient analysis was done by ICP-AES (Spectroflame, Analytical Instruments, 1994) at the laboratory of Dr. Henri Dinel, AAC Ottawa, ON. Approximately 500 mg of dried sample were digested in 20 mL of HNO_3 and 8 mL of $HClO_4$ overnight, then placed on a 90 °C hot plate. This solution was allowed to simmer overnight, then 10 mL of HF was added and the solution transferred to a 200 °C hot plate until it turned clear. If there was no residue, the liquid was boiled down to ~ 2 mL and made up to 25 mL with deionized water. The solution was then analyzed by the ICP-AES for K, P, Ca, Mg. Samples consisted of the dried powder of 3 plants combined. Three samples were analyzed per variety/treatment for both root and shoot portions.

13.0 - Chromosome Count

The number of chromosomes for each variety was determined microscopically on dividing root tips by Ms. Svetlana Kritenko in the lab of Dr. George Fedak at AAC, Ottawa, ON. Roots were placed in ice water to arrest division, then fixed in a 3:1 alcohol acetic acid mixture and hydrolyzed in 1N HCl at 60° C for 10 min. Roots were stained with Felgin stain and observed under a light microscope.

14.0 - Statistical Analyses

Statistics were done using the SYSTAT 9 program (SPSS, 1998). Two-way ANOVAs were done to compare the treatment effects with variety and cultivation level for each measured variable. When necessary, the non-parametric Kruskall-Wallis test was done instead. If an interaction was found between variety and treatment, a Tukey's test was performed to determine varietal differences. If treatment was significant with an interaction, t-tests were performed on each variety to assess mycorrhizal treatment effect.

A cluster analysis was performed on the varieties to determine their relationship to each other with regards to mycorrhizal effect. The effect size of the mycorrhizal treatment was determined by finding the difference between the mycorrhizal and non-mycorrhizal averages and dividing by the total standard error, which includes both mycorrhizal and non-mycorrhizal samples.

$$\text{Effect Size} = \frac{\text{Mycorrhizal Average} - \text{Non-Mycorrhizal Average}}{\text{Total Standard Error}}$$

Effect size was calculated and included in the cluster analysis for: height, tiller numbers, total fresh weight, total dry weight, dry root weight, dry shoot weight, R/S, acid phosphatase level, total root length, average root diameter, number of root tips, % small roots, % large roots, % N shoot, % N root, % C shoot and % C root. For the second experiment, macronutrient concentrations of P, K, Ca, and Mg for both root and shoot sections were included in the analysis. All six types of hierarchical cluster joining (Single, Complete, Average, Median, Centroid and Ward) were performed. The most representative and statistically valid test, Ward joining was selected for both first and the second experiments. Stepwise Canonical Discriminant Analysis (CDA) was performed to determine cluster significance.

Figure 2.1 (A-D). Switchgrass plants. A) Experiment 1 growth conditions. B) Experiment 2 growth conditions. C) The wild variety Ojibway M- and M+ plants. D) The cultivar Forestburg M- and M+ plants.

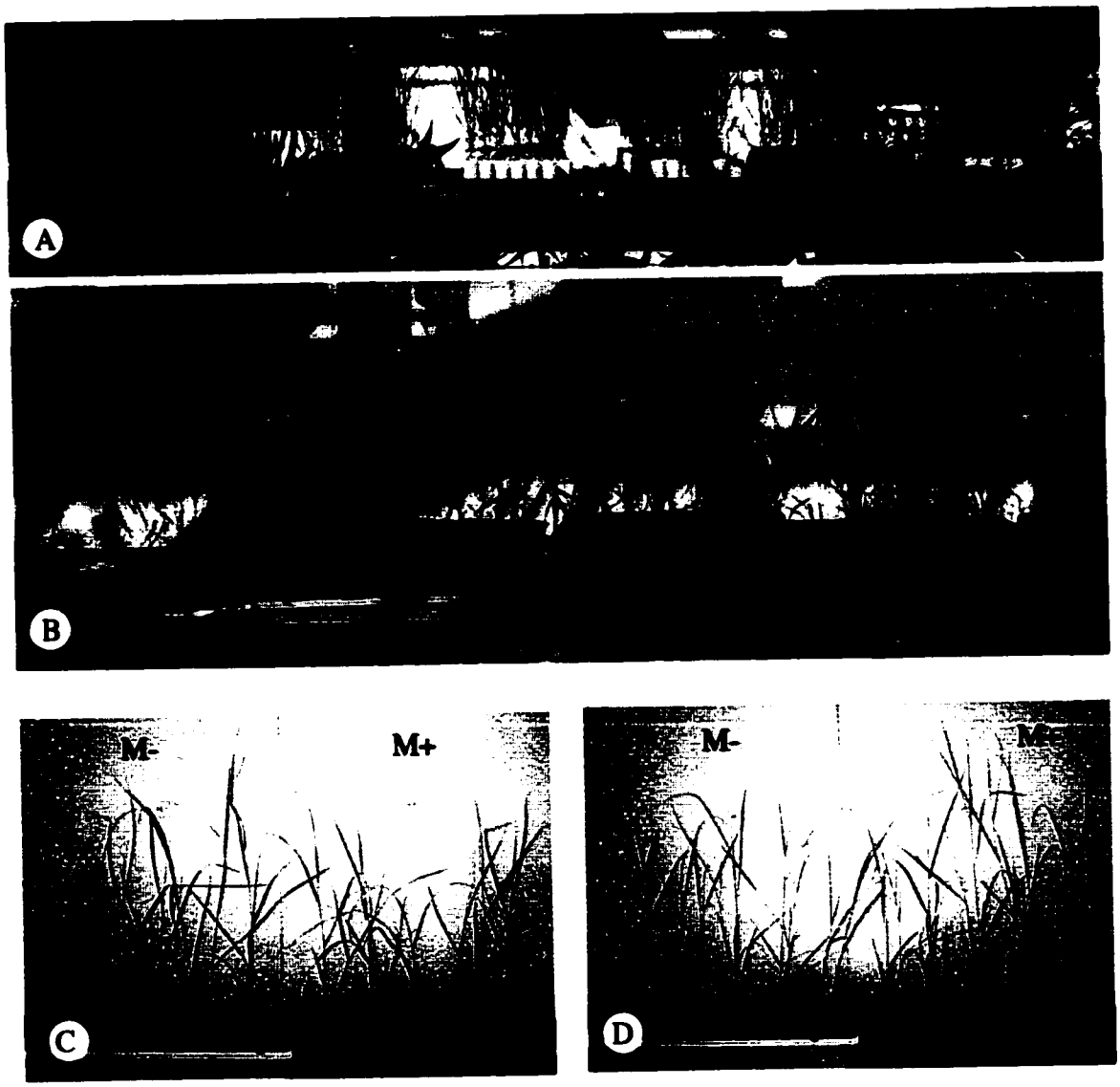
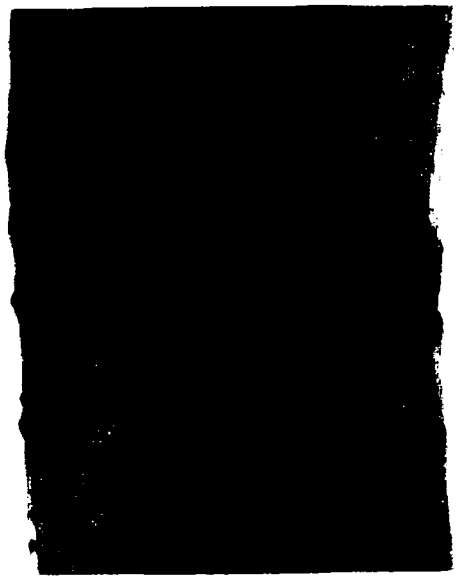

A
B
M-
M+
C
M-
D

Figure 2.2 (A-D). *Glomus intraradices* colonization. A) A heavily colonized root portion seen at 100× magnification. B) Colonized root showing intercellular hyphae and vesicles seen at 100× magnification. C) An arbuscule shown at 400× magnification. D) A vesicle seen at 400× magnification.



A



B

Results

1.0 - Germination

The cultivars of switchgrass had a significantly higher germination percentage than the wild varieties (t-test, p= 0.000) (Table 3.1). There was a significant effect of variety on germination % (one-way ANOVA, p=0.000). Forestburg and Caddo had very high germination rates while Pterophylla and ONP were poor germinators. The wild varieties, except Pterophylla, had husks attached that may hinder seed germination while the cultivars were bare.

2.0 - Chromosome Numbers

Switchgrass has a base number of 9 chromosomes and is commonly found in the tetraploid (2n=36), hexaploid (2n=54) and octoploid (2n=72) forms. The most recently published chromosome counts, and the chromosome numbers we observed are reported in Table 3.2. Wild varieties in this study were all found to be hexaploid. Caddo and Northern Upland were reported to be octoploid, but only 54 chromosomes were counted in these cultivars. Forestburg and Summer had 36 chromosomes as expected. Shelter was correctly reported to be octoploid.

Table 3.1. Germination percentage ± standard error of switchgrass seeds.

Cultivation Level	Variety	% Germination	
Wild	Prairie Habitat	54 % ± 1.7	abcd
	ONP	18 % ± 1.0	cd
	Ojibway	26% ± 2.2	bc
	Pterophylla	16% ± 1.4	d
Cultivated	Shelter	62% ± 1.5	ab
	Forestburg	80% ± 1.1	a
	Summer	44% ± 2.2	abcd
	Caddo	74% ± 1.2	a
	Trailblazer	60% ± 1.7	ab
	Northern Upland	58% ± 2.0	abc
Wild		29%	β
Cultivated		63%	α

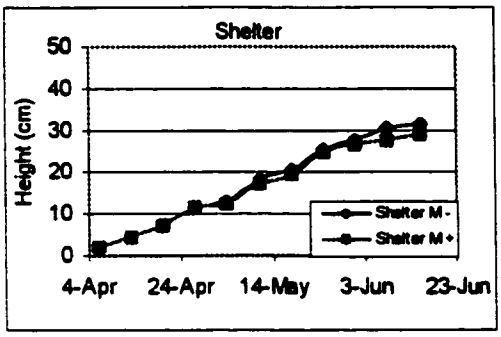
Significant differences between varieties as determined by a Tukey's test at $\alpha = 0.05$ following a one-way ANOVA at $\alpha = 0.05$ indicated by different letters.
Significant differences between cultivation level as determined by a t-test at $\alpha = 0.05$ indicated by α, β.

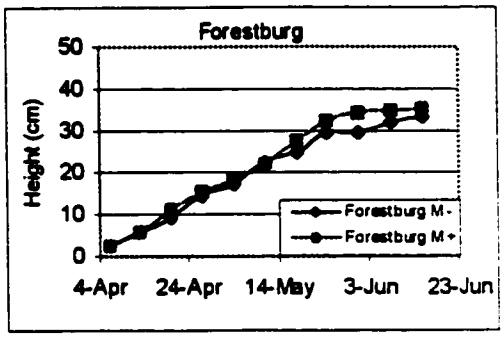
Table 3.2. Chromosome numbers in switchgrass varieties.

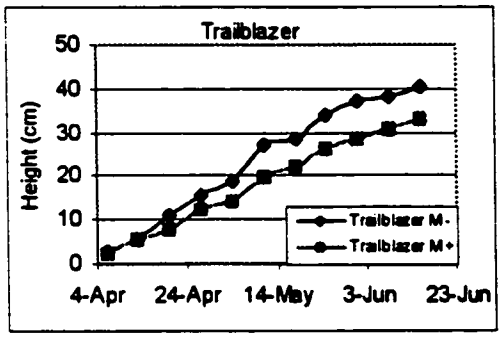
Cultivation Level	Variety	Ploidy	Previously Reported Ploidy
Wild	Prairie Habitat	54	-
	ONP	54	-
	Ojibway	54	-
	Pterophylla	54	-
Cultivated	Shelter	72	72ϕ
	Forestburg	36	36π
	Summer	36	36$\phi\gamma$
	Caddo	54	72ϕ
	Trailblazer	54	54γ
	Northern Upland	54	72λ

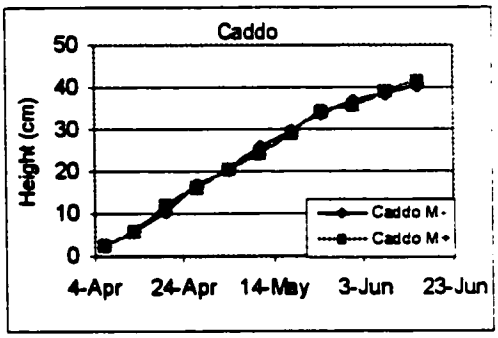
ϕ - Hopkins et al. (1996)
γ - Riley and Vogel (1982)
λ - C. Taliaferro (personal communication)
π - Elbersen et al. (1999)

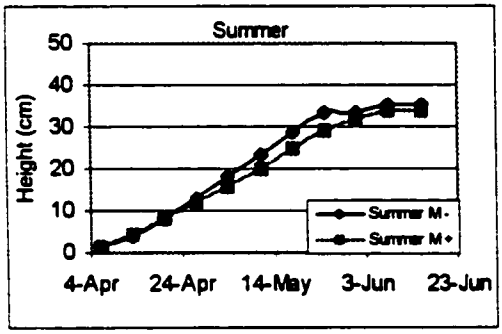
Figures 3.1-3.10. Growth rate as measured by plant height. Graphs show weekly height from basal knot to tallest leaf tip for 14 plants from each variety/ treatment. M+ plants are indicated by ■ and M- plants with ♦. Cultivars are Shelter, Forestburg, Trailblazer, Caddo, Summer and NU. Wild varieties are ONP, Pterophylla, PH, and Ojibway.

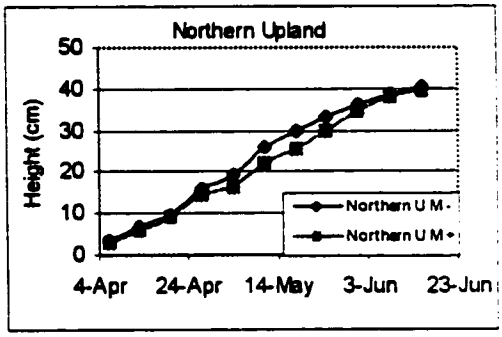

Shelter
Height (cm)
50
40
30
20
10
0
4-Apr
24-Apr
14-May
3-Jun
23-Jun
Shelter M-
Shelter M+

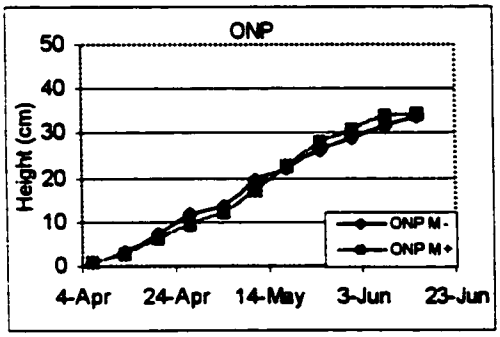

Forestburg
Height (cm)
50
40
30
20
10
0
4-Apr
24-Apr
14-May
3-Jun
23-Jun
Forestburg M-
Forestburg M+

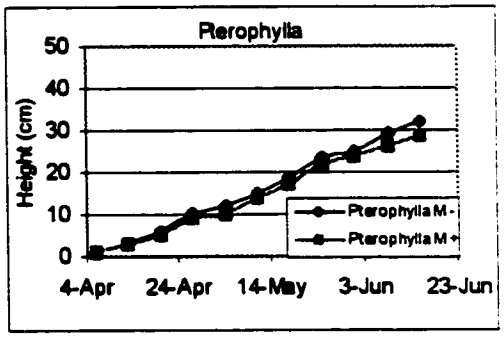

Trailblazer
Height (cm)
50
40
30
20
10
0
4-Apr
24-Apr
14-May
3-Jun
23-Jun
Trailblazer M-
Trailblazer M+

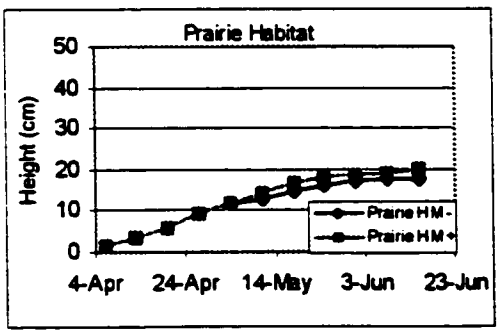

Caddo
Height (cm)
50
40
30
20
10
0
4-Apr
24-Apr
14-May
3-Jun
23-Jun
Caddo M-
Caddo M+

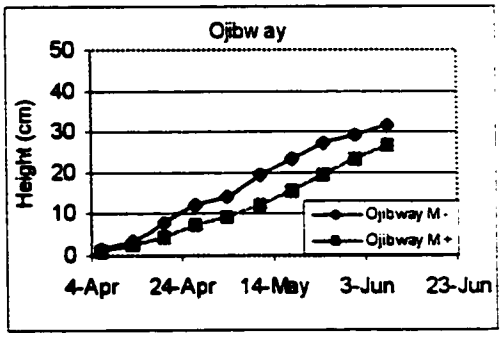

Summer
Height (cm)
50
40
30
20
10
0
4-Apr
24-Apr
14-May
3-Jun
23-Jun
Summer M-
Summer M+


Northern Upland
Height (cm)
50
40
30
20
10
0
4-Apr
24-Apr
14-May
3-Jun
23-Jun
Northern U M-
Northern U M+


ONP
Height (cm)
50
40
30
20
10
0
4-Apr
24-Apr
14-May
3-Jun
23-Jun
ONP M-
ONP M+


Pterophylla
Height (cm)
50
40
30
20
10
0
4-Apr
24-Apr
14-May
3-Jun
23-Jun
Pterophylla M-
Pterophylla M+


Prairie Habitat
Height (cm)
50
40
30
20
10
0
4-Apr
24-Apr
14-May
3-Jun
23-Jun
Prairie H M-
Prairie H M+


Ojibway
Height (cm)
50
40
30
20
10
0
4-Apr
24-Apr
14-May
3-Jun
23-Jun
Ojibway M-
Ojibway M+

3.0- Growth Rate

Growth rate was measured by the height difference of the first experiment plants. The cultivars grew to a greater height than the wild varieties (Figs. 3.1-3.10). The difference in plant height between cultivars and wild varieties was present from the germination stage. The growth curve profiles for mycorrhizal (M+) and non-mycorrhizal (M-) treatments were similar for most of the varieties. Trailblazer and Ojibway showed a divergence where the M- plants grew taller. These differences were significant at the time of harvest (see section 6.3). By the end of the 12-week growth period, most of the plants had reached a growth plateau.

4.0 - Chlorophyll Concentration

Overall, the chlorophyll concentration was quite similar among varieties and M+/M- treatments (Table 3.3). Variety ($p= 0.000$) and mycorrhizal treatment ($p=0.000$) were both significant effects, without significant interaction ($p= 0.913$). Caddo, Summer, NU and Trailblazer had lower chlorophyll levels than Pterophylla, Forestburg and Shelter. The M- treatment had a significantly higher average chlorophyll concentration (24.11) than the M+ treatment (22.93). Wild plants were found to have significantly higher (24.47) chlorophyll concentrations than cultivars (22.93) ($p= 0.040$), with no interaction between cultivation level and treatment ($p=0.873$).

Table 3.3. Chlorophyll concentration ± standard error in switchgrass varieties.

Cultivation Level	Variety	Chlorophyll Concentration		
		M-	M+	
Wild	Prairie Habitat	26.14 ±1.21	24.87 ±1.15	ab
	ONP	24.11 ±0.47	22.00 ±0.61	bc
	Ojibway	23.65 ±0.56	23.06 ±0.79	bc
	Pterophylla	26.92 ±0.94	24.98 ±1.04	a
Cultivated	Shelter	24.70 ±0.54	22.97 ±0.59	abc
	Forestburg	26.19 ±0.98	23.74 ±1.47	ab
	Summer	23.19 ±0.86	21.38 ±0.88	c
	Caddo	21.74 ±1.33	21.43 ±0.58	c
	Trailblazer	22.21 ±0.91	22.18 ±0.72	c
	Northern Upland	22.94 ±0.98	22.52 ±0.67	c
		♦		
Wild		25.20	23.73	α
Cultivated		23.49	22.37	β
		♦		

Significant differences between M-/+ treatments for all varieties determined by a two-way ANOVA at $\alpha = 0.05$ with no interaction between myc and variety indicated by ♦.
Significant differences between varieties as determined by a Tukey's test at $\alpha = 0.05$ indicated by different letters.
Significant differences between cultivation level as determined by a two-way ANOVA at $\alpha = 0.05$ indicated by α, β.

5.0 - Mycorrhizal Colonization

First Experiment

On average, 37% of the plant roots were colonized by *Glomus intraradices* (Fig 3.11). There was a significant effect of variety on colonization level (p=0.007), which was due to a difference between Caddo (59%) and ONP (21%) based on a Tukey's test. The cultivation level did not affect the percentage of mycorrhizal colonization. No mycorrhizae were observed in the non-mycorrhizal plants.

Second Experiment

The colonization percentage was similar in the second experiment, with the average being 36% (Fig 3.12). No statistical differences were detected among the varieties or between the cultivation levels. As in the first experiment, no mycorrhizae were observed in the non-mycorrhizal plants.

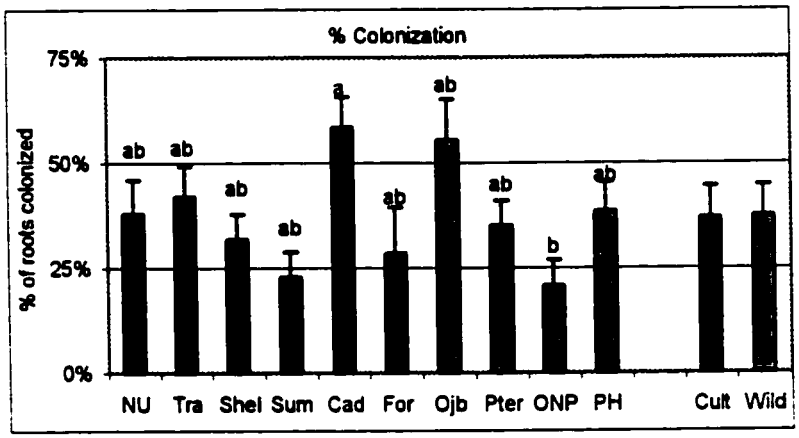


Figure 3.11. The colonization percentage with *G. intraradices* from the first experiment. Switchgrass varieties are indicated on the x-axis. Different letters indicate significant differences as detected by a Tukey's test at $\alpha = 0.05$.

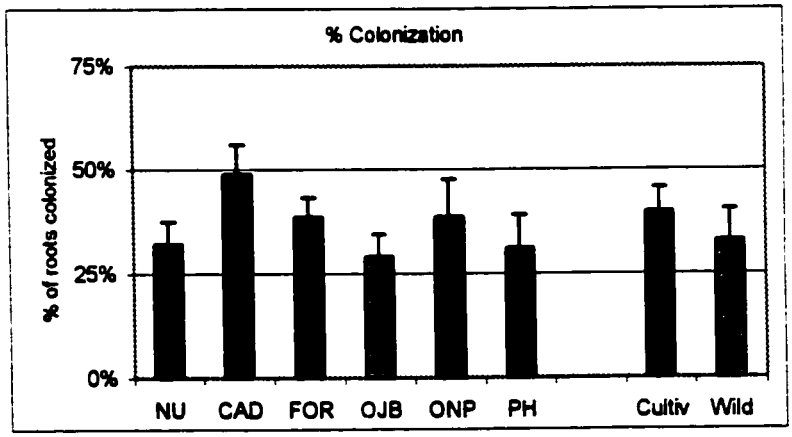


Figure 3.12. The colonization percentage with *G. intraradices* from the second experiment. Switchgrass varieties are indicated on the x-axis. No significant differences were detected.

6.0 - Physiological Data

6.1- Mass

First Experiment

The cultivars had significantly higher fresh and dry masses than wild varieties (Table 3.4). The mycorrhizal treatment decreased the fresh and dry masses for cultivars as a group, due to a decrease in the shoots rather than the roots (Fig. 3.13). Significant differences among varieties were found for fresh, dry, shoot and root masses (Table 3.4.1). Because of significant interactions, each variety was analyzed separately for mycorrhizal effect. Dry mass decreased significantly in the Ojibway, Pterophylla, Caddo, Trailblazer and NU varieties and increased in the ONP and Shelter varieties with the mycorrhizal treatment (Table 3.4). Fresh mass decreased significantly in PH, Ojibway, Pterophylla, Shelter, Caddo and Trailblazer and increased in Summer under the M+ treatment.

Second Experiment

As in the first experiment, cultivars had significantly higher fresh and dry masses than the wild varieties (Table 3.5). A significant varietal effect was found for fresh, dry, shoot and root masses (Tables 3.5 and 3.4.1). The mycorrhizal treatment effect was not significant and did not affect mass for most of the varieties (Table 3.4.1). Varieties were analyzed separately because of a significant interaction with

mycorrhizae. Significant increases in mass with the mycorrhizal treatment were seen in Forestburg for dry, root and shoot mass; NU for fresh and root masses; and ONP for fresh mass (Table 3.5). The increases in mass with the mycorrhizal treatment were not significant over cultivation levels (Fig. 3.17 and Table 3.4.1).

6.2 – Root to Shoot Ratio

First Experiment

The wild varieties had significantly higher root to shoot ratios (R/S) than the cultivated varieties (Table 3.4). Mycorrhizae significantly increased the R/S for both cultivation levels (Fig. 3.14). Variety and mycorrhizal treatment had significant effects and interaction (Table 3.4.1). The increased R/S with mycorrhizae was significant for ONP, Ojibway, Summer, and Caddo (Table 3.4). Among the varieties, PH and Forestburg had the highest R/S while ONP and Caddo had the lowest. Other varieties were intermediate.

Second Experiment

Varieties had significantly different R/S ratios (Table 3.5). The R/S of PH and Forestburg were found to be significantly higher than the others. There was no significant mycorrhizal treatment effect or interaction detected for the R/S (Table 3.4.1). The wild and cultivated groups were not different from each other (Fig. 3.19).

6.3 – Plant Height

First Experiment

Cultivated plants had a superior height compared to wild plants after the 12-week growth period (Table 3.4). Significant varietal, mycorrhizal and interaction effects were found for height (Table 3.4.1). PH plants were much smaller than the others; NU and Caddo grew to the greatest height (Table 3.4). For ONP and Forestburg the height increased with the M+ treatment, in Ojibway, Pterophylla, Shelter, Caddo, and Trailblazer the height decreased with mycorrhizae. Over both cultivation levels, the mycorrhizae decreased the total height of the plants (Fig. 3.15).

Second Experiment

The height of cultivated plants was greater than that of the wild varieties (Table 3.5; Fig. 3.19). A significant effect of variety on height was detected, while mycorrhizal effect and the interaction were non-significant (Table 3.4.1). PH was found to be shorter than all the rest; NU and Caddo were the tallest varieties (Table 3.5).

6.4 – Number of Tillers

First Experiment

The number of vegetative shoots was found to be higher in the wild varieties than the cultivars; this difference was greater in the M+ treatment (Table 3.4). A decrease in tiller numbers with mycorrhizae was found over both cultivation levels (Fig. 3.16). Variety, mycorrhizal treatment and their interaction all had significant effects (Table 3.4.1). PH, ONP, Caddo and Trailblazer had the largest amounts of tillers, while Shelter had the least (Table 3.4). Tiller numbers decreased significantly in M+ plants for Ojibway, Pterophylla, Summer, Caddo, Trailblazer, and NU.

Second Experiment

The cultivated and wild plants had similar tiller numbers in this experiment (Table 3.5). The decrease in tiller numbers with mycorrhizae for both wild and cultivated groups seen in the first experiment was also found in the second experiment (Fig 3.20). There was no significant effect of variety, but there was a significant mycorrhizal effect and an interaction term (Table 3.4.1). The tiller number decrease was significant in Caddo and NU (Table 3.5).

6.5- Summary of Physiological Parameters

First Experiment

Cultivated varieties had higher mass and height than the wild plants. However, root/ shoot ratio was higher in the wild plants than the cultivars. The shoot mass, height and tiller numbers all decreased with the mycorrhizal treatment as a general trend while R/S increased. There was a significant varietal effect for all the physiological factors. The effect of mycorrhizae was dependent upon variety because of significant interactions. Ojibway, Pterophylla, Caddo, and Trailblazer showed a consistent significant negative effect to the mycorrhizal treatment. ONP and Forestburg had positive responses to mycorrhizae. The varieties PH, Shelter, Summer and NU had no clear response to the mycorrhizal treatment.

Second Experiment

Varieties were quite widespread in their physiological parameters and general response to the mycorrhizal treatment. Cultivated plants had higher mass, height and than wild plants. Mycorrhizae tended to increase the mass but decrease the tiller numbers. Significant varietal differences were found for all the physiological parameters except tiller numbers. The mycorrhizal treatment did not have a significant effect in most cases. ONP and Forestburg responded positively to mycorrhizae, other varieties showed no clear response trend.

Table 3.4. Physiological data for experiment 1 switchgrass plants ± standard error.

	Variety	Fresh Mass (g)			Dry Mass (mg)			Root Mass (mg)			Shoot Mass (mg)		
		M-	M+		M-	M+		M-	M+		M-	M+	
Wild	Prairie Habitat	0.93 ±0.04 *	0.56 ±0.03	f	189 ±12	179 ±16	e	105 ±9	89 ±11	e	84 ±5	90 ±16	e
	ONP	1.20 ±0.04	1.28 ±0.03	bcd	248 ±11 *	340 ±25	cd	93 ±6 *	167 ±25	d	155 ±6	173 ±6	cd
	Ojibway	1.55 ±0.06 *	1.11 ±0.04	bc	312 ±16 *	286 ±8	bcd	134 ±8	133 ±5	cd	177 ±8 *	153 ±4	cd
	Pterophylla	1.31 ±0.04 *	1.07 ±0.02	de	322 ±8 *	266 ±7	d	146 ±5 *	126 ±4	cd	176 ±5 *	140 ±3	d
Cultivated	Shelter	1.26 ±0.04 *	0.93 ±0.03	e	275 ±22 *	283 ±8	c	144 ±7 *	131 ± 4	bc	164 ±7	152 ±5	cd
	Forestburg	1.63 ±0.05	1.72 ±0.04	a	336 ± 7	351 ±8	a	164 ±6 *	179 ±5	a	172 ±5	172 ±5	bcd
	Summer	1.13 ±0.02 *	1.32 ±0.03	cd	348 ±7	332 ±7	ab	150 ±5	156 ±5	b	198 ±5 *	176 ±5	ab
	Caddo	1.58 ±0.03 *	1.12 ±0.02	b	371 ±9 *	294 ±8	ab	130 ±4	121 ±4	c	241 ±7 *	173 ±5	a
	Trailblazer	1.45 ±0.07 *	1.03 ±0.05	bcd	342 ±23 *	261 ±16	bcd	148 ±9 *	109 ±7	cd	195 ±14 *	152 ±9	bc
	Northern Upland	1.29 ±0.03	1.33 ±0.02	bc	376 ±7 *	335 ±7	a	157 ±5	148 ±4	ab	219 ±5 *	188 ±4	a
Wild		1.25 *	1.00	α	268	268	α	120	129	α	148 *	139	α
Cultivated		1.39 *	1.24	β	341 *	309	β	149	141	β	198 *	169	β

Significant differences between M-/+ treatments determined by a t-test at α =0.05 indicated by *.
Significant differences between varieties as determined by a Tukey's test at α =0.05 indicated by different letters.
Significant differences between cultivation level as determined by a two-way ANOVA at α =0.05 indicated by α, β.

Table 3.4 continued. Physiological data for experiment 1 switchgrass plants ± standard error.

	Variety	R/S			Height (cm)			# Tillers		
		M-	M+		M-	M+		M-	M+	
Wild	Prairie Habitat	1.36 ±0.22	1.18 ±0.10	a	19.78 ±0.64	18.98 ±0.75	f	2.37 ±0.14	2.22 ±0.10	a
	ONP	0.59 ±0.03 *	1.05 ±0.22	de	34.26 ±0.59 *	36.40 ±0.62	c	2.22 ±0.11	2.24 ±0.12	a
	Ojibway	0.75 ±0.02 *	0.87 ±0.03	bc	34.35 ±0.53 *	31.69 ±0.39	de	2.33 ±0.08 *	1.96 ±0.09	abc
	Pterophylla	0.84 ±0.03	0.90 ±0.03	b	33.94 ±0.52 *	30.09 ±0.44	de	2.29 ±0.10 *	1.92 ±0.08	abc
Cultivated	Shelter	0.87 ±0.04	0.88 ±0.03	b	32.86 ±0.62 *	29.28 ±0.45	e	1.96 ±0.11	1.80 ±0.10	c
	Forestburg	0.98 ±0.04	1.06 ±0.04	a	32.90 ±0.55 *	35.29 ±0.61	cd	1.88 ±0.10	1.90 ±0.10	bc
	Summer	0.78 ±0.04 *	0.90 ±0.04	bc	35.45 ±0.38	35.70 ±0.44	c	2.29 ±0.11 *	1.71 ±0.10	abc
	Caddo	0.55 ±0.02 *	0.70 ±0.02	e	41.77 ±0.66 *	39.42 ±0.52	ab	2.49 ±0.08 *	2.00 ±0.10	a
	Trailblazer	0.81 ±0.06	0.71 ±0.03	cd	39.62 ±1.02 *	35.67 ±0.94	b	2.56 ±0.15 *	2.10 ±0.11	a
	Northern Upland	0.73 ±0.03	0.80 ±0.02	cd	41.63 ±0.65	41.80 ±0.52	a	2.53 ±0.10 *	1.90 ±0.10	ab
Wild		0.89 *	1.00	α	30.58 *	29.29	α	2.30 *	2.09	α
Cultivated		0.78 *	0.84	β	37.37 *	36.19	β	2.28 *	1.90	β

Significant differences between M-/+ treatments determined by a t-test at α =0.05 indicated by *.
Significant differences between varieties as determined by a Tukey's test at α =0.05 indicated by different letters.
Significant differences between cultivation level as determined by a two-way ANOVA at α =0.05 indicated by α, β.

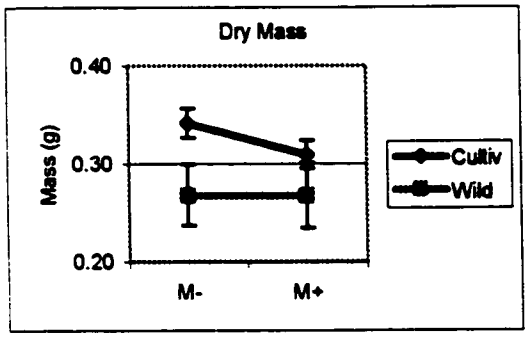


Fig 3.13. Dry mass

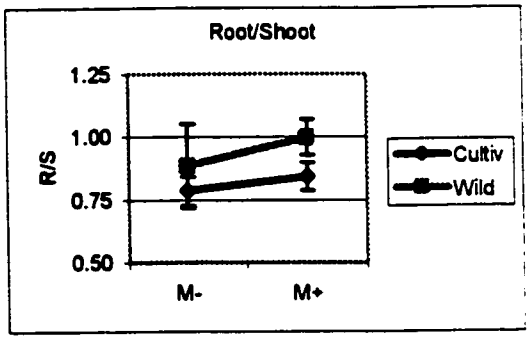


Fig 3.14. Root/Shoot Ratio

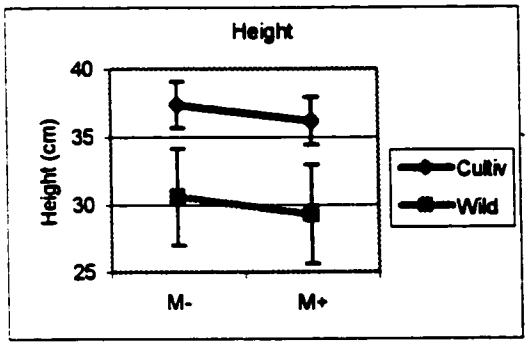


Fig 3.15. Plant Height

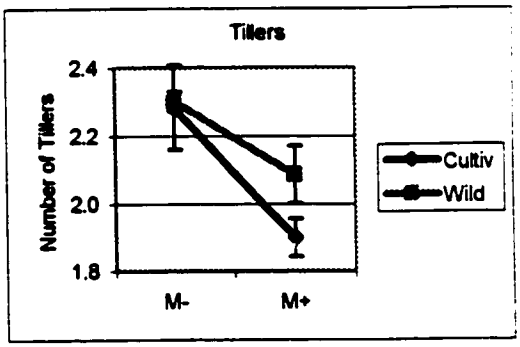


Fig 3.16. Number of Tillers

Figures 3.13-3.16. Physiological parameters of experiment 1 switchgrass plants. Differences between cultivation level and mycorrhizal treatment are shown for dry mass, R/S, height and tiller numbers. Standard error bars are indicated. Line graphs are used to visualize the interaction between cultivation level and mycorrhizal treatment; it is not intended to suggest any continuous variable.

Table 3.4.1. Results of two-way ANOVAs at $\alpha = 0.05$ for physiological parameters of the switchgrass plants from the first and second experiments. The first ANOVA was done for variety and mycorrhizal treatment, the second for cultivation level and mycorrhizal treatment.

	First ANOVA				Second ANOVA			
	R^2	V, p=	T, p=	V×T, p=	R^2	C, p=	T, p=	C×T, p=
Exp 1, n= 980								
Dry Mass*	0.437	0.000	0.000	0.000	0.172	0.000	0.004	0.005
R/S*	0.388	0.000	0.000	0.000	0.049	0.013	0.000	0.381
Height	0.714	0.000	0.000	0.000	0.216	0.000	0.005	0.804
Tillers*	0.102	0.000	0.000	0.000	0.050	0.026	0.000	0.091
Exp 2, n= 220								
Dry Mass*	0.572	0.000	0.270	0.007	0.308	0.000	0.229	0.897
R/S*	0.570	0.000	0.610	0.861	0.020	0.220	0.636	0.561
Height*	0.513	0.000	0.335	0.093	0.364	0.000	0.212	0.905
Tillers	0.236	0.508	0.000	0.000	0.089	0.905	0.000	0.038

Use of a non-parametric test indicated by *.
V= variety, T= mycorrhizal treatment, C= cultivation level.

Table 3.5. Physiological data for experiment 2 switchgrass plants ± standard error.

	Variety	Fresh Mass (g)			Dry Mass (mg)			Root Mass (mg)			Shoot Mass (mg)		
		M-	M+		M-	M+		M-	M+		M-	M+	
Wild	Prairie Habitat	1.56 ±0.13	1.94 ±0.14	d	376 ±37	357 ±36	c	169 ±17	161 ±14	b	207 ±27	196 ±25	c
	ONP	2.12 ±0.11 *	2.88 ±0.27	bc	461 ±17	615 ±80	b	133 ±11	161 ±25	b	328 ±10	454 ±59	b
	Ojibway	2.23 ±0.26	1.77 ±0.17	cd	336 ±58	286 ±14	c	92 ±22	82 ±8	c	244 ±40	204 ±69	c
Cultivated	Forestburg	2.93 ±0.11	3.47 ±0.24	a	635 ±52 *	972 ±61	a	232 ±18 *	340 ±15	a	403 ±35 *	632 ±51	a
	Caddo	2.92 ±0.23	2.69 ±0.16	ab	604 ±45	505 ±34	b	181 ±19	139 ±13	b	423 ±31	366 ±26	ab
	Northern Upland	2.89 ±0.13 *	2.25 ±0.20	b	563 ±35	619 ±53	b	154 ±9 *	165 ±16	b	409 ±28	454 ±37	ab
Wild		1.97	2.20	α	391	419	α	131	134	α	260	285	α
Cultivated		2.92	2.80	β	601	699	β	189	215	β	412	484	β

Significant differences between M-/+ treatments determined by a t-test at α =0.05 indicated by *.
Significant differences between varieties as determined by a Tukey's test at α =0.05 indicated by different letters.
Significant differences between cultivation level as determined by a two-way ANOVA at α =0.05 indicated by α, β.

Table 3.5 continued. Physiological data for experiment 2 switchgrass plants ± standard error.

	Variety	R/S			Height (cm)			# Tillers		
		M-	M+		M-	M+		M-	M+	
Wild	Prairie Habitat	0.90 ±0.12	0.91 ±0.11	a	30.78 ± 1.99	40.69 ± 2.75	c	2.3 ±0.14	1.9 ±0.18	a
	ONP	0.41 ±0.03	0.36 ±0.04	b	50.57 ± 1.70	49.31 ± 2.07	b	1.9 ±0.19	2.0 ±0.20	a
	Ojibway	0.36 ±0.07	0.42 ±0.07	b	49.98 ±3.51	47.46 ± 2.09	b	2.1 ±0.31	1.6 ±0.27	a
Cultivated	Forestburg	0.58 ±0.02	0.56 ±0.05	a	54.98 ± 1.71	53.65 ± 2.30	b	1.6 ±0.13	2.1 ±0.21	a
	Caddo	0.43 ±0.03	0.39 ±0.04	b	58.69 ± 1.45	63.02 ± 1.43	a	2.5 ±0.11 *	1.5 ±0.14	a
	Northern Upland	0.38 ±0.02	0.36 ±0.01	b	60.62 ± 1.58	64.30 ± 2.39	a	2.9 ±0.21 *	1.4 ±0.13	a
Wild		0.56	0.56	α	43.78	45.82	α	2.1 *	1.8	α
Cultivated		0.47	0.44	α	58.10	60.32	β	2.3 *	1.6	α

Significant differences between M-/+ treatments determined by a t-test at $\alpha = 0.05$ indicated by *.
Significant differences between varieties as determined by a Tukey's test at $\alpha = 0.05$ indicated by different letters.
Significant differences between cultivation level as determined by a two-way ANOVA at $\alpha = 0.05$ indicated by α, β.

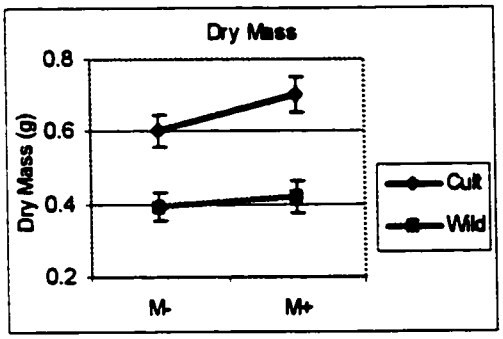


Fig 3.17. Dry Mass

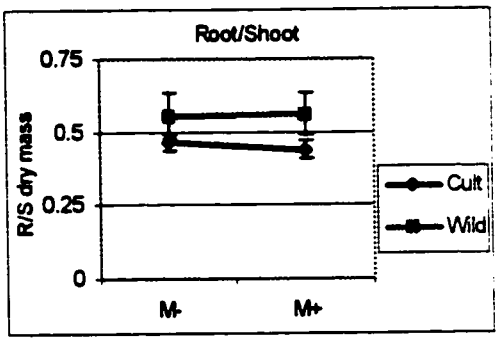


Fig 3.18. Root /Shoot ratio

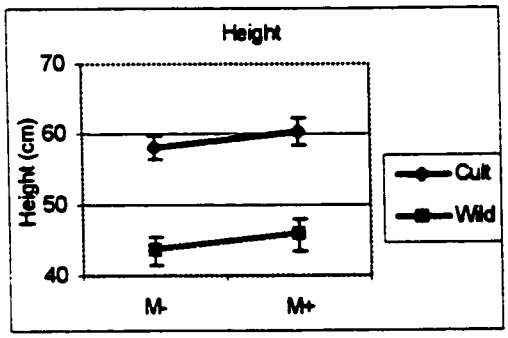


Fig 3.19. Plant Height

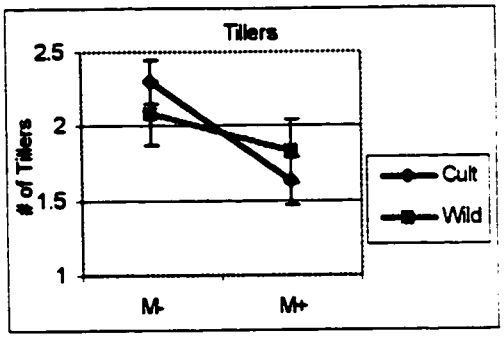


Fig 3.20. Number of Tillers

Figures 3.17-3.20. Physiological parameters of switchgrass from experiment 2 plants. Differences between cultivation level and mycorrhizal treatment are shown for dry mass, R/S, height and tiller numbers. Standard error bars are indicated. Line graphs are used to visualize the interaction between cultivation level and mycorrhizal treatment; it is not intended to suggest any continuous variable.

7.0 – Acid Phosphatase

First Experiment

Cultivated plants had significantly (p=0.012) lower phosphatase levels in the roots than wild plants (Table 3.6). The acid phosphatase levels decreased significantly in the mycorrhizal treatment for the cultivated group only (p=0.000) (Fig. 3.21). Significant differences were found for variety, mycorrhizal treatment and their interaction (p=0.000,0.000,0.000). The highest levels of acid phosphatase were found in the varieties PH, ONP, Ojibway, Shelter, and Trailblazer (Table 3.6). The acid phosphatase level decreased significantly in mycorrhizal Shelter, Forestburg, Trailblazer and NU.

Second Experiment

Cultivars had significantly (p=0.010) lower acid phosphatase levels than wild varieties in the second experiment as well (Table 3.7; Fig. 3.22). The mycorrhizal treatment did not have a significant effect (p=0.614) on the acid phosphatase level in any of the varieties. Differences between varieties were detected (p=0.000) but the interaction with mycorrhizae was not significant (p=0.170). The highest acid phosphatase levels were found in PH and Caddo.

Table 3.6. Acid phosphatase levels ± standard error for experiment 1 switchgrass plants expressed in ΔOptical Density at 400 nm/g FM roots/hr.

	Variety	ΔO.D 400 nm/g/hr		
		M-	M+	
Wild	Prairie Habitat	2.14 ± 0.20	2.57 ± 0.22	a
	ONP	2.17 ± 0.14	1.69 ± 0.13	a
	Ojibway	1.70 ± 0.16	2.46 ± 0.28	a
	Pterophylla	1.34 ± 0.09	1.12 ± 0.06	c
Cultivated	Shelter	2.76 ± 0.18	1.83 ± 0.09 *	a
	Forestburg	2.16 ± 0.20	1.12 ± 0.11 *	b
	Summer	1.47 ± 0.12	1.18 ± 0.10	bc
	Caddo	1.29 ± 0.16	1.29 ± 0.08	bc
	Trailblazer	2.36 ± 0.15	1.57 ± 0.11 *	a
	Northern Upland	1.89 ± 0.12	1.04 ± 0.06 *	bc
Wild		1.84 ± 0.02	1.96 ± 0.05	α
Cultivated		2.30 ± 0.01	1.51 ± 0.01 *	β

Significant differences between M-/+ treatments determined by a t-test at α =0.05 indicated by *.
Significant differences between varieties as determined by a Tukey's test at α =0.05 indicated by different letters.
Significant differences between cultivation level as determined by a two-way ANOVA at α =0.05 indicated by α, β.

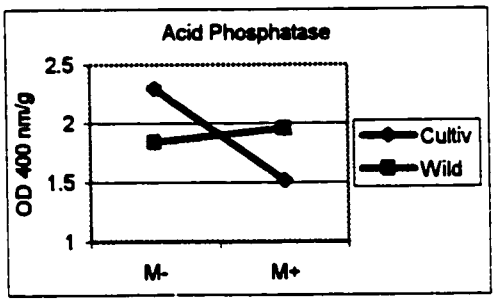


Fig 3.21. Acid phosphatase levels, experiment 1 switchgrass plants

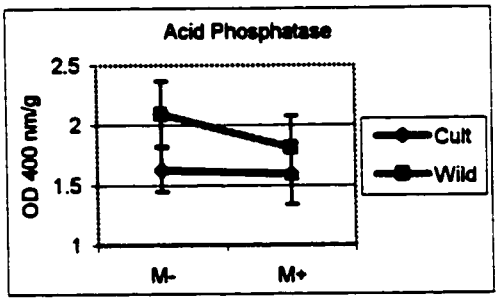


Fig. 3.22. Acid phosphatase levels, experiment 2 switchgrass plants

Figures 3.21-3.22. Acid phosphatase levels of switchgrass expressed in ΔOptical Density at 400 nm/g FM roots/hr. Comparisons of cultivation level and mycorrhizal treatment are shown for both experiments. Standard errors bars are shown. Line graphs are used to visualize the interaction between cultivation level and mycorrhizal treatment; it is not intended to suggest any continuous variable.

Table 3.7. Acid phosphatase levels ± standard error for experiment 2 switchgrass plants expressed in ΔOptical Density at 400 nm/g FM roots/hr.

	Variety	ΔO.D 400nm/g/hr		
		M-	M+	
Wild	Prairie Habitat	2.59 ± 0.34	1.98 ± 0.19	a
	ONP	1.85 ± 0.18	1.52 ± 0.24	ab
	Ojibway	1.83 ± 0.30	1.94 ± 0.36	ab
Cultivated	Forestburg	1.06 ± 0.07	1.41 ± 0.16	b
	Caddo	2.65 ± 0.43	1.86 ± 0.23	a
	Northern Upland	1.17 ± 0.13	1.49 ± 0.15	b
Wild		2.09± 0.27	1.81 ± 0.26	α
Cultivated		1.63 ± 0.18	1.59 ± 0.25	β

Significant differences between varieties as determined by a Tukey's test at α =0.05 indicated by different letters.
Significant differences between cultivation level as determined by a two-way ANOVA at α =0.05 indicated by α, β.

8.0 - Root Architecture

The root parameters most relevant to this study were total root length, average root diameter, number of root tips (to represent branching), the length of roots with small diameter (0-0.1mm) and the length of roots with large diameter (0.9 mm and over).

8.1 – Total Root Length

First Experiment

The cultivated and wild groups did not differ in their root lengths (Table 3.8). Significant mycorrhizal and varietal effects were found, with no interaction (Table 3.8.1). The mycorrhizal treatment significantly increased total root length for both the wild and cultivated groups (Fig 3.23) and over all varieties. The longest root lengths were in Pterophylla and Trailblazer, while NU and PH had the shortest root lengths.

Second Experiment

Mycorrhizae did not have a significant effect on root length in the second experiment, though the M+ plants tended to have longer root lengths than M- plants (Fig. 3.28). Significant differences among varieties and a marginally significant interaction with mycorrhizae were detected (Table 3.8.1). Forestburg had the longest root lengths while Ojibway had the shortest (Table 3.9). Cultivation level did not have a

significant effect on root length. The overall root lengths were higher in the second experiment than in the first because of the increased soil volume available.

8.2 – Average Root Diameter

First Experiment

Root diameter varied significantly amongst varieties (Table 3.8). The varieties Shelter and Trailblazer had the largest diameters while PH had the smallest. Significant effects of variety, mycorrhizal treatment and a marginally significant interaction were detected (Table 3.8.1). There was a significant decrease in diameter with the mycorrhizal treatment overall and for Shelter, Summer, and NU (Table 3.8). Cultivars had a larger diameter than wild varieties in the M- treatment (Fig. 3.24). The mycorrhizal treatment effect was no longer significant when assessed over cultivation level (Table 3.8.1).

Second Experiment

Variety was the most important factor determining root diameter in the second experiment (Table 3.9). A significant effect of variety, and an interaction with mycorrhizal treatment was detected (Table 3.8.1). Forestburg, Caddo and PH had the largest root diameters while Ojibway had the smallest (Table 3.9). Mycorrhizae significantly increased root diameter for Forestburg and NU, and significantly decreased it for Ojibway. A root diameter increase with mycorrhizae was found to be

significant for the cultivars but not the wild varieties (Fig. 3.29). As in the first experiment, cultivars had overall a significantly larger root diameter than wild varieties in the second experiment.

8.3 – Number of Root Tips

First Experiment

Cultivation level did not affect the number of root tips, but mycorrhizae significantly increased root tip numbers for the wild varieties (Table 3.8; Fig. 3.25). The number of root tips was affected by significant effects of variety, mycorrhizal treatment and their interaction (Table 3.8.1). The greatest number of root tips was found in Pterophylla, the least in PH (Table 3.8). This result may be expected since Pterophylla has the largest root system and PH the smallest. The addition of mycorrhizae tended to increase root tip numbers. This increase was significant in Pterophylla, PH, and Caddo; a significant decrease occurred in NU.

Second Experiment

Root tip numbers tended to decrease in the M+ plants in this experiment (Fig. 3.30). No significant effects were found although variety was a marginally significant factor (Table 3.8.1). ONP had the greatest number of root tips (Table 3.9). The large standard error associated with this parameter might have contributed to the lack of significant differences.

8.4 – Percentage of Small Roots

Small roots include the length of root in the category of the smallest root diameter (< 0.1 mm). To make this length relative, it was divided by the total root length for a specific plant and multiplied by 100%. Of the total root length, 7.5% to 16% fell into this category in the first experiment, and 11 % to 16% in the second experiment.

First Experiment

Significant effects of variety, mycorrhizal treatment and their interaction were found for percentage of small root length (Table 3.8.1). Shelter and NU had the highest % small roots PH had the lowest (Table 3.8). Mycorrhizae significantly increased % small root length overall. Ojibway, Shelter, experienced significant increases with mycorrhizal treatment and NU experienced a decrease. Wild varieties had a significant increase with mycorrhizae but cultivars did not (Fig. 3.26). Cultivars had an overall larger % small root length.

Second Experiment

Varietal differences in small roots was less pronounced in the second experiment, but still statistically significant (Tables 3.9 and 3.8.1). ONP, Forestburg and NU had the highest % of small roots while PH had the lowest (Table 3.9). Mycorrhizae had the expected effect of decreasing % of small root length in the second experiment. This

effect was not dependent upon variety (Table 3.8.1). As in the first experiment, cultivars had significantly higher % small roots than wild varieties (Fig. 3.31).

8.5 – Percentage of Large Roots

Large roots include the length of root in the category of the largest root diameter ($\geq$ 0.9 mm). To make this length relative, it was divided by the total root length for a specific plant and multiplied by 100%. The percentage of roots that fell into this category was much less than the % of small roots. Percent of large roots were found to be between 1.0% - 3.2% in the first experiment, and 0.4% - 1.2% in the second.

First Experiment

Significant effects of variety, mycorrhizal treatment and their interaction were found for the length of large roots (Table 3.8.1). Trailblazer had the most % large roots while PH and ONP had the least (Table 3.8). Forestburg, Caddo and NU experienced a significant increase in % of large root length with mycorrhizae. The other varieties tended to increase % large roots in the M+ treatment, except for ONP and Shelter. Mycorrhizae significantly increased % large roots for cultivars. Cultivars had a higher % of large roots than wild varieties (Fig. 3.27).

Second Experiment

The percentage of large roots was much smaller in the second experiment (Table 3.9). A significant varietal effect (Table 3.8.1) was due to NU having a higher proportion of large roots than the others (Table 3.9). Mycorrhizae had no significant overall effect, but some varieties were affected. Ojibway had a significant decrease in the % large roots and Forestburg had a significant increase. No significant differences were found between the cultivation levels, although cultivars tended to increase their % large roots with M+ while wild varieties decreased theirs (Fig. 3.32).

6.6- Summary of Root Parameters

First Experiment

In general, mycorrhizae had the effect of increasing root length, the amount of root tips, and % small roots while decreasing root diameter. For % large roots, no definite trend was observed. Varietal differences existed for all the root parameters. Pterophylla, Shelter and Trailblazer had large root systems; PH had a small root system. In root tips and % small roots, where interactions with cultivation level and treatment occurred, wild varieties responded more dramatically to mycorrhizae.

Second Experiment

Fewer differences were detected in this experiment. For root length, number of root tips and % large roots, no definite trend was observed. Average root diameter decreased with mycorrhizae in wild varieties and significantly increased in cultivars; this was the only significant cultivation level interaction. An overall decrease in the % small roots was observed with mycorrhizae. Varietal differences existed for all root parameters except root tips. Forestburg had the largest root system; Ojibway's was smaller and had finer roots.

Table 3.8. Root architecture data ± standard error from experiment 1 switchgrass plants.

	Variety	Root Length (cm)			Average Root Diameter (mm)			Number of Root Tips		
		M-	M+		M-	M+		M-	M+	
Wild	Prairie Habitat	463.9 ±45.6 *	712.8 ±70.9	cd	0.342 ±0.009	0.350 ±0.007	c	919 ±108 *	1531 ±160	c
	ONP	615.1 ±96.3	907.7 ±128.2	bcd	0.647 ±0.055	0.564 ±0.053	ab	1351 ±224	2223 ±343	bc
	Ojibway	704.0 ±71.6	792.0 ±99.9	bcd	0.436 ±0.050	0.427 ±0.061	bc	1532 ±219	1719 ±266	abc
	Pterophylla	1041.6 ±95.7	1176.5 ±103.0	a	0.583 ±0.058	0.647 ±0.058	ab	2122 ±220 *	3107 ±350	a
Cultivated	Shelter	784.4 ±40.5	895.0 ±81.9	bc	0.760 ±0.023 *	0.520 ±0.059	a	2141 ±94	2366 ±191	ab
	Forestburg	750.0 ±74.6	868.5 ±99.2	bc	0.647 ±0.046	0.521 ±0.067	ab	1598 ±183	1670 ±134	abc
	Summer	595.6 ±49.2 *	743.9 ±30.9	bcd	0.452 ±0.046 *	0.377 ±0.009	bc	1475 ±169	1782 ±200	abc
	Caddo	598.5 ±35.5 *	868.1 ±112.8	bcd	0.380 ±0.006	0.522 ±0.074	b	1185 ±118 *	1949 ±315	bc
	Trailblazer	892.0 ±59.5	939.2 ±69.1	ab	0.750 ±0.022	0.603 ±0.076	a	1664 ±161	1827 ±138	abc
	Northern Upland	524.8 ±61.1	568.7 ±42.0	d	0.626 ±0.065 *	0.412 ±0.013	ab	1991 ±227 *	1096 ±98	bc
Wild		706.1 ♦	897.2	α	0.505	0.497	α	1481 *	2145	α
Cultivated		690.9 ♦	813.9	α	0.603	0.493	β	1676	1782	α

Significant differences between M-/+ treatments for all varieties determined by a two-way ANOVA at $\alpha = 0.05$ with no interaction between myc and variety indicated by ♦.
Significant differences between M-/+ treatments determined by a t-test at $\alpha = 0.05$ indicated by *.
Significant differences between varieties as determined by a Tukey's test at $\alpha = 0.05$ indicated by different letters. Significant differences between cultivation level as determined by a two-way ANOVA at $\alpha = 0.05$ indicated by α, β.

Table 3.8 continued. Root architecture data from experiment 1 switchgrass plants ± standard error.

	Variety	% Small Root Length			% Large Root Length		
		M-	M+		M-	M+	
Wild	Prairie Habitat	7.5 ±1.33	9.2 ±1.68	c	0.8 ±0.16	0.8 ±0.12	c
	ONP	9.8 ±1.94	11.7 ±2.31	bc	1.3 ±0.22	1.0 ±0.23	c
	Ojibway	8.9 ±1.68	12.2 ±2.12 *	bc	1.5 ±0.32	2.5 ±0.70 .	ab
	Pterophylla	9.8 ±1.48	12.5 ±1.55	abc	1.1 ±0.25	1.7 ±0.25	bc
Cultivated	Shelter	10.1 ±0.60	13.8 ±1.39 *	a	3.0 ±0.40	2.0 ±0.15	ab
	Forestburg	8.6 ±1.36	9.6 ±1.12	bc	1.0 ±0.09	2.1 ±0.29 *	bc
	Summer	11.4 ±1.74	12.9 ±1.34	ab	1.8 ±0.22	2.2 ±0.29	ab
	Caddo	9.5 ±1.44	11 ±1.71	bc	1.5 ±0.23	2.7 ±0.45 *	ab
	Trailblazer	8.9 ±1.3	9.5 ±0.59	bc	2.5 ±0.22	2.8 ±0.14	a
	Northern Upland	16 ±2.14	9.8 ±1.41 *	a	1.6 ±0.57	3.2 ±0.37 *	ab
Wild		9.0	11.4 *	α	1.2	1.5 *	α
Cultivated		10.7	11.1	β	1.9	2.5 *	β

Significant differences between M-/+ treatments determined by a t-test at $\alpha = 0.05$ indicated by *.
Significant differences between varieties as determined by a Tukey's test at $\alpha = 0.05$ indicated by different letters.
Significant differences between cultivation level as determined by a two-way ANOVA at $\alpha = 0.05$ indicated by α, β.

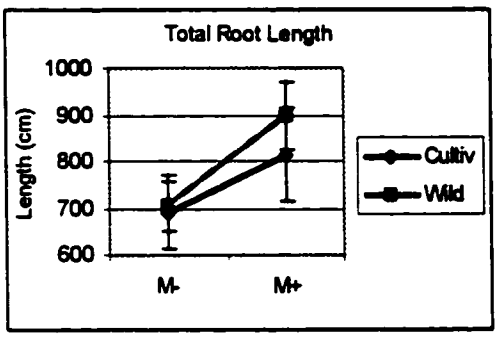


Fig 3.23. Total root length

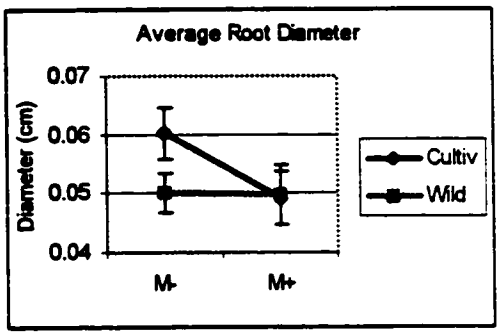


Fig 3.24. Average root diameter

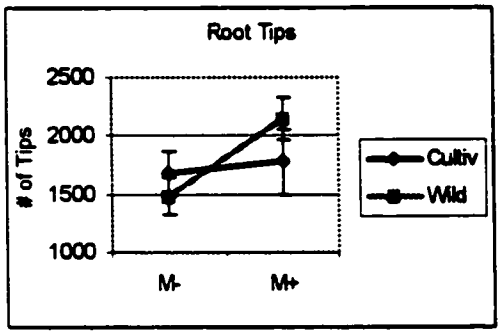


Fig 3.25. Number of root tips

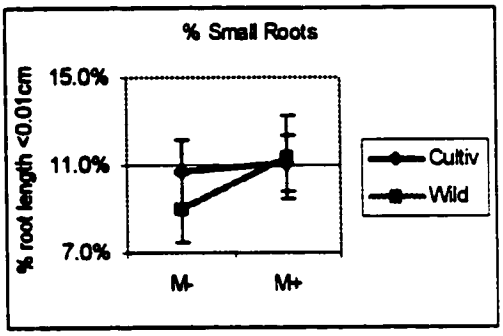


Fig 3.26. % Small roots

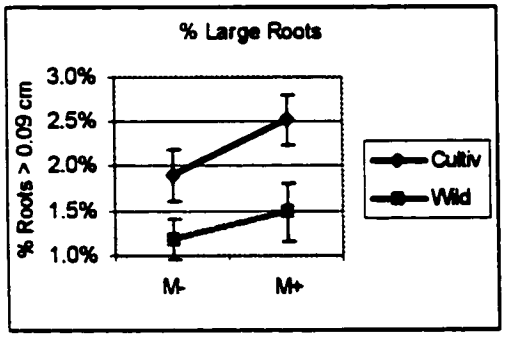


Fig 3.27. % Large roots

Figures 3.23-3.27. Root architecture parameters of switchgrass for experiment 1 plants. Differences between cultivation level and mycorrhizal treatment are shown for root length, root diameter, number of root tips, % small and large roots. Standard error bars are shown. Line graphs are used to visualize the interaction between cultivation level and mycorrhizal treatment; it is not intended to suggest any continuous variable.

Table 3.8.1. Results of two way ANOVAs at $\alpha = 0.05$ for switchgrass root parameters. First ANOVA was done for variety and mycorrhizal treatment, second results are for cultivation level and mycorrhizal treatment.

	First ANOVA				Second ANOVA			
	R^2	V, p=	T, p=	V×T, p=	R^2	C, p=	T, p=	C×T, p=
Exp 1, n=140								
Root Length	0.458	0.000	0.000	0.731	0.095	0.224	0.000	0.442
Root Diameter*	0.500	0.000	0.005	0.060	0.106	0.003	0.055	0.174
Root Tips	0.466	0.000	0.001	0.001	0.088	0.812	0.002	0.018
% Small Roots	0.603	0.000	0.025	0.000	0.090	0.045	0.015	0.036
% Large Roots*	0.551	0.000	0.000	0.000	0.293	0.000	0.007	0.482
Exp 2, n=110								
Root Length	0.290	0.001	0.060	0.054	0.043	0.412	0.068	0.534
Root Diameter	0.429	0.000	0.128	0.000	0.198	0.017	0.075	0.000
Root Tips	0.153	0.088	0.119	0.367	0.028	0.565	0.165	0.464
% Small Roots	0.322	0.000	0.000	0.189	0.160	0.014	0.001	0.200
% Large Roots	0.367	0.000	0.567	0.001	0.057	0.435	0.920	0.190

Use of a Non-Parametric test indicated by *.
V= variety, T= mycorrhizal treatment, C= cultivation level.

Table 3.9. Root architecture data from experiment 2 switchgrass plants ± standard error.

	Variety	Root Length (cm)			Average Root Diameter (mm)			Number of Root Tips		
		M-	M+		M-	M+		M-	M+	
Wild	Prairie Habitat	974.1 ±101.1	1032.1 ±97.4	ab	0.359 ±0.007	0.359 ±0.003	a	2037 ±205	1961 ±515	a
	ONP	987.0 ±99.5	1094.1 ±65.5	ab	0.356 ±0.036	0.348 ±0.008	ab	2760 ±347	2788 ±237	a
	Ojibway	818.7 ±52.33	827.7 ±40.9	b	0.347 ±0.008 *	0.313 ±0.002	b	2407 ±146	1933 ±275	a
Cultivated	Forestburg	1011.3 ±51.0	1446.4 ±122.8	a	0.347 ±0.004 *	0.431 ±0.011	a	2527 ±126	2698 ±428	a
	Caddo	999.1 ±124.4	912.9 ±65.2	b	0.357 ±0.006	0.361 ±0.007	a	2378 ±291	2230 ±251	a
	Northern Upland	870.4 ±69.7	930.2 ±82.3	b	0.336 ±0.006 *	0.370 ±0.007	ab	2821 ±165	1850 ±162	a
Wild		926.4	984.6	α	0.353	0.340	α	2401	2227	α
Cultivated		960.3	1096.5	α	0.347 *	0.387	β	2575	2259	α

Significant differences between M-/+ treatments for all varieties determined by a two-way ANOVA at $\alpha = 0.05$ with no interaction between myc and variety indicated by ♦.
Significant differences between M-/+ treatments determined by a t-test at $\alpha = 0.05$ indicated by *.
Significant differences between varieties as determined by a Tukey's test at $\alpha = 0.05$ indicated by different letters.
Significant differences between cultivation level as determined by a two-way ANOVA at $\alpha = 0.05$ indicated by α, β.

Table 3.9 continued. Root architecture data from experiment 2 switchgrass plants ± standard error.

	Variety	% Small Root Length			% Large Root Length		
		M-	M+		M-	M+	
Wild	Prairie Habitat	13 ±0.90	11 ±0.74	b	0.7 $\pm 7\times10^{-4}$	0.8 $\pm 8\times10^{-4}$	ab
	ONP	14 ±1.37	15 ±0.71	a	0.9 $\pm 8\times10^{-4}$	0.8 $\pm 7\times10^{-4}$	a
	Ojibway	14 ±1.61	12 ±0.36	ab	1.1 $\pm 78\times10^{-4}$ *	0.4 $\pm 93\times10^{-4}$	ab
Cultivated	Forestburg	15 ±0.56	14 ±2.14	a	0.8 $\pm 9\times10^{-4}$ *	1.2 $\pm 7\times10^{-4}$	a
	Caddo	14 ±1.33	13 ±0.73	ab	0.9 $\pm 7\times10^{-4}$	0.9 $\pm 10\times10^{-4}$	a
	Northern Upland	16 ±0.8	13 ±1.0	a	0.6 $\pm 4\times10^{-4}$	1.1 $\pm 10\times10^{-4}$	b
		♦					
Wild		14	12	α	0.9	0.7	α
Cultivated		15	13	β	0.8	1.1	α
		♦					

Significant differences between M-/+ treatments for all varieties determined by a two-way ANOVA at $\alpha = 0.05$ with no interaction between myc and variety indicated by ♦.
Significant differences between M-/+ treatments determined by a t-test at $\alpha = 0.05$ indicated by *.
Significant differences between varieties as determined by a Tukey's test at $\alpha = 0.05$ indicated by different letters.
Significant differences between cultivation level as determined by a two-way ANOVA at $\alpha = 0.05$ indicated by α, β.

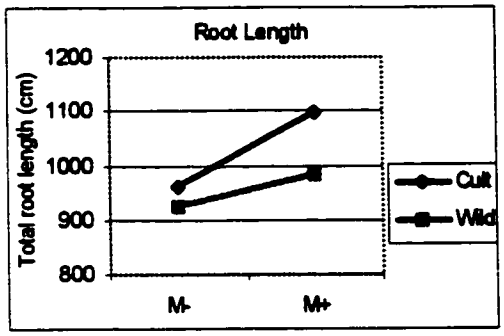


Fig 3.28. Total root length

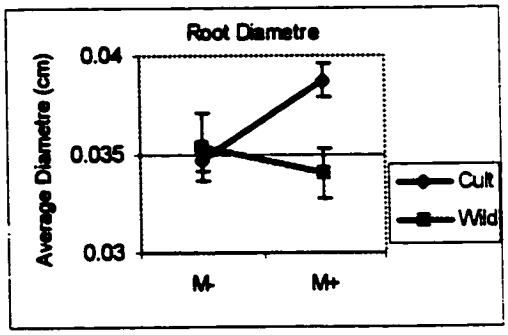


Fig 3.29. Average root diameter

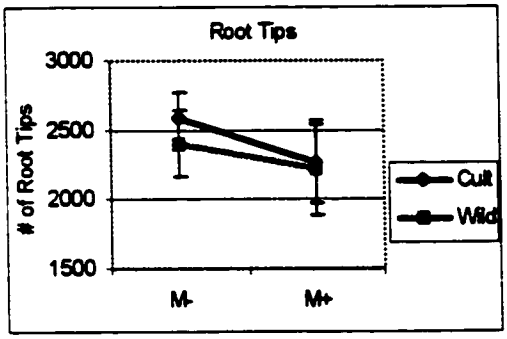


Fig 3.30. Number of root tips

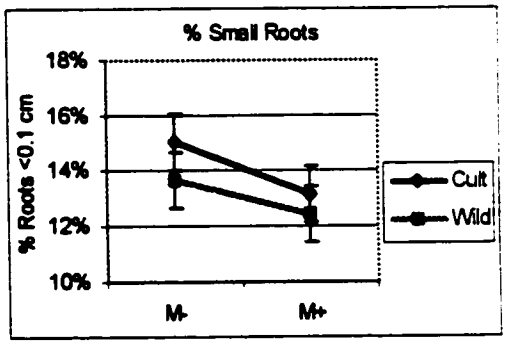


Fig 3.31. % Small roots

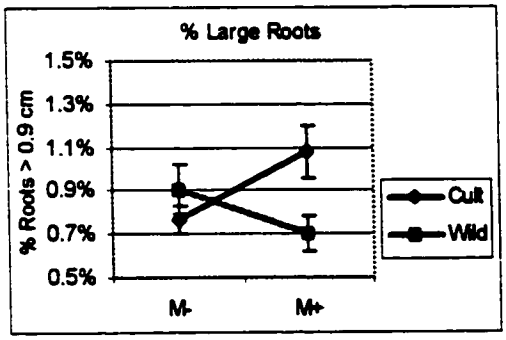


Fig 3.32. % Large roots

Figures 3.28-3.32. Root architecture parameters of switchgrass for experiment 2 plants. Differences between cultivation level and mycorrhizal treatment are shown for root length, root diameter, number of root tips, % small and large roots. Standard error bars are shown. Line graphs are used to visualize the interaction between cultivation level and mycorrhizal treatment, it is not intended to suggest any continuous variable.

9.0- Mineral Content

9.1 – Carbon

First Experiment

Roots and shoots contained similar carbon contents (%C), 45% and 44% respectively (Table 3.10). Significant effects of variety and mycorrhizal treatment were found for both roots and shoots; their interaction was significant in roots only (Table 3.11.1). A significant decrease in %C occurred in the mycorrhizal plants, even though the difference between M+ and M- was less than 1% in all cases, except in Forestburg roots (Table 3.10). Varieties differed by similarly small amounts. Carbon levels of cultivars were slightly higher than those of wild varieties; this difference was statistically significant (Figs. 3.33 and 3.34).

Second Experiment

As in the first experiment, %C was similar in roots (45%) and shoots (43%) (Table 3.11). In the shoots, a significant effect of variety and an interaction with the mycorrhizal treatment were found; in the roots, all factors were significant (Table 3.11.1). Caddo and Ojibway had a higher %C in shoots than ONP or PH. In roots, no differences among varieties were found despite the significant varietal effect. Mycorrhizae significantly increased %C in the shoots of Caddo. Wild varieties and cultivars were not significantly different from each other (Table 3.11) but %C of

roots significantly increased in wild varieties and decreased in cultivars (Figs. 3.37 and 3.38).

9.2 – Nitrogen

First Experiment

The total nitrogen content (%N) was higher in shoots (1.2%) than roots (0.8%) (Table 3.10). Significant effects of variety, mycorrhizal treatment and interaction were found for shoot %N; in roots all except the mycorrhizal effect were significant (Table 3.11.1). PH had high %N in both portions, while Caddo and Pterophylla had low %N (Table 3.10). Decreases in % N were noted in the shoots of mycorrhizal PH, Pterophylla, and Shelter. No mycorrhizal effect was found over cultivation levels (Table 3.11.1). Wild varieties had higher %N than cultivars, and these tended to decrease with the M+ treatment (Figs. 3.35 and 3.36).

Second Experiment

Similarly to the first experiment, shoots had a greater % N (1.4%) than roots (0.9%) (Table 3.11). Total N contents tended to be higher in the second experiment than the first. Significant effects were found for variety, mycorrhizal treatment and interaction in both shoots and roots (Table 3.11.1). In the shoots, PH had the highest %N, Forestburg and NU the lowest (Table 3.11). In the roots, Ojibway had the highest %N and Forestburg the lowest. Mycorrhizae significantly increased the %N in the shoots

of Caddo, and the roots of ONP, Caddo and Ojibway. The wild varieties had higher %N than the cultivars in both shoots and roots; significant increases in %N with M+ were seen in both cultivation levels (Figs. 3.39 and 3.40).

9.3- Phosphorus

Shoots contained higher P levels (970 μg/g) than roots (680 μg/g). The varietal effect and interaction were significant for roots and shoots; the mycorrhizal effect was significant for roots only (Table 3.11.2). The highest P concentrations were found in Ojibway (shoots) and PH (roots), and the lowest in Caddo (Table 3.12). Mycorrhizae significantly increased root P levels overall, and in Forestburg. Root P level significantly decreased with mycorrhizae in PH. The mycorrhizal effect was not significant over cultivation level groups (Table 3.11.2). The P concentration was significantly higher in wild varieties than cultivars (Figs. 3.41 and 3.42).

9.4- Calcium

Average calcium concentration was higher in shoots (1.5 mg/g) than roots (1.1 mg/g). A significant effect of variety was found for both roots and shoots while the interaction was significant for shoots (Table 3.11.2). No effects of the mycorrhizal treatment were found. Interestingly, PH had the highest Ca concentration in shoots and lowest in roots, while Caddo had the lowest Ca in shoots and highest in roots (Table 3.12). Wild varieties had higher shoot Ca concentrations than cultivars, no differences were found in roots (Figs. 3.43 and 3.44).

9.5- Potassium

Potassium levels were higher in the roots (21.8 mg/g) than the shoots (19.2 mg/g). Significant varietal effects and interactions are found in both roots and shoots; mycorrhizal treatment effect was significant in shoots only (Table 3.11.2). Varietal differences are due to PH (shoots and roots) and Forestburg (roots) having lower K levels than the others (Table 3.13). Increases with M+ were found in the shoots of ONP and the roots of Caddo, while a significant decrease was found in the root K level of Caddo. No differences were found between cultivars and wild varieties in either roots or shoots (Figs. 3.45 and 3.46).

9.6- Magnesium

Root sections contained higher Mg levels (4.0 mg/g) than shoots (2.2 mg/g) (Table 3.13). No significant effects were detected for variety, mycorrhizal treatment or interactions (Table 3.11.2). Cultivars and wild varieties had similar Mg concentrations in both roots and shoots (Figs. 3.47 and 3.48).

9.7 – Overall Mineral Effects

First Experiment

An increase in mineral content with mycorrhizae was not observed in the first experiment plants, when available soil volume was low. The N content tended to decrease in mycorrhizal plants, though this was not significant. Mycorrhizal plants had a significantly lower %C than the non-mycorrhizal ones, with the absolute difference between M- and M+ being very small in each case.

Second Experiment

Mineral levels were all in the expected range, except for high Mg. Roots contained on average, higher levels of N, K and Mg while shoots contained higher P and Ca levels. Mycorrhizae increased the concentrations of N, P, K and Ca depending on variety. Wild plants tended to respond positively to mycorrhizae with increased Ca concentrations while cultivars had higher P levels than wild types in the mycorrhizal treatment. Wild plants tended to have higher mineral concentrations than cultivated plants. No differences in %C were observed with the mycorrhizal treatment.

Table 3.10. Total percentage of C and N ± standard error for experiment 1 switchgrass plants.

	Variety	% C Shoot			% C Root			% N Shoot			%N Root		
		M-	M+		M-	M+		M-	M+		M-	M+	
Wild	Prairie Habitat	43.61 ±0.21	43.54 ±0.06	b	45.52 ±0.12	45.80 ±0.02	a	1.82 ±0.05 *	1.40 ±0.08	a	0.95 ±0.04	0.89 ±0.07	a
	ONP	44.32 ±0.06	43.91 ±0.27	ab	46.28 ±0.64	44.59 ±0.15	ab	1.09 ±0.05	1.16 ±0.02	bcd	1.56 ±0.39	0.80 ±0.02	a
	Ojibway	44.08 ±0.29	43.54 ±0.12	ab	44.14 ±0.45	44.62 ±0.11	b	1.31 ±0.01	1.19 ±0.04	b	0.76 ±0.02	0.66 ±0.03	ab
	Pterophylla	43.77 ±0.12	43.42 ±0.30	ab	45.09 ±0.10	45.00 ±0.17	ab	1.18 ±0.04 *	0.99 ±0.00	bcd	0.61 ±0.04	0.62 ±0.01	b
Cultivated	Shelter	43.71 ±0.32	43.58 ±0.11	b	44.93 ±0.48	44.47 ±0.61	ab	1.44 ±0.05 *	1.06 ±0.04	b	0.95 ±0.04	0.63 ±0.02	ab
	Forestburg	44.77 ±0.21	43.47 ±0.14	ab	46.88 ±1.06	42.13 ±3.04	ab	1.15 ±0.06	1.23 ±0.12	bc	1.32 ±0.51	0.75 ±0.04	ab
	Summer	43.97 ±0.49	43.60 ±0.13	ab	45.75 ±0.19	44.92 ±0.24	ab	1.11 ±0.08	1.25 ±0.11	bc	0.57 ±0.01	0.82 ±0.01	ab
	Caddo	44.10 ±0.21	43.78 ±0.11	ab	45.60 ±0.15	45.40 ±0.25	ab	0.86 ±0.06	0.89 ±0.03	d	0.58 ±0.04	0.64 ±0.02	b
	Trailblazer	44.58 ±0.12	44.09 ±0.12	ab	46.53 ±0.81	45.54 ±0.27	a	1.02 ±0.07	1.07 ±0.04	bcd	1.26 ±0.59	0.66 ±0.09	ab
	Northern Upland	44.69 ±0.09 ♦	44.22 ±0.19	a	45.64 ±0.14	45.87 ±0.07	a	0.96 ±0.05	0.98 ±0.03	cd	0.78 ±0.12	0.69 ±0.03	ab
Wild		43.95	43.60	α	45.26 *	45.00	α	1.35	1.19	α	0.97	0.74	α
Cultivated		44.30 ♦	43.79	α	45.89 *	44.72	β	1.09	1.08	β	0.91	0.70	α

Significant differences between M-/+ treatments for all varieties determined by a two-way ANOVA at α =0.05 with no interaction between myc and variety indicated by ♦.
Significant differences between M-/+ treatments determined by a t-test at α =0.05 indicated by *.
Significant differences between varieties as determined by a Tukey's test at α =0.05 indicated by different letters.
Significant differences between cultivation level as determined by a two-way ANOVA at α =0.05 indicated by α, β.

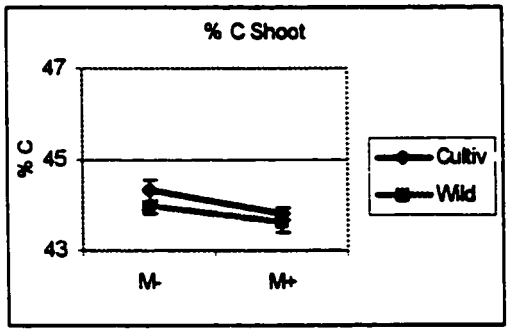


Fig 3.33. Carbon in shoot

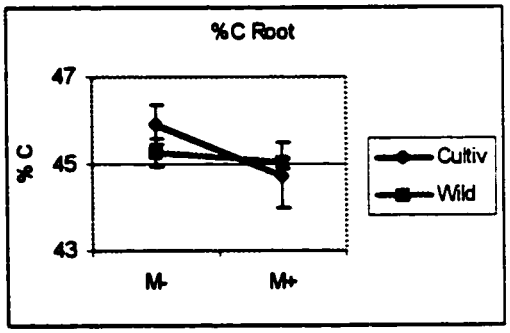


Fig 3.34. Carbon in root

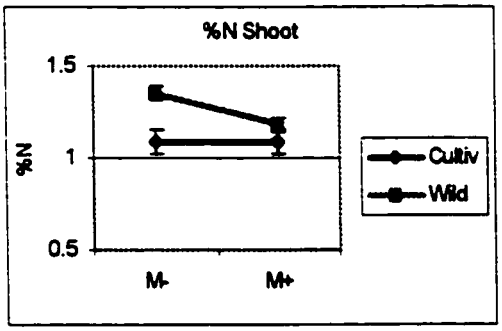


Fig 3.35. Nitrogen in shoot

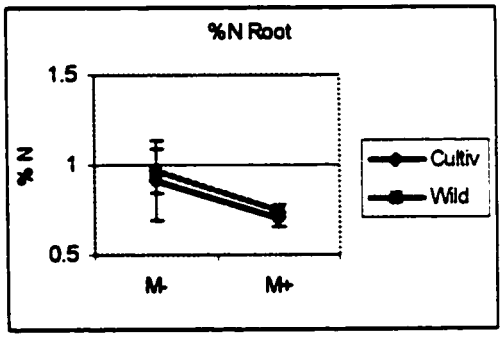


Fig 3.36. Nitrogen in root.

Figures 3.33-3.36. Total %C and %N in switchgrass of experiment 1 plants. Differences in cultivation level and mycorrhizal treatment are shown for carbon and nitrogen in roots and shoots. Standard error bars are shown. Line graphs are used to visualize the interaction between cultivation level and mycorrhizal treatment; it is not intended to suggest any continuous variable.

Table 3.11. Total percentage of Carbon and Nitrogen ± standard error for experiment 2 switchgrass plants.

	Variety	% C Shoot			% C Root			% N Shoot			%N Root		
		M-	M+		M-	M+		M-	M+		M-	M+	
Wild	Prairie Habitat	41.95 ±0.24	41.45 ±0.09	c	44.77 ±0.41	44.39 ±0.14	a	1.49 ±0.08	1.96 ±0.14	a	0.88 ±0.13	1.19 ±0.03	ab
	ONP	42.48 ±0.27	40.20 ±0.50	c	45.08 ±0.01	43.66 ±0.52	a	1.44 ±0.08	1.16 ±0.23	ab	0.93 ±0.04 *	1.18 ±0.10	ab
	Ojibway	43.57 ±0.28	43.45 ±0.62	ab	44.52 ±0.23	46.60 ±0.56	a	1.58 ±0.05	1.72 ±0.22	ab	0.92 ±0.04 *	1.49 ±0.05	a
Cultivated	Forestburg	41.72 ±0.46	43.06 ±0.46	b	45.31 ±0.07	44.27 ±0.12	a	1.24 ±0.05	1.03 ±0.06	b	0.65 ±0.03	0.75 ±0.03	c
	Caddo	43.26 ±0.05 *	44.39 ±0.20	a	44.64 ±0.35	43.85 ±0.38	a	1.12 ±0.02 *	1.76 ±0.05	ab	0.59 ±0.02 *	1.06 ±0.04	abc
	Northern Upland	42.48 ±0.18	41.55 ±0.33	bc	45.43 ±0.28	43.64 ±0.39	a	1.18 ±0.05	1.46 ±0.09	b	0.62 ±0.05	0.81 ±0.02	bc
Wild		42.67	41.70	α	44.79 *	44.88	α	1.50 *	1.61	α	0.91 *	1.29	α
Cultivated		42.49	43.00	α	45.12 *	43.92	α	1.18 *	1.42	β	0.62 *	0.87	β

Significant differences between M-/+ treatments determined by a t-test at α =0.05 indicated by *.
Significant differences between varieties as determined by a Tukey's test at α =0.05 indicated by different letters.
Significant differences between cultivation level as determined by a two-way ANOVA at α =0.05 indicated by α, β.

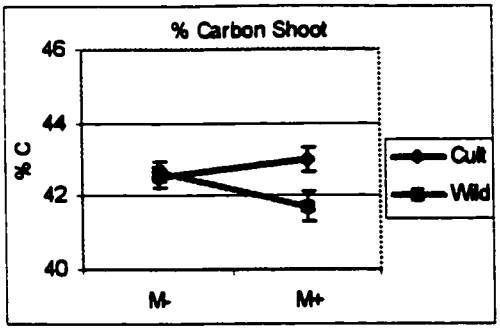


Fig 3.37. Carbon in shoots

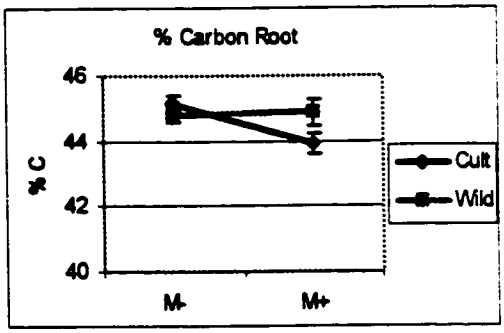


Fig 3.38. Carbon in roots.

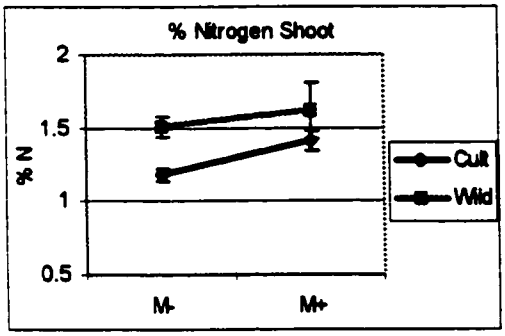


Fig 3.39. Nitrogen in shoots.

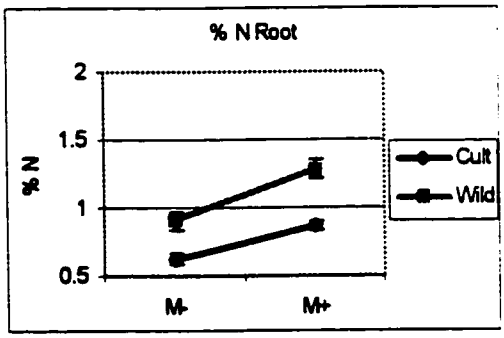


Fig 3.40. Nitrogen in roots.

Figures 3.37-3.40. Total %C and %N in switchgrass of experiment 2 plants. Differences in cultivation level and mycorrhizal treatment are shown for carbon and nitrogen in roots and shoots. Standard error bars are shown. Line graphs are used to visualize the interaction between cultivation level and mycorrhizal treatment; it is not intended to suggest any continuous variable.

Table 3.11.1. Results of two way ANOVAs at $\alpha = 0.05$ for C and N analysis of switchgrass roots and shoots. The first ANOVA was done for variety and mycorrhizal treatment, the second for cultivation level and mycorrhizal treatment.

	First ANOVA				Second ANOVA			
	R^2	V, p=	T, p=	V×T, p=	R^2	C, p=	T, p=	C×T, p=
Exp 1, n= 60								
%C Shoot *	0.632	0.001	0.000	0.340	0.239	0.050	0.001	0.648
% C Root *	0.671	0.000	0.005	0.015	0.153	0.033	0.046	0.513
% N Shoot	0.868	0.000	0.010	0.000	0.205	0.002	0.125	0.162
% N Root *	0.730	0.000	0.092	0.000	0.050	0.218	0.254	0.667
Exp 2, n= 36								
%C Shoot	0.799	0.000	0.079	0.006	0.174	0.103	0.185	0.142
% C Root*	0.684	0.029	0.002	0.003	0.240	0.247	0.014	0.175
% N Shoot	0.829	0.000	0.000	0.001	0.498	0.000	0.004	0.386
% N Root	0.921	0.000	0.000	0.000	0.752	0.000	0.000	0.091

Use of a Non-Parametric test indicated by *.
V= variety, T= mycorrhizal treatment, C= cultivation level.

Table 3.11.2. Results of two way ANOVA at $\alpha = 0.05$ for P, Ca, K and Mg analysis of experiment 2 switchgrass roots and shoots. The first ANOVA was done for variety and mycorrhizal treatment, the second for cultivation level and mycorrhizal treatment.

	First ANOVA				Second ANOVA			
	R^2	V, p=	T, p=	V×T, p=	R^2	C, p=	T, p=	C×T, p=
Exp 2, n= 36								
%P Shoot	0.841	0.000	0.569	0.000	0.150	0.033	0.774	0.431
% P Root	0.853	0.000	0.011	0.000	0.404	0.000	0.223	0.182
%Ca Shoot	0.815	0.000	0.097	0.047	0.418	0.000	0.269	0.070
% Ca Root	0.668	0.000	0.454	0.604	0.079	0.213	0.343	0.668
% K Shoot*	0.680	0.002	0.024	0.012	0.176	0.644	0.093	0.067
% K Root	0.732	0.000	0.340	0.023	0.018	0.629	0.561	0.969
% Mg Shoot	0.471	0.087	0.069	0.280	0.213	0.208	0.081	0.061
% Mg Root	0.426	0.131	0.170	0.310	0.080	0.960	0.207	0.295

Use of a Non-Parametric test indicated by *.
V= variety, T= mycorrhizal treatment, C= cultivation level

Table 3.12. Concentration of P and Ca ± standard error in experiment 2 switchgrass plants.

	Variety	P shoots (µg/g)			P roots (µg/g)			Ca shoots (µg/g)			Ca roots (µg/g)		
		M-	M+		M-	M+		M-	M+		M-	M+	
Wild	Prairie Habitat	1192 ± 24	909 ± 112	ab	961 ± 10 *	731 ± 54	a	2030 ± 64	2126 ± 239	a	921 ± 99	851 ± 25	c
	ONP	948 ± 45	909 ± 34	bc	710 ± 55	686 ± 22	ab	1393 ± 102 *	1815 ± 66	b	1011 ± 49	1144 ± 89	bc
	Ojibway	1055 ± 52	1189 ± 33	a	579 ± 13	898 ± 94	ab	1190 ± 41	1577 ± 150	bc	1247 ± 82	1287 ± 89	ab
Cultivated	Forestburg	930 ± 57 *	1243 ± 41	ab	556 ± 28 *	810 ± 40	b	1348 ± 101	1483 ± 97	bc	986 ± 162	1150 ± 61	bc
	Caddo	791 ± 46	706 27	d	481 ± 18	535 ± 40	c	1206 ± 85	1016 ± 67	c	1221 ± 168	1341 ± 26	a
	Northern Upland	951 ± 25	811 ± 42	cd	574 ± 27	586 ± 55	bc	1338 ± 65	1166 ± 138	bc	1135 ± 26	1123 ± 52	abc
Wild		1065 ± 40	1002 ± 60	α	750 ± 26	772 ± 57	α	1538 ± 69	1839 ± 152	α	1060 ± 77	1094 ± 68	α
Cultivated		891 ± 43	920 ± 36	β	537 ± 24	644 ± 45	β	1297 ± 84	1222 ± 101	β	1114 ± 118	1205 ± 46	α

Significant differences between M-/+ treatments determined by a t-test at α =0.05 indicated by *.
Significant differences between varieties as determined by a Tukey's test at α =0.05 indicated by different letters.
Significant differences between cultivation level as determined by a two-way ANOVA at α =0.05 indicated by α, β.

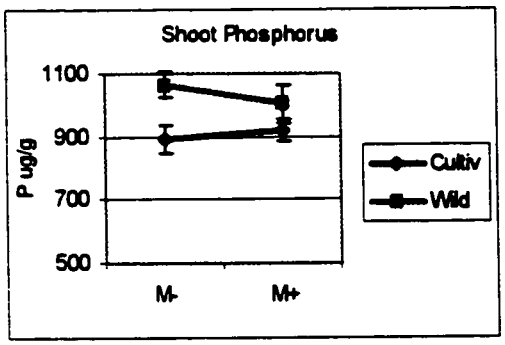


Fig 3.41. Shoot phosphorus

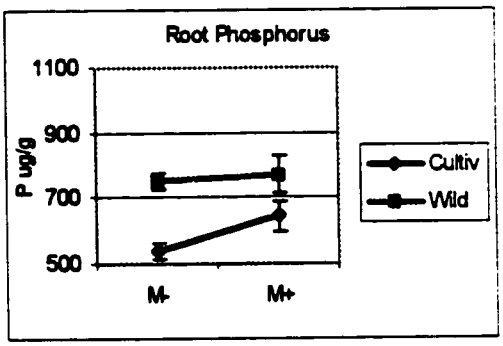


Fig 3.42. Root phosphorus

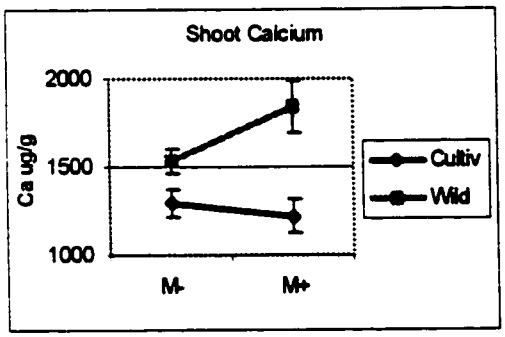


Fig 3.43. Shoot calcium

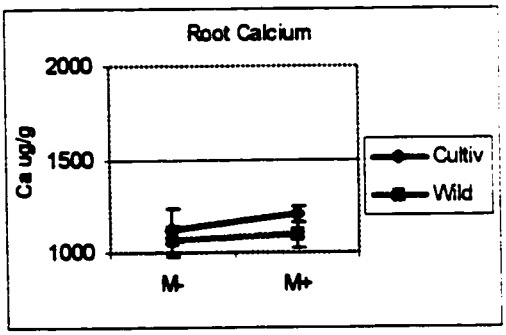


Fig 3.44. Root calcium

Figures 3.41-3.44. Concentration of phosphorus and calcium of switchgrass from experiment 2 plants. Differences in cultivation level and mycorrhizal treatment are shown for P and Ca in shoots and roots. Standard error bars are shown. Line graphs are used to visualize the interaction between cultivation level and mycorrhizal treatment; it is not intended to suggest any continuous variable.

Table 3.13. Concentration of K and Mg ± standard error in experiment 2 switchgrass plants.

	Variety	K shoots (mg/g)			K roots (mg/g)			Mg shoots (μg/g)			Mg roots (μg/g)		
		M-	M+		M-	M+		M-	M+		M-	M+	
Wild	Prairie Habitat	16.26 ± 0.73	14.63 ± 0.44	b	18.70 ± 0.95	19.55 ± 0.53	b	2018 ± 65	2258 ± 220	a	4170 ± 150	4425 ± 358	a
	ONP	18.47 ± 0.43 *	21.82 ± 0.54	a	22.95 ± 0.21	24.46 ± 1.12	a	2230 218	2800 334	a	3272 ± 152	3552 ± 235	a
	Ojibway	21.44 ± 0.53	21.63 ±0.41	a	25.27 ± 0.47	21.11 ± 2.51	a	1890 ±58	2295 ± 152	a	4555 ± 721	4171 ± 579	a
Cultivated	Forestburg	19.03 ± 1.15	18.15 ± 1.45	ab	20.31 ± 0.99	17.16 ± 0.42	b	1999 ± 132	2201 ± 182	a	3627 ± 365	4115 ± 362	a
	Caddo	21.49 ±0. 45 *	14.35 ± 1.35	ab	21.45 ± 0.29 *	23.31 ± 0.36	a	1978 ± 167	1980 ± 28	a	3377 ± 271	4633 ± 164	a
	Northern Upland	23.11 ± 0.68	19.72 ± 2.51	a	23.17 ± 0.85	23.41 ± 1.25	a	2375 ± 186	2125 ± 159	a	4236 ± 97	4089 ± 275	a
Wild		18.72 ± 0.56	19.36 ± 1.71	α	22.31 ±0. 55	21.70 ± 1.39	α	2046 ± 114	2451 ± 235	α	3999 ± 341	4050 ± 391	α
Cultivated		21.21 ± 0.76	17.41 ± 1.77	α	21.65 ± 0.71	21.29 ±0.68	α	2117 ± 161	2101 ± 123	α	3746 ± 244	4279 ± 267	α

Significant differences between M-/+ treatments determined by a t-test at α =0.05 indicated by *.
Significant differences between varieties as determined by a Tukey's test at α =0.05 indicated by different letters.
Significant differences between cultivation level as determined by a two-way ANOVA at α =0.05 indicated by α, β.

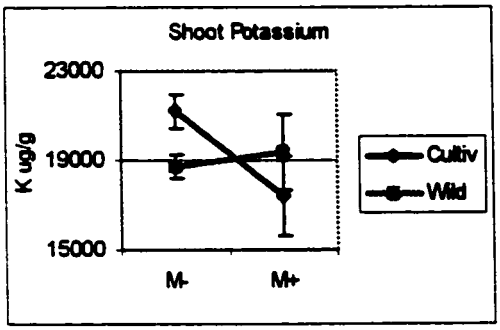


Fig 3.45. Shoot potassium.

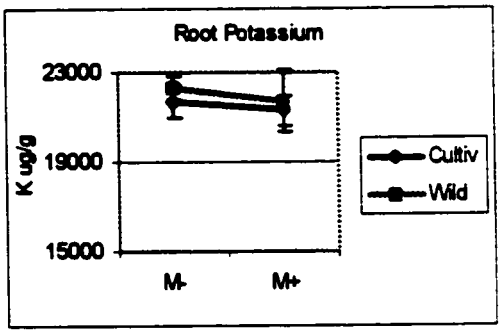


Fig 3.46. Root potassium

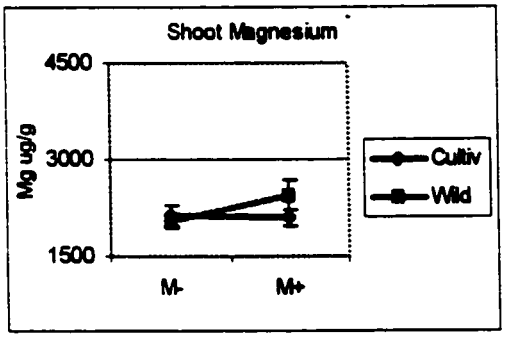


Fig 3.47. Shoot magnesium.

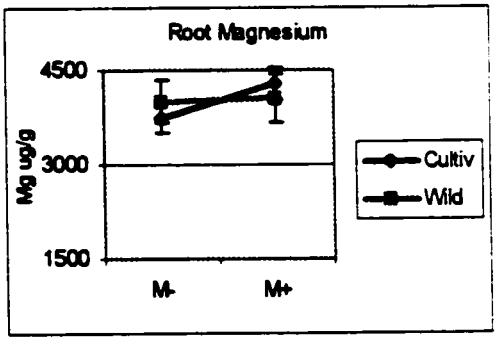


Fig 3.48. Root magnesium.

Figures 3.45-3.48. Concentration of potassium and magnesium of switchgrass from experiment 2 plants. Differences in cultivation level and mycorrhizal treatment are shown for K and Mg in shoots and roots. Standard error bars are shown. Line graphs are used to visualize the interaction between cultivation level and mycorrhizal treatment; it is not intended to suggest any continuous variable.

10.0 – Statistical Analysis

10.1- Cluster Analysis

First Experiment

Two main clusters separate the varieties in the first experiment (Fig 3.49). Clusters seem to separate the varieties that responded positively to the mycorrhizal treatment, from varieties that responded negatively. Cluster 1, the positive responders, includes NU, Summer, Forestburg and ONP. Cluster 2 includes PH, Shelter, Ojibway, Trailblazer as well as Pterophylla and Caddo, which are closely joined to each other and separate from the other varieties in this group. More varieties had negative than positive responses in this experiment. Both Forestburg and ONP, which were closely joined, had positive responses to mycorrhizae almost consistently. Northern Upland and Summer, the other varieties included in this cluster, had positive responses for some of the variables and not for others. Positive effects were found in height, fresh mass, a decrease in root diameter, and an increase in % N shoot that were greater than those of the other varieties. Of the negatively responding group, Caddo and Pterophylla had negative responses in mass, increases in root diameter and a drop in acid phosphatase level, which set them apart from the others. The remaining varieties were very closely joined to each other, and in other types of joining their positions switched. Trailblazer, Shelter, Ojibway and Prairie Habitat had large negative responses in fresh mass, and varied responses to all the other factors. There were no distinct groupings between the wild and cultivated varieties.

Second Experiment

The two main clusters were separated seemingly on root characteristics for the cluster analysis on the second experiment (Fig 3.50). Although two out of the three from the wild and cultivated groups clustered closely together, Forestburg grouped with the wild types and Ojibway with the cultivars. All three of the first cluster, Forestburg, ONP and PH have significantly higher root:shoot ratios than the others. Differences between clusters in effect size were noticed for fresh weight where group 1 had high positive values while the others were negative. No other obvious segregations were noticed.

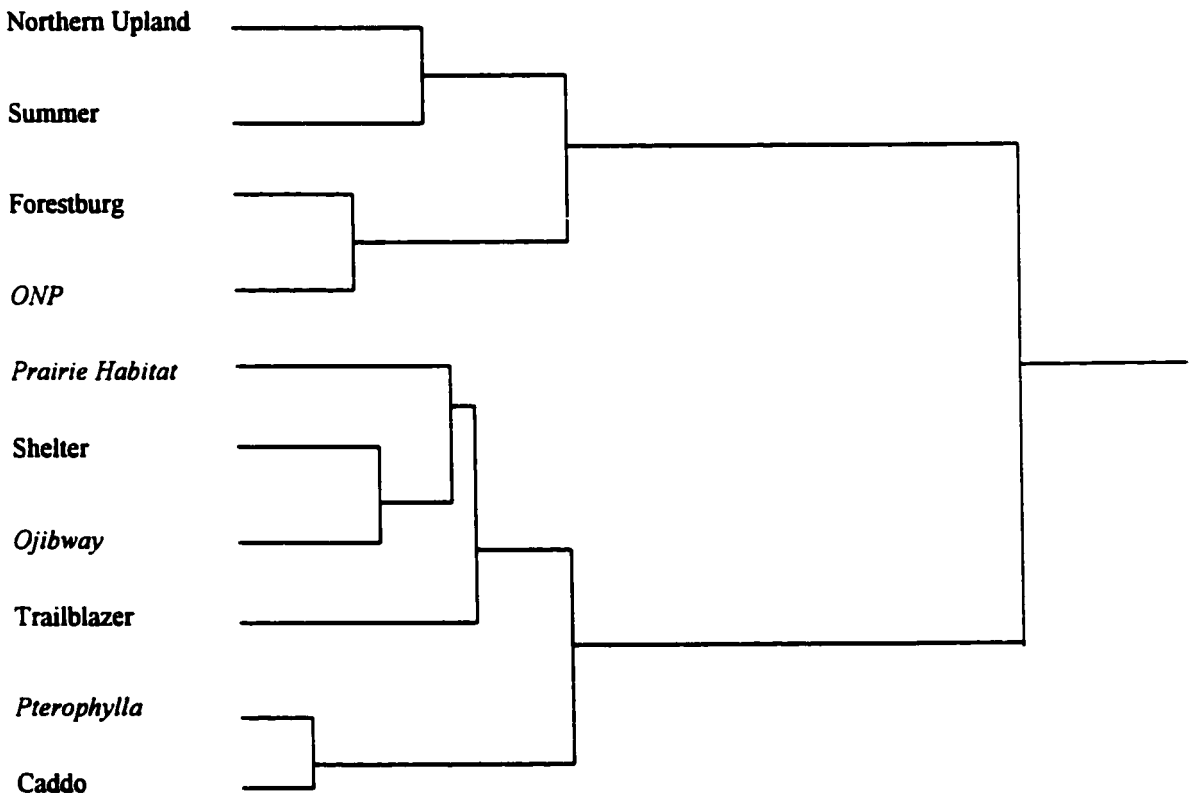


Figure 3.49. Cluster analysis of the mycorrhizal treatment effect size on wild and cultivated varieties of switchgrass for the first experiment plants using Ward joining. Wild varieties are indicated in italics.

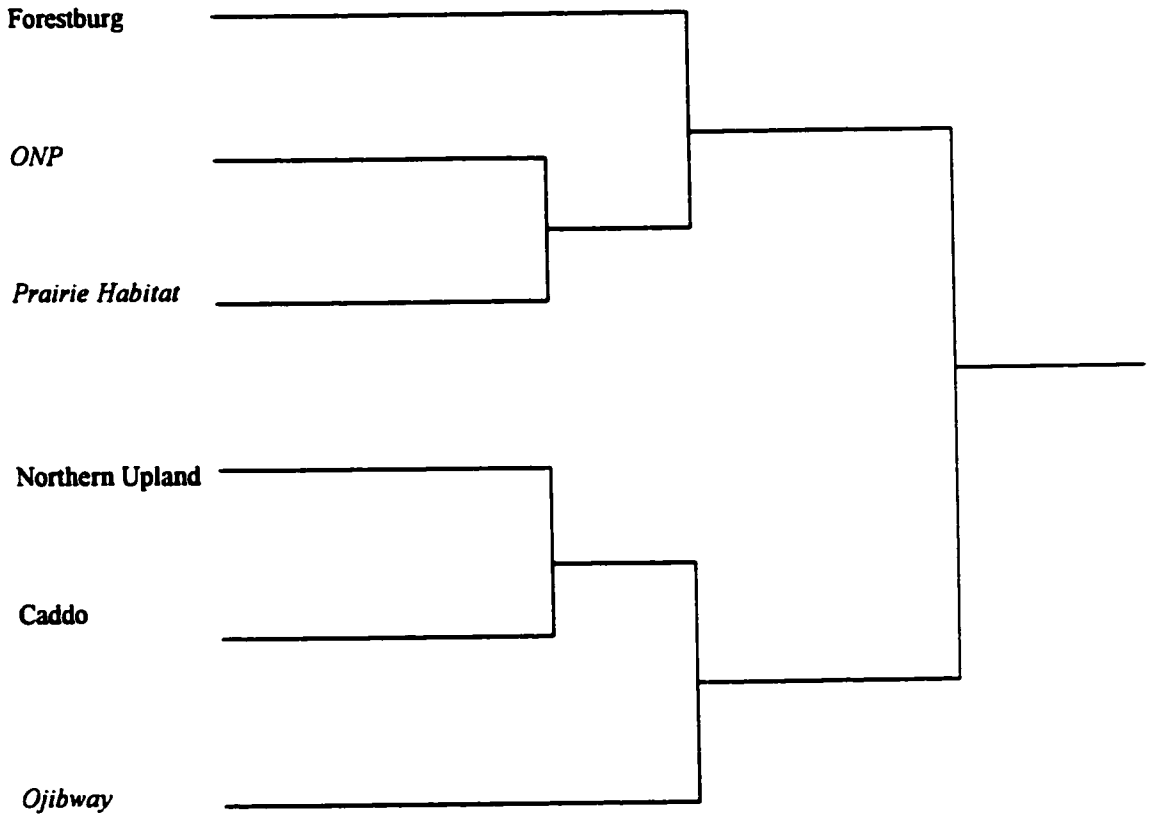


Figure 3.50. Cluster analysis of the mycorrhizal treatment effect size on wild and cultivated varieties of switchgrass for the second experiment plants using Ward joining. Wild varieties are indicated in italics.

10.2- Canonical Discriminant Analysis

Experiment 1

To assess the validity of the patterns found in the cluster analysis, a canonical discriminant analysis (CDA) was performed. Three clusters were identified for this analysis. Cluster 1 included NU, Summer, Forestburg and ONP; cluster 2 included PH, Shelter, Ojibway and Trailblazer; cluster 3 included Pterophylla and Caddo (Fig 3.51). The clusters were separated in this manner because two clusters (cluster 1 and clusters 2&3) had a lower predictive rate than 3 clusters. Stepwise discriminant analysis was done to determine which factors were the most important to cluster grouping. Results from the stepwise CDA showed that 4 variables were the best predictors of cluster grouping. The most important predictors of clustering by effect size were fresh mass, acid phosphatase level, average root diameter, and % of small roots (Table 3.14). The clusters were significantly different from each other using Wilk's lambda, while predictive ability using jackknifed classification, as well as eigenvalues are very good in this analysis (Table 3.15).

Experiment 2

CDA was performed to determine validity of the clusters in experiment 2. Cluster 1 was comprised of Forestburg, ONP and PH; cluster 2 of NU, Caddo and Ojibway. From the forward stepwise CDA (F-to-join values), it was determined that two variables were important in determining clusters. The most important factors were determined to be fresh mass and root length (Table 3.16). Cluster groups were significantly different using Wilk's lambda, predictive values (jackknifed classification %) were good but the eigenvalues were lower in this experiment than in the first because of the smaller sample size (Table 3.17). Only one canonical axis was generated since only two clusters were compared. Canonical scores grouped by cluster are displayed in figure 3.52.

Table 3.14. Mean effect size values for each cluster in experiment 1.

	Cluster 1	Cluster 2	Cluster 3
Fresh Mass	4.426	-10.598	-11.956
Acid Phosphatase	-2.030	-2.159	-4.386
Average Diameter	-3.587	-1.984	2.580
% Small Roots	-0.213	1.823	1.350

Table 3.15. Statistics from canonical discriminant analysis of experiment 1 clusters.

Statistic		Value
Wilks Lambda		p = 0.00000
F-to-Remove	Fresh Mass	144.87
	Acid Phosphatase	90.75
	Average Diameter	10.68
	% Small Roots	4.94
Jackknifed Classification		100%
Eigenvalue		342.39
Canonical Correlation		0.999

Table 3.16. Mean effect size value for each cluster in experiment 2.

	Cluster 1	Cluster 2
Fresh Mass	8.151	-5.869
Root Length	2.775	0.238

Table 3.17. Statistics from canonical discriminant analysis of experiment 2 clusters.

Statistic		Value
Wilks Lambda		p = 0.0021
F-to-Remove	Fresh mass	113.65
	Root Length	8.56
Jackknifed Classification		100%
Eigenvalue		59.31
Canonical Correlation		0.992

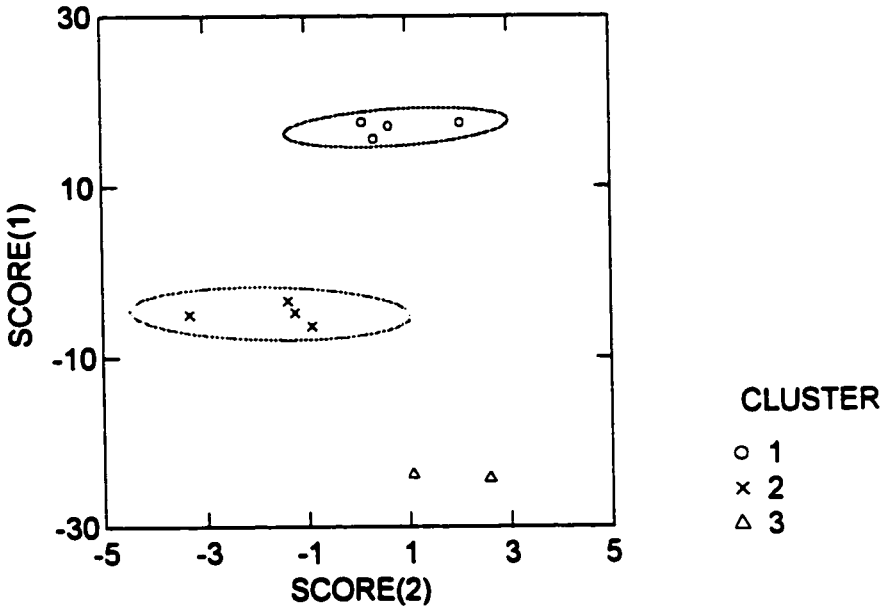


Figure 3.51. Canonical scores of clusters as determined by canonical discriminant analysis of experiment 1 switchgrass plants. Cluster groups are surrounded by p= 0.60 confidence ellipses where statistically possible.

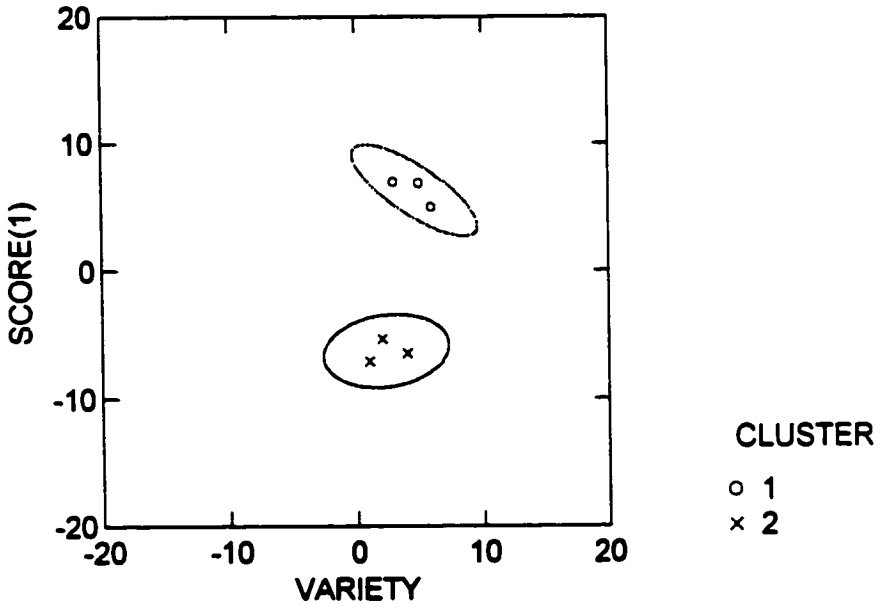


Figure 3.52. Canonical scores of cluster groups as determined by canonical discriminant analysis of experiment 2 switchgrass plants. Cluster groups are surrounded by p= 0.60 confidence ellipses.

Discussion

1.0- Switchgrass Varieties

1.1- Physiological Differences Among Varieties

The cultivars of switchgrass showed some signs of genetic selection for height and mass. All the cultivars were larger than their wild counterparts. Cultivars NU and Caddo were very tall, sparsely branched and had a high biomass. There was more variation among the wild varieties in their physiology. The wild variety PH grew to a low height but was highly branched and produced many tillers. In contrast, Ojibway was mid height and slightly branched, while ONP, a wild variety, was tall and similar to cultivars. Other researchers have also observed variation in appearance, tillering, branching patterns, forage quality, and other physiological characteristics among switchgrass varieties or populations (Nielsen, 1944; Eberhart and Newell, 1959; Hopkins et al., 1995; Madakadze et al, 1998b).

The plants in this greenhouse study did not grow to the usual height of switchgrass, and they also started to flower early. This stunting is most likely caused by growth stress from being in a small soil volume and nutrient stress caused by the low fertilization level.

The height of the tallest leaf was found to be an adequate measure of growth for grass plants in this experiment and in others (Miller et al., 1987). Plants experienced linear growth over time, for the most part. Fluctuations in the growth rate tend to coincide with fertilizer inputs; increases were noted after fertilization. After an 8-week growth period,

many varieties seemed to experience a plateau. The height of cultivars Forestburg and Summer, and wild varieties ONP and PH noticeably level off. It is possible that at this point, their root systems became limited by the cones in which they were grown. Interestingly, the varieties that had the longest root lengths were not the ones that experienced the growth plateau. The varieties that levelled off in growth had in common the fact that they generally benefited from mycorrhizal colonization. Growth limitation could be due to the soil volume being fully exploited by the hyphal network. The growth rate of cultivars was higher than that of wild varieties. Faster growth of cultivars may reduce their potential to experience biomass and height benefits from mycorrhizal colonization, since most plants experience a growth depression when first colonized (Smith and Read, 1997).

Chlorophyll levels of the first experiment plants did not differ much although significant differences were found. A large sample size can reduce the standard error and give statistically significant results when the actual differences are minimal in a biological perspective. The slightly higher values seen in wild plants can be explained since higher chlorophyll concentrations were found in plants with lesser biomass. Resources such as chlorophyll can become diluted in plants that have a large volume (Madakadze et al., 1999). Forestburg, which had the highest chlorophyll concentration, does not fit this assumption as it had a fairly large biomass. The chlorophyll meter readings were lower than those found in field stands of switchgrass fertilized with high N rates (Madakadze et al., 1999b). The lack of response in chlorophyll levels to mycorrhizal colonization was also found in turfgrasses (Charest et al.,1997) and maize (Subramanian and Charest,

1995). Other researchers have found increases in chlorophyll levels when plants are mycorrhizal (Allen et al., 1981).

1.2- Chromosome Numbers

The ploidy of switchgrass varieties found in this study in most cases did not concur with the amount of chromosomes expected. Other studies have provided conflicting reports on the ploidy of several of the cultivars used in the present study. The chromosome numbers of varieties that are either hexaploid or octoploid are very difficult to distinguish in dividing root tips because of their small size. In this study, the wild variety ONP actually had been counted with 2 ploidy levels, 54 and 72. It has been suggested that octoploid varieties may experience 'stickiness' during mitosis that makes these chromosomes even more difficult to count (Dr. Ken Vogel, pers. comm.). Contrary to expectations (McMillan and Weiler, 1959), the wild varieties in this study were all found to be hexaploid and not tetraploid. There may be several reasons for this discrepancy, the ploidy of the plants might have changed since 1959 or Canadian varieties, which were not included by McMillan and Weiler, are different ecotypes or the wild varieties might have interbred with cultivated types. Ploidy level does not seem to correspond with any physiological character such as height, biomass or tillering in this study.

2.0- Mycorrhizal Response

2.1- Colonization

The levels of colonization of switchgrass roots with *Glomus intraradices* found in this study did not vary significantly among varieties or between the two greenhouse experiments. The average colonization level of 36% corresponds to the levels found for switchgrass in other studies, under both greenhouse and field conditions (Hetrick and Bloom, 1983; Brejda et al., 1993; Johnson, 1998). It is possible that ~ 40% is the threshold colonization level for switchgrass.

Various studies on the responsiveness of switchgrass to AM fungi have shown that a certain level of fungus-plant compatibility may occur. Different AM fungal species were found to vary widely in their relative spore numbers when associated with switchgrass (Hetrick and Bloom, 1983). This suggests that some AM fungal species may survive better than others when associated with switchgrass in the wild. Species diversity of AM fungal spores was found to be low in the vicinity of switchgrass plants in comparison with the diversity under other tallgrass prairie species (Eom et al., 2000). This suggests that switchgrass may be a more compatible host for some AM fungi than others. Plant response to AM fungi varies as well, suggesting that the host plant also has some preference in its symbiotic partner. Brejda et al. (1993) found that an unspecific inoculum that contained many species of AM fungi produced more benefits to switchgrass than did a pure isolate of *Glomus deserticola*. When switchgrass was grown with several separate

species of mycorrhizal fungi, *Glomus clarum* and *Glomus diaphanum* were the most effective species at increasing plant biomass while *Glomus intraradices* was less compatible (Clark et al., 1999; Eom et al., 2000). However, positive growth responses were observed in switchgrass plants colonized with *G. intraradices* when compared to control plants (Clark et al., 1999). It is possible that the lack of many significant mycorrhizal responses in this study was due to a relatively low symbiotic compatibility of *Panicum virgatum* with *Glomus intraradices*.

It has been suggested that plants with higher ploidy levels may have higher mycorrhizal colonization levels. In the present study, no differences in colonization percentage were found among the varieties, regardless of their ploidy. Differences between cultivars of various ploidies were found by Jun and Allen (1991), but no clear relationship between ploidy and colonization was found. Kapulnik and Kushnir (1991) found that polyploid wheat had higher % colonization than diploid wheat; however this was not related to mycorrhizal efficiency. Berta et al. (2000) found that tomato cultivars with higher ploidies had higher colonization levels.

2.2- Physiological Responses to Mycorrhizae

Greenhouse growth conditions were important in determining plant response to mycorrhizal colonization. In the first experiment, where soil volume was limiting, a negative response was seen for almost all the varieties. Therefore, in the first experiment of this study, the carbon cost is likely to have exceeded the benefits of AM fungal symbiosis. The soil volume was increased in the second experiment, and the responses to

colonization became either neutral or positive. Baath and Hayman (1984) also noted that mycorrhizal symbiosis in onion (*Allium cepa* L.) with both *Glomus mosseae* and *G. caledonia* went from beneficial to antagonistic when soil volume was reduced. In small soil volumes it is likely that the root system of the plant is sufficient for mineral absorption.

Biomass is often used as the benchmark of mycorrhizal response. In this study, mass differences due to mycorrhizae were not as dramatic as others reported in the literature (reviewed in Smith and Read, 1997). Biomass decreases observed in the first experiment were not unexpected, when the stressful growth conditions are taken into account. The responses of the host plants seen in the second experiment are more reliable than those of the first experiment; therefore, the second experiment plants are considered more closely. The cultivated group had a larger biomass response to mycorrhizal colonization, mainly because of growth increases in Forestburg. Dry mass decreases in the wild varieties PH and Ojibway, the two smallest varieties, also contributed to this difference. A similar study by Clark et al. (1999) found the total dry mass of switchgrass decreased when colonized by *Glomus intraradices* at pH 4; the present study was performed at pH 4.5. However, most studies with switchgrass reported biomass increases when mycorrhizal (Hetrick et al., 1988; Bredja et al., 1998; Johnson, 1998; Entry et al., 1999).

Root to shoot ratio is a good measure of plant health, when compared within a species, and life history, when compared between species. A large R/S indicates a higher investment in underground structures for resource gathering. The R/S usually increases if

a plant is mineral or water stressed (Hopkins, 1995). The average R/S of near 1.0 in the first greenhouse experiment indicates the stressed condition of these plants. The R/S decreased to 0.5 in the second experiment. In both experiments of this study, the soil pH was 4.5. The R/S in the second experiment is similar to what was found in the study of Clark et al. (1999), where switchgrass colonized by *G. intraradices* had R/S ratios of 0.58 in soil at pH 4 and 0.35 at pH 5.

Since the hyphal network serves the same function as small roots, fewer roots need to be produced and the R/S should decrease in mycorrhizal plants. The R/S was found to be higher in mycorrhizal plants of the first experiment of this study. This is likely because of increased stress level and C transferred to AM fungi, making the shoots grow more slowly (Tinker et al., 1994). This R/S increase was more pronounced in wild plants, which already invested more into root growth. The wild varieties continued to allocate more resources for root growth, but both wild and cultivated groups maintained a steady R/S under M+ and M- treatments in the second experiment. Bredja et al. (1993) found that switchgrass responded to mycorrhizal colonization by decreasing R/S. This trend has been observed in many species such as onion (Baath and Hayman, 1984), wheat (Vierheilig and Ocampo, 1991), *Salix viminalis* L. (Tinker et al., 1994), and lettuce (Koide et al., 2000). Cultivated varieties need to invest fewer resources towards root growth since they are less likely to experience nutrient or drought stress than wild plants. Lower R/S in cultivars may be a sign that agricultural selection has changed the physiology of switchgrass. Shoot biomass is a characteristic selected for in breeding programs, so it is not unexpected that shoots are more emphasized in the cultivated

group. It is interesting to note that Forestburg, a cultivar that has not undergone selection, maintains a high R/S while Ojibway, a wild variety, has a fairly low R/S.

In grasses, growth can occur either vertically through the stem and leaves, or laterally via tillers. While height was consistently greater in the cultivated plants, tiller numbers were not different between the cultivation levels when non-mycorrhizal. An interesting phenomenon was found in this study: a significant decrease in tiller numbers occurred in mycorrhizal plants of both experiments and cultivation levels. The only variety unaffected by this decline was the cultivar Forestburg, which seems to compensate by having a lower height when mycorrhizal. However, no correlation between height and tiller number was found in this study. A lack of vegetative reproduction could indicate that the plants were stressed by the growth conditions. Our results are contrary to those of Bredja et al. (1993) who observed increased tiller numbers in mycorrhizal switchgrass. Mycorrhizae increased tiller numbers but decreased height of the prairie grass *Agropyron smithii* L., while an overall mass increase was noted (Miller et al., 1987). Overall, mycorrhizae did not affect the height of plants in the second experiment. Height increases due to mycorrhizal colonization were observed in switchgrass by Johnson (1998).

In the first experiment, under stress conditions, the wild varieties experienced smaller declines in mass, height and tillers than the cultivars when mycorrhizal. ONP had the effect of buffering the detrimental mycorrhizal effects in the collective results of the wild varieties, since Ojibway and Pterophylla both had significant declines for most of the physiological parameters examined. Cultivars Caddo, Summer and Trailblazer declined

significantly for almost every physiological parameter when colonized by mycorrhizae. In the second experiment, under less limiting growth conditions, the wild variety ONP and the non-selected cultivar Forestburg benefited from mycorrhizal colonization while the other varieties had neutral or negative responses. The agriculturally selected cultivars, Caddo and NU, both had declining mass and tiller numbers, suggesting negative physiological responses of cultivars to mycorrhizal colonization.

2.3- Root Architecture Responses to Mycorrhizae

Since mycorrhizal hyphae are in some ways a replacement for root hairs, root architecture is an important factor in determining mycorrhizal response (Baylis, 1970; Manjunath and Habte, 1991). As well, mycorrhizal colonization can induce changes in roots that may determine the symbiotic efficiency. Since mycorrhizal symbiosis seems to be the ancestral condition in plants, root morphology that lowers mycorrhizal dependency suggests an environmental adaptation. Special physiological adaptations such as bulbous root hairs were related to the non-mycorrhizal state of some Carex species (Miller et al., 1999). Grasses have adventitious root systems, which are highly branched and have small average diameters. Because the existing root systems should be able to adequately exploit the soil volume without the aid of the mycorrhizal hyphal network, mycorrhizae have been theorized to be neutral or commensal symbionts for grasses (Francis and Read, 1995). Nonetheless, many grass species, especially C4 grasses including *Andropogon gerardii* Vitm., *Sorghastrum nutans* L., and *Bouteloua cutipendula* Michx., are highly mycorrhizal dependent (Hetrick et al., 1988,1990 &1992).

An increase in total root length was observed in the mycorrhizal plants in both experiments. Root length increases due to AM mycorrhizal fungi have also been observed in other studies with switchgrass (Hetrick et al., 1991; Clark et al., 1999), and with *Plantago lanceolata* (Hodge et al., 2000). Schweiger et al. (1995) found no differences in either root length or diameter between mycorrhizal and non-mycorrhizal treatments in some pasture species. It is possible that increased root length occurs because resources no longer need to be allocated to root hairs and branching.

In this study, large differences in root diameter were detected between varieties in the first experiment, while the root diameters were smaller and fairly similar among varieties in the second experiment. Cultivars were found to have larger root diameters than wild varieties in both experiments. Since plants with coarser roots would find functional complementarity with the fine mycorrhizal hyphae, it is predicted that they should benefit most from the symbiosis. Fine roots and mycorrhizae were suggested to be alternative methods of acquiring minerals (Gahoonia et al., 1999), with the former being at a selective advantage under agricultural conditions. Switchgrass cultivars do not seem to be lowering their mycorrhizal dependency by producing finer diameter roots. Root diameter did not affect the % colonization of roots in this study, in contrast to the study of Reinhardt and Miller (1990) where mycorrhizal colonization decreased as average root diameter decreased. The average root diameter should theoretically increase when plants are colonized by AM fungi, since small diameter absorptive structures would be redundant. In this study, root diameter only increased in the mycorrhizal plants of the cultivars in the second greenhouse experiment. Root diameter increases in mycorrhizal

plants were found in several warm season grasses including switchgrass, tall grama-grass, indiangrass, big bluestem, and little bluestem Hetrick et al. (1991). Kaldorf and Ludwig-Muller (2000) found that smaller diameter roots increased in maize plants colonized by *Glomus intraradices*, possibly because of the effect of a growth hormone (indole-3-butyric acid) and its derivatives.

Root tip number, as a measurement of branching, was not as good a determinant of mycorrhizal effect as expected. Increased root branching in the mycorrhizal plants was noted in the first experiment and no differences in root tip numbers were observed in the second. It is likely that sub-optimal growth conditions in the first experiment caused stress, which induced increased branching. However, Hetrick et al. (1991) found that mycorrhizae suppressed root branching in C4 grasses. The determination of % large roots did not reveal any interesting results. There was a very small amount of roots with large diameters, and no clear patterns emerged from this variable. The % of small roots is more interesting. Schweiger et al. (1995) found that the length of root hairs was more important in determining mycorrhizal effect than root length, diameter and R/S. Graminaceous plants have adventitious root systems; therefore it is expected that most of the roots would have small to medium root diameters. As this category of roots would be directly replaced by mycorrhizal hyphae, their % of the total root length should decrease when colonized. Also, when plants are stressed they tend to produce more root hairs and small diameter roots. The increase in % small roots seen in some of the first experiment varieties is most probably due to a stress response. The response in small root production turns out to be a determining factor in the clustering of the first experiment varieties. In

the second experiment, when mycorrhizae were more effective, an overall decrease in the % of small roots was measured for all the varieties.

Root architecture has been used as a predictor of mycorrhizal dependency in many studies (Manjunath and Habte, 1991; Khalil et al., 1994). Criteria for mycorrhizal dependency list low R/S, little branching, large root diameter and root plasticity as factors that should indicate high responsiveness (Smith, 2000). Root plasticity refers to the variation in root morphology of a species or genotype and its ability to adapt its physiological parameters, such as root length, branching or diameter, to specific environmental conditions. In this study, wild varieties seemed to show more plasticity in their root morphological responses to mycorrhizal colonization. According to the scheme of Smith (2000), the cultivar Caddo should be the most mycorrhizal responsive variety, but this was not the case. The cultivar Forestburg had the highest MD, it also had high R/S, a large root length, and large diameter.

2.4- Mineral Responses to Mycorrhizae

Average mineral concentrations (μg/g) found in this study were P- 970 shoot, 676 root, Ca –1474 shoot, 1118 root, Mg –2179 shoot, 4018 root, K –19,175 shoot, 21,738 root. Mineral concentrations in μg/g for mycorrhizal switchgrass shoots at pH 4-5 were reported to be: P 100-1500, Ca 1000-8000, Mg 1000-1500, K 3000-25 000 (Clark et al., 1999). Concentrations are within the normal range for all minerals except for high

magnesium contents. High magnesium contents (890-3370 μg/g) were also noted in switchgrass under fertilized field conditions (Jung et al., 1988).

Varietal differences were found for all minerals except for magnesium. Highest mineral levels were found in wild varieties, possibly because of their smaller biomass. Caddo continuously had the lowest mineral content, as well as showing increases when mycorrhizal. This could indicate that it has low mineral requirements.

Plant mineral content is expected to increase with the addition of mycorrhizal fungi to the soil microflora (Gianinazzi-Pearson and Gianinazzi, 1983; Smith and Read, 1997). The effect is most pronounced when minerals are sparse in the soil environment. The plants in both experiments of this study were put on a low nutrient fertilizer regime and so were reasonably mineral stressed in order to emphasize the mycorrhizal effect. The P levels measured in this study were lower than those reported for switchgrass under field conditions (Jung et al., 1988). The only variety to experience an increase in P content with mycorrhizae was Forestburg, while PH experienced a decrease in P. Levels of K increased in ONP and Caddo, and Ca increased in ONP. Overall, mycorrhizae did not seem to increase uptake of minerals to any great extent. A study using the switchgrass cultivar 'Cave-in-Rock' found that P, K and Cu increased when *Glomus intraradices* was introduced (Clark et al., 1999). Johnson (1998) found P levels increased in switchgrass inoculated with three *Glomus* species (including *G. intratradices*); in addition P content was correlated with mass and height, effects not detected in this study. Increased mineral

concentrations were also observed in switchgrass colonized by *G. intraradices* in the study of Entry et al. (1999).

Concentrations of N in switchgrass are often reported to be around 10 g/kg or 1% by weight (Staley et al., 1991; Madakadze et al., 1999). Lower N concentrations of 5.1-9.2 g/kg or higher concentrations of 5.4-20.2 g/kg may also be found, depending upon variety and N availability in the soil (Madakadze et al., 1999c). In the field, N uptake tended to plateau early when fertilizer application was very high (Staley et al., 1991). It is common to find that the % N and other minerals decrease throughout the growing season as they are diluted through the plant tissue (Talbert et al., 1983; Balasko et al., 1984; Madakadze et al., 1999). A decrease in crude protein corresponding to that of N, was reported by Anderson and Matches (1983).

Nitrogen concentrations in this study (0.58-1.96%) were comparable to field values despite the low fertilization regime. The highest N concentration was found in the wild variety PH, which had the smallest mass. This result may correspond with the dilution effect mentioned by Madakadze et al. (1999). The mycorrhizal treatment tended to have opposite effects on % N in the two experiments. A decrease was found in the first experiment while an increase was observed in the second. The model proposed by Bago et al. (2001) on N translocation from mycorrhizal fungi to roots leaves allowance for N to move back into the fungus, and tends to link N movement to C flux. If the fungus was stressed for either C or N in the first experiment, translocation of both would have

decreased, which was observed. Increases in N with mycorrhizae in the second experiment were significant for the wild varieties ONP, Ojibway and the cultivar Caddo.

Wild varieties had higher acid phosphatase levels, except for the non-mycorrhizal plants of the first experiment; high levels in cultivars Shelter and Trailblazer skew the results of this treatment. This result is not surprising as it is more likely that wild plants would need to convert inorganic P from rock phosphorus than plants that receive PO_4 in fertilizer. No clear pattern emerges between the MD of plants and acid phosphatase levels in this study. The correlation between acid phosphatase level and P concentration was poor (non-significant), even though acid phosphatase should increase P uptake and therefore internal concentration. The contribution of mycorrhizal fungi to acid phosphatase levels has been found to depend on the AM fungal species (Boddington and Dodd, 1998). Acid phosphatase levels in this study were found either to decrease in the mycorrhizal treatment or to remain constant. Other studies attempting to determine mycorrhizal contribution to acid phosphatase have come up with contradictory or non-significant results (Gianinazzi-Pearson and Gianinazzi, 1976; Rubio et al., 1990). Acid phosphatase has also been reported to increase up to 60% with the addition of mycorrhizae (Tarafdar and Marschner, 1994; Khalil et al., 1999). Decreased production of acid phosphatase has been linked to higher phosphorus use efficiency (McLauchlan, 1980) and increases indicate a deficiency of P in the plant (Marschener, 1998). In this study, wild varieties of switchgrass had higher levels of acid phosphatase than the cultivars. This is in contrast to the findings of Khalil et al. (1994) where wild varieties of corn and soja had lower acid phosphatase levels and higher MD.

2.5- Cost/Benefit Analysis

Carbon content in plants can be used as a measure of their energetic status. As mycorrhizae are acting as a carbon sink for the plant, the amount of shoot carbon may decrease. It is known that mycorrhizae induce the plant to divert C resources from the shoots to the roots after colonization (Douds et al., 1988; Graham et al., 1997). Instead of the expected increase as C is diverted to the roots, root C levels decreased with in the mycorrhizal treatment in the present study. A decrease in % C was also noted by Entry et al., (1999) in switchgrass colonized with *G. intraradices*, but this did not deter biomass increases in the plant. It is possible that we are not detecting the entire carbon pool available to the roots and fungi, as the hyphal network could not be analyzed. The cost of the symbiosis may be exceeding the benefits to the plant in this particular case, even though the % colonization observed was not excessive. Carbon limitation, due to a large carbon sink or low production of photosynthate for example, is one of the conditions that can cause the symbiosis to be antagonistic to the plant (Johnson et al., 1997; Smith, 2000).

ONP, a wild variety, and Caddo, a cultivar, seem to have overall improved mineral uptake when mycorrhizal. Caddo did not show any physiological effects of this improvement in mineral nutrition, but ONP did have growth increases when mycorrhizal. Forestburg showed increased mineral concentration with mycorrhizae only for P, but did show positive mycorrhizal responses by increasing physiological parameters such as mass and root length. These data suggest that, in this study, improved mineral nutrition was not the sole causal factor for plant physiological responses to mycorrhizae. Mineral

uptake was also found to be distinct from biomass increase in Hetrick et al. (1996). The high cost of symbiosis under greenhouse growth conditions may be contributing to the lack of physiological benefits observed. In this study, a relationship between C cost and mineral gain was not observed.

The level of benefit that plants gain from mycorrhizal colonization depends upon the environmental conditions. Plants grown in soil with a high mineral content tend to show less benefit or even depressed growth under mycorrhizal conditions (Menge et al., 1978; Francis and Read, 1995). The phenomenon of high soil P levels having a negative effect on mycorrhizal colonization is widely known. A list of hypotheses for plant control of colonization mediated by P availability has been reviewed in Schwab et al. (1991). A recent model (Bago et al., 2000) proposes that P transport into host cells is linked directly to carbohydrate transfer to fungal hyphae through a common translocator. Such a system would limit the amount of carbohydrate spent on mycorrhizal symbiosis in excess mineral conditions, keeping their cost/benefit ratio manageable.

Responses to mycorrhizal colonization varied depending upon variety of switchgrass and experimental conditions. Responses ranged from negative to positive, spanning the symbiotic continuum. The ability to show a range of responses within a symbiotic association is fully described by Bronstein (1994). Selective breeding under fertilized agricultural conditions would tend to propagate genotypes that can respond quickly and strongly to increased mineral availability, thus their response to mycorrhizae would be very sensitive to soil mineral availability. Wild plants tend to have slower growth and

responsiveness to mineral availability. Plants in natural environments are more likely to maintain mycorrhizae, even under conditions when symbiosis is detrimental, in order to gain the benefits of mineral nutrition and stress reduction under certain environmental conditions. Continuing to maintain mycorrhizal symbiosis even when the carbon cost is higher than the mineral benefits has been theoretically modeled as a plausible evolutionary strategy if the potential exists for a beneficial relationship when conditions change (Herre et al., 2000; Tuomi et al., 2001). Under agricultural conditions with high fertilizer input, quick responsiveness of plants is favoured, and plant mycorrhizal colonization would be reduced, even eliminated. Symbiotic associations are less likely to be maintained in agriculture; therefore, wild plants are theoretically more likely to be mycorrhizal dependent.

2.6- Mycorrhizal Responsiveness of Switchgrass Varieties

In this study, the varieties with the overall most positive responses to mycorrhizal colonization are Forestburg, a non-agriculturally selected cultivar, and ONP, a wild variety. They cluster closely together in both experiments. PH, a wild variety that clusters in this group, is not responsive to mycorrhizae, except for a decrease in root P. The cultivars NU and Caddo group together in the second experiment cluster analysis and are generally not responsive to mycorrhizae, except for positive responses by NU in dry mass and root diameter and by Caddo in C, N, and K levels. The wild variety Ojibway that clusters with this group seems to have a negative response to mycorrhizal colonization in both experiments, especially in root parameters. Cluster analysis on effect size does seem

to be adequate for separating varieties based on an overall MD. Clearly, cluster separation is not due to the wild or cultivated status of the variety.

It is interesting to note that fresh mass differences were most important in determining clusters for both experiments while dry mass, the usual biomass parameter, was not. Cluster groups had a clear positive or negative mean mycorrhizal effect size for fresh mass in both cluster analyses, while the other determining factors had more of a gradient response. In the first experiment, the mycorrhizal effect on acid phosphatase level, root diameter and % of small roots were also statistically important in distinguishing clusters. Root length was found to be the only other important factor for cluster determination in the second experiment; both clusters had increased root length with mycorrhizae. In this study, mass was the most important determinant of MD based on effect size, while root architecture was also found to be important.

Other research has shown that MD is closely related to mass differences, root architecture, and P responsiveness (Hetrick et al., 1992 and 1996). Increased shoot mass, acid phosphatase and root diameter, as well as R/S decreases were further pinpointed as determinants of MD in soja (Khalil et al., 1994 and 1999). It has also been suggested that C drain might be more important in determining biomass increases, and therefore MD, than mineral uptake (Haynes et al., 1991; Hetrick et al., 1996). The results of the present study generally correspond with the findings of other researchers. The R/S was found to correspond well to MD in the second experiment of this study, varieties with higher R/S having higher MD. Our attempt to assess MD by amalgamating the effect of mycorrhizae

on many different parameters simultaneously, found fewer factors to be important than the literature would suggest because so few factors significantly affected cluster makeup.

Despite the small number of factors influencing clustering, some trends can be noted by observing the data patterns in the second experiment. Forestburg and ONP are most MD for the physiological factors; Forestburg and the wild varieties have the most root architecture characteristics that indicate high MD; ONP, Caddo and Forestburg responded best to mycorrhizae for mineral uptake. Overall, the cultivated group tends to respond more dramatically to mycorrhizae, mostly based on the influence of Forestburg, the cultivar that has not undergone agricultural selection. The genetic variety of switchgrass was the most important determinant of MD in this study. These experiments have shown that varieties respond to mycorrhizae by altering different parameters. It was also noted that by changing one growth condition, soil volume, mycorrhizal responsiveness differed. This suggests that the specific environment does mediate response.

2.7- Wild vs. Cultivated

The varieties in this study have shown distinct responsiveness patterns to mycorrhizal colonization. Because of the strong responses of some varieties for some of the parameters, most of the comparisons between wild and cultivated groups suggest that the cultivated group is the more mycorrhizal responsive group. However, this trend is most likely due to the positive responses to mycorrhizae that were seen in the Forestburg variety. Since Forestburg is a cultivar that has not undergone selection under an

agricultural environment, it is reasonable that in the cluster analysis, it groups with the wild varieties more closely than with the other cultivars.

The highly selected varieties NU and Caddo have different responses to mycorrhizae. NU does not respond consistently with either positive or negative effects. In the first experiment its responses are more positive, and in the second they are neutral. It is not clear whether selection under an agricultural environment has affected the MD of the variety NU. Caddo does not respond to mycorrhizae by altering its physiological or root parameters, but does show a positive response in terms of mineral uptake. Mycorrhizal functioning as determined by ^{32} P uptake was also found not to be related to biomass increases in wheat cultivars (Hetrick et al., 1996). It could be that the mineral efficiency, i.e. the ability to use minerals to increase biomass or fitness, is low in this switchgrass cultivar. The high level of acid phosphatase in the cultivar Caddo also suggests low P efficiency (McLauchlan, 1980). The overall mineral levels were lower in Caddo, possibly allowing for low mineral efficiency. It is possible that mycorrhizae are benefiting Caddo plants, but the poor mineral conversion efficiency lowers its mycorrhizal dependency. In this case I would conjecture that MD is decreasing for Caddo.

The response of the wild switchgrass varieties is puzzling in terms of expected results. Previous research on wild populations of switchgrass has indicated high dependency on mycorrhizae for growth (Hetrick et al., 1988; Bredja et al., 1998; Entry et al., 1999). In this study, mycorrhizal colonization was detrimental for Ojibway in almost all categories, while PH had unclear responses, some positive and some negative. The only wild variety

that shows high MD is ONP. Because of this, we cannot affirm that the natural condition of switchgrass is to have a high MD. However, since the studies mentioned above took place in the U.S mid west, the place of origin for the switchgrass populations that were made into agricultural cultivars, their comparisons of wild and cultivated plants may be less tainted by the possibility of geographic ecotype differences. It has been shown that the shorter growing season found in the colder climate of Canada limits sporulation of *Glomales* in association with warm season grasses; this may induce the symbiosis to be less beneficial during colder times of the year (Bentivenga and Hetrick, 1991). The effect of climate may be limiting the MD of wild switchgrass in Canada.

Based on mass and root architecture, cultivated wheat was shown to be more MD than its wild ancestors in several studies (Kapulnik and Kushnir, 1991; Hetrick et al., 1992; Hetrick et al., 1993). In highly cultivated species such as wheat, corn and soja, the responses of the less selected varieties were not consistent with each other. This corresponds to what was found for wild switchgrass varieties in this study. For soybean plants, higher MD was based on root architecture, increases in P and acid phosphatase (Khalil et al., 1999). Although mass responses were higher in certain non-selected varieties of soja and corn, mineral responses were in general greater in the selected cultivars. These results led the authors to conclude that although cultivars were not more MD than non-selected varieties, the latter were quite different from each other in their responses to mycorrhizae (Khalil et al., 1994).

Wild oats, which were less MD than cultivars based on mass, were observed to have higher R/S, %N, and P efficiency, as well as slower growth (Haynes et al., 1991). Cultivated tomatoes were determined to be more MD based on higher % colonization, mass increases, P content and seed number increases, while root plasticity was greater and days to flowering was shorter in mycorrhizal wild varieties (Bryla and Koide, 1990a & b). The wild switchgrass in this study was similar to the wild varieties in those studies as they all have higher R/S, %N, slower growth and more root plasticity than cultivars. Forestburg, the variety with the highest MD, has a high R/S but fast growth and low %N, in contrast to the wild types. These traits of wild switchgrass are likely adapted to deal with unfavourable soil conditions. Since cultivars are less innately prepared to cope with adverse conditions, based on root morphology and growth rate, cultivars should be either more MD, which contradicts predictions from cost/benefit analyses, or did not encounter environmental stress during selection. It is possible that successful non-selected cultivars, such as Forestburg, have both high mycorrhizal responsiveness and large root systems as a functional redundancy, and this is one of the reasons why they grow successfully in marginal conditions.

Corn, soja and wheat are 'high input' crops which receive lots of fertilizer and pesticides when grown under agricultural conditions. Less fertilizer or chemicals are applied to oats and possibly tomatoes. The lower input crops (oats, tomato) have been found to have an increased MD in cultivation, while wild varieties are more MD for corn, soja and wheat. Switchgrass, as a low input crop, is more likely to respond similarly to oats than to wheat.

2.8- Implications for Agriculture

The inclusion of mycorrhizae in sustainable agriculture has been a vision to many mycorrhizal researchers (Hayman, 1982; Hamel, 1995; Plenchette and Strullu, 1995). Selective breeding for MD may have a place in sustainable agriculture practices. This study has shown that there are variations in the MD of switchgrass varieties, and that the manner by which cultivars are affected by mycorrhizae depends on variety and environment. For example, based on the results of this study, if the goal was to maximize biomass production with mycorrhizae, Forestburg or ONP would be the best varieties. If the goal were to grow on low mineral soil, Caddo or ONP, which gain the most by mycorrhizae in terms of mineral content and have low mineral requirements, would be more appropriate. The genetic diversity among switchgrass varieties remains high and therefore selective breeding for MD based on the variances already seen in varieties is feasible. It should be noted however, that plant-fungal species preference might be important in determining response of switchgrass to mycorrhizal fungi.

Farming practices that are less likely to disrupt natural mycorrhizal populations are being promoted as part of a sustainable agriculture management plan. In Europe, where organic farming is more widespread and there is some aversion to using genetically modified organisms, mycorrhizae-friendly farming practices are more likely to take hold. Organic farms which did not till and used crop rotation, experienced a higher than expected yield, only 10% less than conventional systems, and this was attributed to mycorrhizae (Mader et al., 2000).

Low-input farming has the additional advantage of maintaining natural levels of nutrients, mycorrhizal populations, and biodiversity in the soil (Podesfinski et al., 2001). Under these conditions, inadvertent selection against mycorrhizal symbiosis should not occur. Lower MD of cultivated wheat and lower MD of corn selected for fungal resistance, indicate that selection against mycorrhizae does occur in agricultural systems. The fungi themselves are also subject to inadvertent selection. It has been suggested that AM fungal species that can colonize new roots from colonized root particles in the soil or mycelium broken off from the network are at a selective advantage in tilled systems. *Glomus* species can colonize from spores, hyphae or infected roots while *Gigaspora* species depend more on their large spores for propagation (Smith and Read, 1997). Pot experiments by Boddington and Dodd (2000b) led them to suggest that *Glomus* species are 'aggressive' and selected for under agricultural environments while *Gigaspora* species are not. A similar effect was found by Hendrix et al. (1995) where *Gigaspora* species dominated under a monoculture of soybean, while *Glomus* species dominated under crop rotations.

While some *Glomus* species have been shown to be compatible symbiotic partners for switchgrass (Clark and Zeto, 1999; Eom et al., 2000), *Gigaspora* species are more common in some agricultural soils (Podeszfinski et al., 2001) and in mature tallgrass prairie soils (R.M Miller, pers. comm.). The possible inadvertent selection under conventional agricultural systems for certain mycorrhizal species may limit the responsiveness of switchgrass, since symbiont preference has been shown in this species. This is another argument in favour of non-disruptive practices for sustainable agriculture.

3.0- Conclusions

As a general rule, we cannot conclude that the wild varieties of switchgrass are more MD than cultivars, as was stated in our original hypothesis. The present study concurred with other studies on cultivation and mycorrhizae, which found that wild varieties are inconsistent in their MD. Therefore, it is not justified to compare wild varieties to cultivars as a unit. In the cluster analysis of the second experiment, the cluster of ONP, PH and Forestburg encompasses the varieties that responded positively to mycorrhizal colonization for most of the parameters. Among these, the first two are wild and the third one is cultivated but has not undergone any selection. NU and Caddo, the selected cultivars, clustered together and with the wild variety Ojibway. Varieties in this group tended to have neutral or negative responses to mycorrhizae, but not for every parameter.

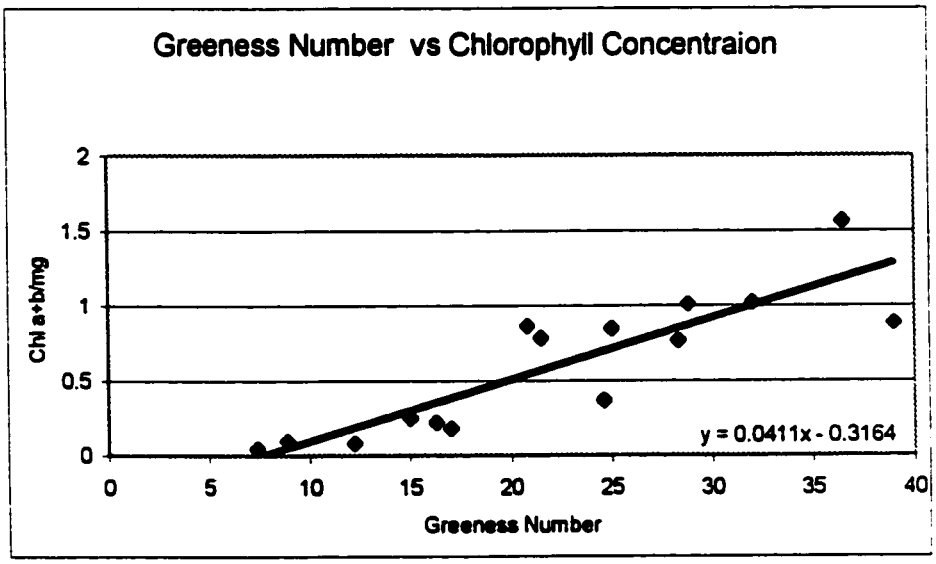
The most important factors in determining mycorrhizal response were the variety of switchgrass and the growth conditions, especially the soil volume, not the cultivation level. It was noted that not only do varieties differ in MD, but in the type of response they have when mycorrhizal. Varieties respond with either increased physiological, root or mineral parameter responses. Mineral increases did not correlate with biomass increases in this study. Assessment of MD by CDA on effect size found in both experiments that fresh mass was the most important parameter by which varieties were clustered. It was also observed the R/S was a good predictor of MD in the second experiment. These results are in agreement with other studies, which determined that biomass increases and root architecture are important determinants of MD in grasses.

There are definite differences in MD among switchgrass varieties in this study. Plant breeders may enhance these genotype differences to increase the MD of certain cultivars. A cultivar that is highly mycorrhizal efficient would be attractive for sustainable agriculture practices.

To further determine the impact of agriculture on the ability of crop plants to benefit from mycorrhizal associations, similar studies should be done for other species for which a breeding history record exists. It might also be prudent to use species that have been cultivated for longer than switchgrass. It would also be interesting to pursue the nature of the MD differences found among the switchgrass varieties used in this study. A field study would help to determine if mycorrhizal responsiveness differs under a more natural environment. An examination of the natural soil types and fungal populations at the wild sites might indicate why wild varieties ONP and Ojibway showed opposite mycorrhizal responses despite being geographically close. The question of switchgrass ploidy remains elusive. It is recommended that flow cytometry be used to determine switchgrass ploidies. This technique would be useful to compare cultivar ploidies to those found by other researchers, and it would be both interesting and important to assess the ploidy of wild Canadian populations of switchgrass for prairie revegetation efforts.

Appendix

1.0- Standard curve for greenness number vs. chlorophyll concentration.



References

Alderson J, Sharp W.C. 1994. Grass Varieties in the United States. Agriculture Handbook No. 170. Soil Conservation Service, USDA, Washington D.C.

Allen M.F, Sexton J.C, Moore T.S, Christensen M. 1981. Influence of Phosphate Source on Vesicular-Arbuscular Mycorrhizae of *Bouteloua gracilis*. New Phytologist 87: 75-82

Anderson A.J. 1988. Mycorrhizae- Host Specificity and Recognition. The American Phytopathological Society 78:375-378

Anderson B, Matches A.G. 1983. Forage, Yeild, Quality and Persistence of Switchgrass and Caucasian Bluestem. Agronomy Journal 75: 119-124

Anderson R.C, Hetrick B.A.D, Wilson G.W.T. 1994. Mycorrhizal Dependence of *Andropogon gerardii* and *Schizachyrium scoparium* in Two Prairie Soils. American Midland Naturalist 132: 366-376

Atsatt P.R. 1988. Are Vascular Plants 'Inside-Out' Lichens? Ecology 69: 17-23

Azcon R, Ocampo J.A. 1981. Factors Affecting the Vesicular-Arbuscular Colonization and Mycorrhizal Dependency of Thirteen Wheat Cultivars. New Phytologist 87:677-685

Baath E, Hayman D.S. 1984. Effect of Soil Volume and Plant Density on Mycorrhizal Infection and Growth Response. Plant and Soil 77: 373-376

Bago B, Pfeffer P, Shachar-Hill Y. 2000. Carbon Metabolism and Transport in Arbuscular Mycorrhizas. Plant Physiology 124: 949-957

Bago B, Pfeffer P, Shachar-Hill Y. 2001. Could the Urea Cycle be Translocating Nitrogen in the Arbuscular Mycorrhizal Symbiosis? New Phytologist 149: 4-8

Balasko J.A, Burner D.M, Thayne W.V. 1984. Yield and Quality of Switchgrass Grown Without Soil Amednments. Agronomy Journal 76: 204-208

Baon J.B, Smith S.E, Alston A.M, Wheeler R.D. 1992. Phosphorus Efficiency of Three Cereals as Related to Indigenous Mycorrhizal Infection. Australian Journal of Agricultural Research 43: 479-491

Baon J.B, Smith S.E, Alston A.M. 1993. Mycorrhizal Responses of Barley Cultivars Differing in P Efficiency. Plant and Soil 157: 97-105

Barker R.E, Haas R.J, Jacobson E.T, Berdahl J.D. 1988. Registration of 'Forestburg' Switchgrass. Crop Science. 28: 192-193

Bauhus J, Messier C. 1999. Evaluation of Fine Root Length and Diameter Measurements obtained using RHIZO Image Analysis. Agronomy Journal 91: 142-147

Baylis G.T.S. 1970. Root Hairs and Phycomycetous Mycorrhizas in Phosphorus-Deficient Soil. Plant and Soil 33: 713-716

Bentivenga S.P, Hetrick B.A.D. 1992. Seasonal and Temperature Effects on Mycorrhizal Activity and Dependence of Cool- and Warm-Season Tallgrass Prairie Grasses. Canadian Journal of Botany 70: 1596-1602

Berta G, Fusconi A , Sampo S, Lingua G, Perticone S, Repetto O. 2000. Polyploidy in Tomato Roots as Affected by Arbuscular Mycorrhizal Colonization. Plant and Soil 226: 37-44

Bertheau V, Gianinazzi-Pearson V, Gianinazzi, S. 1980. Développement et Expression de l'Association Endomycorhizienne chez le Blé I. – Mise en Évidence d'un Effet Variétal. Annales de l'Amélioration des Plantes. 30: 67-78

Bever J.D, Morton J.B, Antonovics J, Schultz P.A. 1996. Host-Dependent Sporulation and Species Diversity of Arbuscular Mycorrhizal Fungi in a Mown Grassland. Journal of Ecology 84: 71-82

Blilou I, Ocampo J.A, Garcia-Garrido J. 1999. Resistance of Pea Roots to Endomycorrhizal Fungus or *Rhizobium* Correlates with Enhanced Levels of Endogenous Salicylic Acid. Journal of Experimental Biology 50:1663-1668

Boddington C.L, Dodd J.C. 1998. A Comparison of the Development and Metabolic Activity of Mycorrhizas Formed by Arbuscular Mycorrhizal Fungi from Different Genera on Two Tropical Forage Legumes. Mycorrhiza 8: 149-157

Boddington C.L, Dodd J.C. 2000. The Effect of Agricultural Practices on the Development of Indigenous Arbuscular Mycorrhizal Fungi.I. Field Studies in an Indonesian Ultisol. Plant and Soil 218: 137-144

Boddington C.L, Dodd J.C. 2000b. The Effect of Agricultural Practices on the Development of Indigenous Arbuscular Mycorrhizal Fungi. II. Studies in an Experimental Microcosm. Plant and Soil 218: 145-157

Boucher A, Dalpé Y, Charest C. 1999. Effect of Arbuscular Mycorrhizal Colonization of Four Species of *Glomus* on Physiological Responses of Maize. Journal of Plant Nutrition 22: 783-797

Bredja J.J, Yocom D.H, Moser L.E, Waller S.S. 1993. Dependence of 3 Nebraska Sandhills Warm-Season Grasses on Vesicular-Arbuscular Mycorrhizae. Journal of Range Management 46: 14-20

Bredja J.J, Moser L.E, Vogel K.P. 1998. Evaluation of Switchgrass Rhizoshpere Microflora for Enhancing Seedling Yield and Nutrient Uptake. Agronomy Journal 90: 753-758

Bronstein J.L. 1994. Conditional Outcomes in Mutualistic Interactions. Trends in Ecology and Evolution 9: 214-217

Bruinsma, J.1963.The Quantitative Analysis of Chlorophyll a and b in Plant Extracts. Photochemical Photobiology 72: 241-249.

Brundrett M, Melville L, Peterson L. 1994. Practical Methods in Mycorriza Research. Mycologue Publications.

Bryla D.R, Koide R.T. 1990. Regulation of Reproduction in Wild vs Cultivated *Lycopersion esculentum* Mill. By Vesicular –Arbuscular Mycorrhizal Infection. Oecologia 84:74-81

Bryla D.R, Koide R.T. 1990b. Role of Mycorrhizal Colonization in the Growth and Reproduction of Wild vs Cultivated Plants II. Oecologia 84:82-92

Charest C, Clark G, Dalpé Y. 1997. The Impact of Arbuscular Mycorrhizae and Phosphorus Status on Growth of Two Turfgrass Species. Journal of Turfgrass Management 2: 1-14

Clark R.B, Zobel R.W. Zeto S.K. 1999. Effects of Mycorrhizal Fungus Isolates on Mineral Acquisition by *Panicum virgatum* in acidic soil. Mycorrhiza 9: 167-176

Clark R.B, Zeto S.K, Zobel R.W. 1999b. Arbuscular Mycorrhizal Fungal Isolate Effectiveness on Growth and Root Colonization of *Panicum virgatum* in Acidic Soil. Soil Biology and Biochemistry 31: 1757-1763

Clark R.B, Zeto S.K. 2000. Mineral Acquisition by Arbuscular Mycorrhizal Plants. Journal of Plant Nutrition. 23: 867-902

Deacon J.W. 1997. Modern Mycology. Blackwell Science Ltd, Oxford .

De Bary A. 1879. Nat. Versam. Cassel . L1 121

Dekkers T.B.M, van der Werff P.A. 2000. Mutualistic Functioning of Indigenous Arbuscular Mycorrhizae in Spring Barley and Winter Wheat After Cessation of Long-Term Phosphate Fertilization. Mycorrhiza 10: 195-201

Dhillion S.S. 1992. Evidence for Host-Mycorrhizal Preference in Native Grassland Species. Mycological Research 96: 359-362

Dickson S, Smith S.E, Smith F.A. 1999. Characterization of two Arbuscular Mycorrhizal Fungi in Symbiosis with *Allium porrum*: colonization, Plant Growth and Phosphate Uptake. New Phytogist 144: 163-172

Dodd J.D. 2000. The Role of Arbuscular Mycorrihzal Fungi in Agro-and Natural Ecosystems. Outlook on Agriculture 29: 55-62

Dore W.G, Mc Neill J. 1980. Grasses of Ontario. Research Branch Agriculture Canada. Hull Qc.

Douds D.D, Johnson C.R, Koch K.E. 1988. Carbon Cost of the Fungal Symbiont Relative to Net Leaf P Accumulation in a Split-Root VA Mycorrhizal Symbiosis. Plant Physiology 86: 491-496

Drijber R.A, Doran J.W, Parkhurst A.M, Lyon D.J. 2000. Changes in Soil Microbial Community Structure with Tillage Under Long-Term Wheat-Fallow Management. Soil Biology and Biochemistry 32: 1419-1430

Duc G, Trouvelot A, Gianinizzi-Pearson V, Gianinazzi S. 1989. First Report of Non-Mycorrhizal Plant Mutants (Myc-) Obtained in Pea (*Pisum sativum* L.) and Fababean (*Vicia faba* L.). Plant Science 60: 215-222

Eberhart S.A, Newell L.C. 1959. Variation in Domestic Collections of Switchgrass, Panicum virgatum L. Agronomy Journal 51: 613-616

Elbersen H.W, Ocumpaugh W.R, Hussey M.A, Sanderson M.A, Tischler C.R. 1999. Field Evaluation of Switchgrass Seedlings Divergently Selected for Crown Node Placement. Crop Science 39: 475-479

Entry J.A, Watrud L. S, Reeves M. 1999. Accumulation of ^{137}Cs and ^{90}Sr from Contaminated Soil by Three Grass Species Inoculated with Mycorrhizal Fungi. Environmental Pollution 104: 449-457

Eom A-H, Hartnett D.C, Wilson G.W.T. 2000. Host Plant Species Effects on Arbuscular Mycorrhizal Fungal Communities in Tallgrass Prairie. Oecologia 122: 435-444

Estaún V, Calvet C, Hayman D.S. 1987. Influence of Plant Genotype on Mycorrhizal Infection: Response of Three Pea Cultivars. Plant and Soil 103: 295-298

Filion M, St-Arnaud M, Fortin J.A. 1999. Interaction Between the AM Fungus *Glomus intraradices* and Different Rhizosphere Microorganisms. Proceedings from Mycorhizes 1999, pg 13

Fitter A.H. 1991. Costs and Benefits of Mycorrhizas: Implications for Functioning Under Natural Conditions. Experientia 47: 350-355

Francis R, Read D.J. 1984. Direct Transfer of Carbon Between Plants Connected by Mycorrhizal Mycelium. Nature 307: 53-56

Francis R, Read D.J. 1995. Mutualism and Antagonism in the Mycorrhizal Symbiosis, With Special Reference to Impacts on Plant Community Structure. Canadian Journal of Botany 73(S1): S1301-S1309

Franke-Snyder M, Douds D.D, Galvez L, Phillips J.G, Wagoner P, Drinkwater L, Morton J.B. 2001. Diversity of Communities of Arbuscular Mycorrhizal (AM) Fungi Present in Conventional Versus Low-Input Agricultural Sites in Eastern Pennsylvania. Applied Soil Ecology 16: 35-48

Gahoonia T.S, Nielsen N.E, Lyshede O.B. 1999. Phosphorus (P) Acquisition of Cereal Cultivars in the Field at Three Levels of P Fertilization. Plant and Soil 211: 269-281

Galvez L, Douds D.D, Drinkwater L.E, Wagoner P. 2001. Effect of Tillage and Farming System Upon VAM Fungus Populations and Mycorrhizas and Nutrient Uptake of Maize. Plant and Soil 228: 299-308

Gange A.C. 1999. On the Relation Between Arbuscular Mycorrhizal Colonization and Plant 'Benefit'. Oikos 87: 615-621

Gemma J.N, Koske R.E, Roberts E.M, Jackson N, DeAntonis K.M. 1997. Mycorrhizal Fungi Improve Drought Resistance in Creeping Bentgrass. Journal of Turfgrass Science 73:15-29

Gianinazzi-Pearson V, Gianinazzi, S. 1976. Enzymatic Studies on the Metabolism of Vesicular-Arbuscular Mycorrhiza. Physiologie Végétale 14:833-841

Gianinazzi-Pearson V, Gianinazzi, S. 1983.The Physiology of Vesicular-Arbuscular Mycorrhizal Roots. Plant and Soil 71: 197-209

Godshalk E.B, Timothy D.H, Burns J.C. 1988. Effectiveness of Index of Selection for Switchgrass Forage Yield and Quality. Crop Science 28: 825- 830

Graham J.H. 2000. Assessing Costs of Arbuscular Mycorrhizal Symbiosis in Agroecosystems. In 'Current Advances in Mycorrhizae Research'. Editors G.K. Podila and D.D. Douds. APS Press, St Paul, US. Pg 127-137

Graham J.H, Duncan L.W, Eissenstat D.M. 1997. Carbohydrate Allocation Patterns in Citrus Genotypes as Affected by Phosphorus Nutrition, Mycorrhizal Colonization and Mycorrhizal Dependency. New Phytologist 135: 335-343

Graham J.H, Abbott L.K. 2000. Wheat Responses to Aggressive and Non-Aggressive Arbuscular Mycorrhizal Fungi. Plant and Soil 220: 207-218

Grierson P.F, Comerford N.B. 2000. Non-Destructive Measurement of Acid Phosphatase Activity in the Rhizosphere using Nitrocellulose Membranes and Image Analysis. Plant and Soil 218: 49-57

Gunter L.E, Tuskan G.A, Wullschleger S.D. 1996. Diversity among Populations of Switchgrass Based on RAPD Markers. Crop Science 36:1017-1022

Guo B.Z, An Z-Q, Hendrix J.W, Dougherty C.T. 1993. Influence of a Change From Tall Fescue to Pearl Millet or Crabgrass on the Mycorrhizal Fungal Community. Soil Science 155: 398-405

Hamel C. 1995. Les Endomycorrhizes a Vésicules et Arbuscules en Agriculture: Chimère où Panacée? In 'La Symbiose Mycorrhizienne'. Editeurs J.A Fortin, C. Charest et Y.Piché. Editions Orbis, Québec. p. 145

Haynes B, Koide R.T, Elliott G. 1991. Phosphorus Uptake and Utilization in Wild and Cultivated Oats. Journal of Plant Nutrition 14: 1105-1118

Haynes J.G, Pill W.G, Evans T.A. 1997. Seed Treatments Improve the Germination and Seedling Emergence of Switchgrass (*Panicum virgatum* L.). HortScience 32; 1222-1226

Heckman J.R, Angle J.S. 1987. Variation Between Soybean Cultivars in Vesicular-Arbuscular Mycorrhiza Fungi Colonization. Agronomy Journal 79: 428-430

Hendrix J.W, Jones K.J, Nesmith W.C. 1992. Control of Pathogenic Mycorrhizal Fungi in Maintenance of Soil Productivity by Crop Rotation. Journal of Produce Agriculture 5: 383-386

Hendrix J.W, Guo B.Z, An Z-Q. 1995. Divergence of Mycorrhizal Fungal Communities in Crop Production Systems. Plant and Soil 170: 131-140

Herre E.A, Knowlton N, Mueller U.G, Rehner S.A. 2000. The Evolution of Mutualisms: Exploring the Paths Between Conflict and Cooperation. Trends in Ecology and Evolution 14: 49-52

Hetrick B.A.D. 1991. Mycorrhizas and Root Architecture. Experientia 47: 355-361

Hetrick B.A.D, Bloom J. 1983. Vesicular-Arbuscular Mycorrhizal Fungi Associated with Native Tall Grass Prairie and Cultivated Winter Wheat. Canadian Journal of Botany 61:2140-2146

Hetrick B.A.D, Kitt D.G, Wilson, G.T. 1988. Mycorrhizal Dependence and Growth Habit of Warm Season and Cool Season Tallgrass Prairie Plants. Canadian Journal of Botany 66:1376-1380

Hetrick B.A.D, Wilson, G.T, Todd T.C. 1990. Differential Responses of C_3 and C_4 grasses to mycorrhizal symbiosis, phosphorus fertilization and soil microorganisms. Canadian Journal of Botany 68:461-467

Hetrick B.A.D, Wilson, G.T, Leslie J.F. 1991. Root Architecture of Warm and Cool Season Grasses: Relationship to Mycorrhizal Dependence. Canadian Journal of Botany 69: 112-118

Hetrick B.A.D, Wilson, G.T, Cox T.S. 1992. Mycorrhizal Dependence of Modern Wheat Varieties, Landraces, and Ancestors. Canadian Journal of Botany 70: 2032-2040

Hetrick B.A.D, Wilson, G.T, Todd T.C. 1992b. Relationships of Mycorrhizal Symbiosis, Rooting Strategy, and Phenology Among Tallgrass Prairie Forbs. Canadian Journal of Botany 70: 1521-1528

Hetrick B.A.D, Wilson, G.T, Cox T.S. 1993. Mycorrhizal Dependence of Modern Wheat Cultivars and Ancestors: A Synthesis. Canadian Journal of Botany 71:512-518

Hetrick B.A.D, Wilson, G.T, Todd T.C. 1996. Mycorrhizal Response in Wheat Cultivars: Relationship to Phosphorus. Canadian Journal of Botany 74: 19-25

Hitchcock, A.S. 1950. Manual of the Grasses of the United States. US Government Printing Office, Washington

Hodge A, Robinson D, Fitter A.H. 2000. An Arbuscular Mycorrhizal Inoculum Enhances Root Proliferation in, but not Nitrogen Capture From, Nutrient-Rich Patches in Soil. New Phytologist 145: 575-584

Hope H.J, McElroy A. 1990. Low-Temperature Tolerance of Switchgrass (*Panicum virgatum* L.). Canadian Journal of Plant Science 70:1091-1096

Hopkins A.A, Vogel K.P, Moore K.J, Johnson K.D, Carlson I.T. 1995. Genotype Effects and Genotype by Environment Interactions for Traits of Elite Switchgrass Populations. Crop Science 35: 125-132

Hopkins A.A, Taliaferro C.M, Murphy C.D, Christian D. 1996. Chromosome Number and Nuclear DNA Content of Several Switchgrass Populations. Crop Science 36: 1192-1195

Hopkins A.A, Taliaferro C.M. 1997. Genetic Variation Within Switchgrass Populations for Acid Soil Tolerance. Crop Science 37: 1719-1722

Hopkins, W.G. 1995. Introduction to Plant Physiology. John Wiley & Sons Inc. Toronto, ON.

Hultquist S.J, Vogel K.P, Lee D.J, Arumuganathan K, Kaeppler S. 1996. Chloroplast DNA and Nuclear DNA Content Variations among Cultivars of Switchgrass, *Panicum virgatum* L. Crop Science 36: 1049-1052

Jasper D.A, Abbott L.K, Robson A.D. 1989. Soil Disturbance Reduces the Infectivity of External Hyphae of Vesicular-Arbuscular Mycorrhizal Fungi. New Phytologist 112: 93-99

Jeffries P. 1987. Use of Mycorrhizae in Agriculture. Critical Reviews in Biotechnology 5: 319-355

Johnson N.C. 1998. Responses of *Salsola kali* and *Panicum virgatum* to mycorrhizal fungi, phosphorus and soil organic matter: implications for reclamation. Journal of Applied Ecology 35:86-94

Johnson N.C, Graham J.H, Smith F.A. 1997. Functionning of Mycorrhizal Associations Along the Mutualism-Parasitism Continuum. New Phytologist 135: 575-585

Joner E.J, Ravnskov S, Jakobsen I. 2000. Arbuscular Mycorrhizal Phosphate Transport under Monoxenic Conditions using Radio-Labelled Inorganic and Organic Phosphate. Biotechnology Letters 22:1705-1708

Jun D.J, Allen E.B. 1991. Physiological Responses of 6 Wheatgrass Cultivars to Mycorrhizae. Journal of Range Management 44: 336-341

Jung G.A, Shaffer J.A, Stout W.L. 1988. Switchgrass and Big Bluestem Responses to Amendments on Strongly Acid Soil. Agronomy Journal 80: 669-676

Kabir Z, O'Halloran I.P, Widden P, Hamel C. 1998. Vertical Distribution of Arbuscular Mycorrhizal Fungi Under Corn (*Zea mays* L.) in No-Till and Conventonal Tillage Systems. Mycorrhiza 8: 53-55

Kaeppler S.M, Parke J.L, Mueller S.M, Senior L, Stuber C, Tracy W.F. 2000. Variation Among Maize Inbred Lines and Detection of Quantitative Trait Loci for Growth at Low Phosphorus and Responsiveness to Arbuscular Mycorrhizal Fungi. Crop Science 40: 358-364

Kaldorf M, Ludwig-Muller J. 2000. AM Fungi Might Affect the Root Morphology of Maize by Increasing Indole-3-Butyric Acid Biosynthesis. Physiologia Plantarum 109: 58-67

Kapulnik Y, Kushnir U. 1991. Growth Dependency of Wild, Primitive and Modern Cultivated Wheat Lines on Vesicular-Arbuscular Mycorrhiza Fungi. Euphytica 56: 27-36

Kendrick B.1991.Fungal Symbioses and Evolutionary Innovations. In Symbiosis as a Source of Evolutionary Innovation Eds L Margulis and R Fester. The MIT Press.

Khalil S, Loynachan T.E, Tabatai M.A. 1999. Plant Determinants of Mycorrhizal Dependency in Soybean. Agronomy Journal 91: 135-141

Khalil S, Loynachan T.E, Tabatai M.A. 1994. Mycorrhizal Dependency and Nutrient Uptake by Improved and Unimproved Corn and Soyabean Cultivars. Agronomy Journal 86: 949- 958

Koide R. 2000. Functional Complementarity in the Arbuscular Mycorrhizal Symbiosis. New Phytologist. 147:233-235

Koide R, Li M, Lewis J, Irby C. 1988. Role of Mycorrhizal Colonization in the Growth and Reproduction of Wild vs Cultivated Plants I. Oecologia 77:537-543

Koide R.T, Goff M.D, Dickie I.A. 2000. Component Growth Efficiencies of Mycorrhizal and Nonmycorrhizal Plants. New Phytologist 148: 163-168

Krishna K.R, Shetty K.G, Dart P.J, Andrews D.J. 1985. Genotype Dependent Variation in Mycorrhizal Colonization and Response to Inoculation of Pearl Millet. Plant and Soil 86: 113-125

Lackie S.M, Bowley S.R, Peterson R.L. 1988. Comparison of Colonization Among Half-Sib Families of *Medicago sativa* L. by *Glomus versiforme* (Daniels & Trappe) Berch. New Phytologist 108: 477-482

Levy D.B, Redente E.F, Uphoff G.D. 1999. Evaluating the Phytotoxicity of Pb-Zn Tailings to Big Bluestem (*Andropogon gerardii* Vitman) and Switchgrass (*Panicum virgatum* L.) Soil Science 164: 363-375

Liu A, Hamel C, Hamilton R.I, Ma B.L, Smith D.L. 2000. Acquisition of Cu, Zn, Mn and Fe by Mycorrhizal Maize (*Zea mays* L.) Grown in Soil at Different P and Micronutrient Levels. Mycorrhiza 9:331-336

Looman J. 1983. 111 Range and Forage Plants of the Canadian Prairies. Research Branch, Agriculture Canada. Ottawa, On. Pg 106

Madakadze I.C, Coulman B.E, Peterson P.R, Stewart K.A, Samson R, Smith D.L. 1998. Leaf Area Development, Light Interception, and Yield among Switchgrass Populations in a Short-Season Area. Crop Science 38: 827-834

Madakadze I.C, Coulman B.E, Stewart K.A, Peterson P.R, Samson R, Smith D.L. 1998b. Phenology and Tiller Characteristics of Big Bluestem and Switchgrass Cultivars in a Short Growing Season Area. Agronomy Journal 90: 489-495

Madakadze I.C, Stewart K.A, Peterson P.R, Coulman B.E, Smith D.L. 1999. Swichgrass Biomass and Chemical Composition for Biofuel in Eastern Canada. Agronomy Journal 91: 696-701

Madakadze I.C, Stewart K.A, Madakadze R.M, Peterson P.R, Coulman B.E, Smith D.L. 1999b. Field Evaulation of the Chlorophyll Meter to Predict Yield and Nitrogen Concenration of Switchgrass. Journal of Plant Nutrition 22: 1001-1010

Mader P, Edenhofer S, Boller T, Wiemken A, Niggli U. 2000. Arbuscular Mycorrhizae in a Long-Term Field Trial Comparing Low-Input (Organic, Biological) and High-Input (Conventional) Farming Systems in a Crop Rotation. Biology and Fertility of Soils 31: 150-156

Maier W, Hammer K, Dammann U, Schulz B, Strack D. 1997. Accumulation of Sesquiterpenoid Cyclohexenone Derivatives Induced by an Arbuscular Mycorrhizal Fungus in Members of the Poacae. Planta 202:36-42

Malloch D.W, Pirozynski K.A, Raven P.H. 1980. Ecological and Evolutionary Significance of Mycorrhizal Symbiosis in Vascular Plants. Proc. Natl. Acad. Sci. 77(4): 2113-2118

Manjunath A, Habte M. 1991. Root Morphological Characteristics of Host Species Having Distinct Mycorrhizal Dependency. Canadian Journal of Botany 69: 671-676

Manjunath A, Habte M. 1991b. Relationship Between Mycorrhizal Dependency and Rate Variables Associated with P Uptake, Utilization, and Growth. Communications in Soil Science and Plant Analysis 22: 1423-1437

Manjunath A, Habte M. 1992. External and Internal P Requirements of Plant Species Differing in Their Mycorrhizal Dependency. Arid Soil Research and Rehabilitation 6: 271-284

Manske G.G.B. 1989. Genetical Analysis of the Efficiency of VA Mycorrhizae with Spring Wheat. Agriculture, Ecosystems, and the Environment 29:273-280

Manske G.G.B, Ortiz-Monasterio J.I, Van Ginkel M, Gonzalez R.M, Rajaram S, Molina E, Vlek P.L.G. 2000. Traits Associated with Improved P-Uptake Efficiency in CIMMYT's Semidwarf Spring Bread Wheat Grown on Acid Andisol in Mexico. Plant and Soil 221: 189-204

McGonigle T.P, Miller M.H. 2000. The Inconsistent Effect of Soil Disturbance on Colonization of Roots by Arbuscular Mycorrhizal Fungi: A Test of the Inoculum Density Hypothesis. Applied Soil Ecology 14: 147-155

McLaughlan K.D. 1980. Acid Phosphatase Activity of Intact Roots and Phosphorus Nutrition in Plants. 1 Assay Conditions and Phosphatase Actvity. Australian Journal of Agricultural Research 31: 429-440

McMillan C, Weiler J. 1959. Cytogeography of *Panicum virgatum* in Central North America. American Journal of Botany 46(8): 590-593

Menge J.A, Johnson E.L.V, Platt R.G. 1978. Mycorrhizal Dependency of Several Citrus Cultivars Under Three Nutrient Regimes. New Phytologist 81:553-578

Menge J.A. 1983. Utilization of Vesicular-Arbuscular Mycorrhizal Fungi in Agriculture. Canadian Journal of Botany 61: 1015-1024

Mercy M.A, Shivashankar G, Bagyraj D.J. 1990. Mycorrhizal Colonization in Cowpea is Host Dependent and Heritable. Plant and Soil 121: 292-294

Miller R.M, Jarster A.G, Pillai J.K. 1987. Biomass Allocation in an *Agropyron smithii-Glomus* Symbiosis. American Journal of Botany 74: 114-122

Miller R.M, Smith C.I, Jastrow J.D, Bever J.D. 1999. Mycorrhizal Status of the Genus *Carex* (Cyperaceae). American Journal of Botany 86:547-553

Monz C.A, Hunt H.W, Reeves F.B, Elliott E.T. 1994. The Response of Mycorrhizal Colonization to Elevated CO_2 and Climate Change in *Pascopyrum smithii* and *Bouteloua gracilis*. Plant and Soil 165: 75-80

Moore K.J, Moser L.E, Vogel K.P, Waller S.S, Johnson B.E, Pedersen J.F. 1991. Describing and Quantifying Growth Stages of Perennial Forage Grasses. Agronomy Journal 83: 1073-1077

Moser L.E, Vogel K.P. 1994. Switchgrass, Big Bluestem and Indiangrass. Chapter 32, Forages Vol 1 : An Introduction to Grassland Agriculure. Eds Barnes R.F, Miller D.A, Nelson C.J. Iowa State University Press.

Mozafar A, Anken T, Ruh R, Frossard E. 2000. Tillage Intensity, Mycorrhizal and Nonmycorrhizal Fungi, and Nutrient Concentrations in Maize, Wheat, and Canola. Agronomy Journal 92: 1117-1124

Newell L.C, Eberhart S.A. 1961. Clone Progeny Evaluation in the Improvement of Switchgrass. *Panicum virgatum* L. Crop Science 1: 117-121

Newman E.I, Reddell P. 1987. The Distribution of Mycorrhizas Among Families of Vascular Plants. New Phytologist 106: 745-751

Nielsen E.L. 1944. Analysis of Variation in *Panicum virgatum*. Journal of Agricultural Research 69: 327-353

Nielsen K.L, Bouma T.J, Lynch J.P, Eissenstat D.M. 1998. Effects of Phosphorus Availability and Vesicular-Arbuscular Mycorrhizas on the Carbon Budget of Common Bean (*Phaseolus vulgaris*). New Phytologist 139: 647-656

Noyd R.K, Pfleger F.L, Russelle M.P. 1995. Interactions Between Native Prairie Grasses and Indigenous Arbuscular Mycorrhizal Fungi: Implications for Reclamation of Taconite Iron Ore Tailing. New Phytologist 129:651-660

O'Bannon J.H, Evans D.W, Peaden R.N. 1980. Alfalfa Varietal Response to Seven Isolates of Vesicular-Arbuscular Mycorrhizal Fungi. Canadian Journal of Plant Science 60: 859-863

Ocampo J.A, Azcon R. 1985. Relationship between the Concentration of Sugars in the Roots and VA Mycorrhizal Infection. Plant and Soil 86:95-100

Ollivier B, Bertheau Y, Diem H.G, Gianinazzi- Pearson V. 1983. Inflence de la Variété de *Vigna unguiculata* dans L'Expression de Trois Associations Endomycorrhiziennes à Vésicules et Arbuscules. Canadian Journal of Botany 61: 354-358

Paradis R, Dalpé Y, Charest C. 1995. The Combined Effect of Arbuscular Mycorrhizas and Short-Term Cold Exposure on Wheat. New Phytologist 129: 637-642

Piepp H, Maier W, Schmidt J, Wray V, Strack D. 1997. Arbuscular Mycorrhizal Fungus-Induced Changes in the Accumulation of Secondary Compounds in Barley Roots. Phytochemisty 44:581-587

Pirozynski K.A. 1980. Interactions Between Fungi and Plants Through the Ages. Canadian Journal of Botany 59:1824-1827

Pleger F.L, Stewart E.L, Noyd R.K. 1994.Role of VAM Fungi in Mine Land Revegetation. In: Mycorrhizae and Plant Health. Editors F.L Pleger and R.G Linderman. APS Press, St Paul Minn, US.

Plenchette C, Strullu D.G. 1995. Mycorhizes et Agriculture Intégrée : Standards Expéerimentaux Dans 'La Symbiose Mycorrhizienne'. Editeurs J.A Fortin, C. Charest et Y.Piché. Editions Orbis, Québec. Pg 126

Podeszfinski C, Dalpé Y, Charest C. 2001. *In Situ* Turfgrass Establishment 1. Responses to Arbuscular Mycorrhizae and Fertilization. Journal of Sustainable Agriculture. In Press.

Puoli J.R, Reid R.L, Belesky D.P. 1992. Photosensitization in Lambs Grazing Switchgrass. Agronomy Journal 84: 1077-1080

Ravnskov S, Nybroe O, Jakobsen I. 1999. Influence of an Arbuscular Mycorrhizal Fungus on *Pseudomonas fluorescens* DF57 in Rhizosphere and Hydrosphere Soil. New Phytologist 142: 113-122

Reeves F.B, Wagner D, Moorman T, Kiel J. 1979. The Role of Encomycorrhizae in Revegetation Practices in the Semi-Arid West.1. A Comparison of Incidence of Mycorrhizae in Severely Disturbed vs. Natural Environments. American Journal of Botany 66:6-13

Reinhardt D.R, Miller R.M. 1990. Size Classes of Root Diameter and Mycorrhizal Fungal Colonization in Two Temperate Grassland Communities. New Phytologist 116: 129-136

Riley R.D, Vogel K.P. 1982. Chromosome Numbers of Released Cultivars of Switchgrass, Indiangrass, Big Bluestem and Sand Bluestem. Crop Science 22: 1082-1083

Rosewarne G.M, Barker S.J, Smith S.E, Smith F.A, Schachtman D.P. 1999. A *Lycopericon esculentum* Phosphate Transporter (LePT1) Involved in Phosphorus Uptake from a Vesicular-Arbuscular Mycorrhizal Fungus. New Phytologist 144: 507-516

Rubio R, Moraga E, Borie F. 1990. Acid Phosphatase Activity and Vesicular Arbuscular Mycorrhizal Infection Associated with Roots of Four Wheat Cultivars. Journal of Plant Nutrition 13: 585-598

Ruiz-Lozano JM, Gianinazzi S, Gianinazzi-Pearson V. 1999. Genes Involved in Resistance to Powdery Mildew in Barley Differently Modulate Root Colonization by the Mycorrhizal Fungus *Glomus mossae*. Mycorrhiza 9: 237-240

Rygiewicz P.T, Andersen C.P. 1994. Mycorrhizae Alter Quality and Quantity of Carbon Allocated Below Ground. Nature 369: 58-60

Samson R, Blais P-A, Mehdi B, Girouard P. 1999. Switchgrass Plant Improvement Program for Paper and Agri-Fibre Production in Eastern Canada. Report for Switchgrass Improvement Program-R.E.A.P- Canada, Ste. Anne de Bellevue PQ.

Sanderson M.A, Reed R.L, McLaughlin S.B, Wullschleger S.D, Conger B.V, Parrish D.J, Wolf D.D, Taliaferro C, Hopkins A.A, Ocumpaugh W.R, Hussey M.A, Read J.C, Tischler C.R. 1996. Switchgrass as a Sustainable Bioenergy Crop. Bioresource Technology 56 : 83-93

Schenck N.C, Kinloch R.A. 1980. Incidence of Mycorrhizal Fungi on Six Field Crops in Monoculture on a Newly Cleared Woodland Site. Mycologia 72 : 445-456

Schwab S.M, Menge J.A, Tinker P.B. 1991. Regulation of Nutrient Transfer Between Host and Fungus in Vesicular-Arbuscular Mycorrhizas. New Phytologist 117 : 387-398

Schweiger P.F, Robson A.D, Barrow N.J. 1995. Root Hair Length Determines Beneficial Effect of a *Glomus* Species on Shoot Growth of Some Pasture Species. New Phytologist 131: 247-254

Schweiger P.F, Jakobsen I. 1999. Direct Measurement of Arbuscular Mycorrhizal Phosphorus Uptake into Field-Grown Winter Wheat. Agronomy Journal 91: 998-1002

Scott E.P, Eason W.R, Scullion J. 1994. Effectivity of Indigenous Arbuscular Mycorrhizal Spore Populations from Contrasting Agricultural Management Regimes. In: Mycorrhizas in Integrated Systems: From Genes to Plant Development, Proceedings of the Fourth European Symposium on Mycorrhizas. Eds C Azcon-Aguilar, J.M Barea. Pg591-594

Simon L, Bousquet J, Levesque R.C, Lalonde M. 1993. Origin and Diversification of Endomycorrhizal Fungi and Coincidence with Vascular Land Plants. Nature 363: 67-69

Smith F.A. 2000. Measuring the Influence of Mycorrhizas. New Phytologist 148: 4-6

Smith F.A, Jakobsen I, Smith S.E. 2000. Spatial differences in Acquisition of Soil Phosphate between Two Arbuscular Mycorrhizal Fungi in Symbiosis with *Medicago trunculata*. New Phytologist 147: 357-366

Smith M.D, Hartnett D.C, Rice C.W. 2000. Effects of Long-Term Fungicide Applications on Microbial Properties in Tallgrass Prairie Soil. Soil Biology and Biochemistry 32: 935-946

Smith M.R, Charvat I, Jacobson R.L. 1998. Arbuscular Mycorrhizae Promote Establishment of Prairie Species in a Tallgrass Prairie Restoration. Canadian Journal of Botany 76:1947-1954

Smith S.E, Robson A.D, Abbott L.K. 1992. The Involvement of Mycorrhizas in Assessment of Genetically Dependent Efficiency of Nutrient Uptake and Use. Plant and Soil 146:169-179

Smith S.E, Read D.J. 1997. Mycorrhizal Symbiosis 2nd edition. Academic Press, Toronto.

Solaiman Z, Saito M. 1997. Use of Sugars by Intraradical Hyphae of Arbuscular Mycorrhizal Fungi Revealed by Radiorespirometry. New Phytologist 136: 533-538

Staley T.E, Stout W.L, Jung G.A. 1991. Nitrogen Use by Tall Fescue and Switchgrass on Acidic Soils of Varying Water Holding Capacity. Agronomy Journal 83: 732-738

St-Arnaud M, Caron H.M, Fortin J.A. 1994. Inhibition of *Phytium ultimum* in Roots and Growth Substrate of Mycorrhizal *Tagetes patula* Colonized with *Glomus Intraradices*. Canadian Journal of Plant Pathology 16: 187-194

Subramanian K.S, Charest C. 1995. Influence of Arbuscular Mycorrhizae on the Metabolism of Maize Under Drought Stress. Mycorrhiza 5: 273-278

Subramanian K.S, Charest C, Dwyer L.M, Hamilton R.I. 1996. Effects of Arbuscular Mycorrhizae on Leaf Water Potential, Sugar Content, and P Content During Drought and Recovery of Maize. Canadian Journal of Botany 75: 1582-1591

Subramanian KS, Charest C. 1998. Arbuscular Mycorrhizae and Nitrogen Assimilation in Maize After Drought and Recovery. Physiologia Plantarum 102:285-296

Talbert L.E, Timothy D.H, Burns J.C, Rawlings J.O, Moll R.H. 1983. Estimates of Genetic Parameters in Switchgrass. Crop Science 23:725-728

Talukdar N.C, Germida J.J. 1993. Occurrence and Isolation of Vesicular-Arbuscular Mycorrhizae in Cropped Field Soils of Saskatchewan, Canada. Canadian Journal of Microbiology 39: 567-575

Tarafdar J.C, Marschner H. 1994. Phosphatase Activity in the Rhizosphere and Hydrosphere of VA Mycorrhizal Wheat Supplied with Inorganic and Organic Phosphorus. Soil Biology and Biochemistry 26: 387-395

Tester M, Smith S.E, Smith F.A. 1987. The Phenomenon of "Nonmycorrhizal" Plants. Canadian Journal of Botany 65:419-431

Tinker P.B, Durall D.M, Jones M.D. 1994. Carbon Use Efficiency in Mycorrhizas: Theory and Sample Calculations. New Phytologist 128: 115-122

Toth R, Page T, Castleberry R. 1984. Differences in Mycorrhizal Colonization of Maize Selections for High and Low Ear Leaf Phosphorus. Crop Science 24: 994-996

Toth R, Toth D, Starke D, Smith D.R. 1990. Vesicular-Arbuscular Mycorrhizal Colonization in *Zea mays* Affected by Breeding for Resistance to Fungal Pathogens. Canadian Journal of Botany 68:1039-1044

Tuomi J, Kytoviita M-M, Hardling R. 2001. Cost Efficiency of Nutrient Acquisition and the Advantage of Mycorrizal Symbiosis for the Host Plant. Oikos 92: 62-70

van der Heijden M.G.A, Klironomos J.N, Ursic M, Moutoglis P, Streitwolf-Engel R, Boller T, Wiemken A, Sanders I.R. 1998. Mycorrhizal Fungal Diversity Determines Plant Biodiversity, Ecosystem Variability and Productivity. Nature 396: 69-72

van der Heijden M.G.A, Boller T, Wiemken A, Sanders I.R. 1998b. Different Arbuscular Mycorrhizal Fungal Species are Potential Determinants of Plant Community Structure. Ecology 79: 2082-2091

Vierheilig H, Ocampo J.A. 1991. Receptivity of Various Wheat Cultivars to Colonization by VA Mycorrhizal Fungi as Influenced by Inoculum Potential and the Relation of VAM-effectiveness to Succinic Dehydrogenase Activity of the Mycelium in the Roots. Plant and Soil 133:291-296

Vierheilig H, Garcia-Garrido J.M, Wyss U, Piché Y. 2000. Systemic Suppression of Mycorrhizal Colonization of Barley Roots already Colonized by AM fungi. Soil Biology and Biochemistry 32: 589-595

Vogel K.P. 1987. Seeding Rates for Establishing Big Bluestem and Switchgrass with Preemergence Atrazine Applications. Agronomy Journal 79: 509-512

Vogel K.P, Haskins F.A, Gorz H.J. 1981.Divergent Selection for In Vitro Dry Matter Digestibility in Switchgrass. Crop Science 21: 39-41

Wilson G.W.T, Hartnett D.C. 1997. Effects of Mycorrhizae on Plant Growth and Dynamics in Experimental Tallgrass Prairie Microcosms. American Journal of Botany 84:578-482

Wilson G.W.T, Hartnett D.C. 1998. Interspecific Variation in Plant Responses to Mycorrhizal Colonization in Tallgrass Prairie. American Journal of Botany 85: 1732-1738

Zhu Y-G, Laidlaw A.S, Christie P, Hammond M.E.R. 2000. The Specificity of Arbuscular Mycorrhizal Fungi in Perennial Ryegrass-White Clover Pasture. Agriculture, Ecosystems and the Environment 77:211-218

MQ

66044

UMI

MICROFILMED 2002

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600



PERMISSION DE REPRODUIRE ET DE DISTRIBUER LA THÈSE

PERMISSION TO REPRODUCE AND DISTRIBUTE THE THESIS

NOM DE L'AUTEUR / NAME OF AUTHOR:	GERVAIS, Marc
ADRESSE POSTALE / MAILING ADDRESS:	3-175 Rue Granville, Vanier, Ontario K1L 6Y3
GRADE / DEGREE:	**ANNÉE D'OBTENTION / YEAR GRANTED**
M. A. [redacted]. (SCIENCE POLITIQUE)	2001

TITRE DE LA THÈSE / TITLE OF THESIS:
CREATING A NEW CITY: MUNICIPAL RESTRUCTURING: THE CASE OF THE OTTAWA TRANSITION BOARD AND THE AMALGAMATION EXERCISE IN THE OTTAWA-CARLETON REGION (2000).

L'auteur permet, par la présente, la consultation et le prêt de cette thèse en conformité avec les règlements établis par le bibliothécaire en chef de l'Université d'Ottawa. L'auteur autorise aussi l'Université d'Ottawa, ses sumccesseurs et cessionnaires, à reproduire cet exemplaire par photograhie ou photocopie pour fins de prêt ou de vente au prix coûtant aux bibliothèques ou aux chercheurs qui en feront la demande.

Les droits de publication par tout autre moyen et pour vente au public demeureront la propriété de l'auteur de la thèse sous réserve des règlements de l'Université d'Ottawa en matière de publication de thèses.

The author hereby permits the consultation and the lending of this thesis pursuant to the regulations established by the Chief Librarian of the University of Ottawa. The author also authorizes the University of Ottawa, its successors and assignees, to make reproductions of this copy by photographic means or by photocopying and to lend or sell such reproductions at cost to libraries and to scholars requesting them.

The right to publish the thesis by other means and to sell it to the public is reserved to the author, subject to the regulations of the University of Ottawa governing the publication of theses.

N.B. LE MASCULIN COMPREND ÉGALEMENT LE FÉMININ

Le 27 novembre 2001

DATE

Marc Gervais

(AUTEUR) SIGNATURE (AUTHOR)



Université d'Ottawa · University of Ottawa

Université d'Ottawa • University of Ottawa

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

GERVAIS, Marc

AUTEUR DE LA THÈSE - AUTHOR OF THESIS

M.A. (Science politique)

GRADE - DEGREE

Département de Science politique

FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

CREATING A NEW CITY

Municipal Restructuring: The Case of the Ottawa Transition Board and the Amalgamation Exercise in the Ottawa-Carleton Region (2000)

Caroline Andrew

DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

L. Cardinal

L. Juillet

J.-M. De Koninck, Ph.D.

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

SIGNATURE

DEAN OF THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

CREATING A NEW CITY

Municipal Restructuring: The Case of the Ottawa Transition Board and the Amalgamation Exercise in the Ottawa-Carleton Region (2000)

by
Marc Gervais
467656

Submitted to
the Faculty of Graduate and Postdoctoral Studies
as a requirement for the course
"Thesis" (POL 7999)

Supervisor: Caroline Andrew, Ph.D.
Department of Political Science

University of Ottawa
November 2, 2001

© Marc Gervais, Ottawa, Canada, 2001

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-66044-3

Canada

ABSTRACT

The issue of municipal amalgamation has been the subject of much political debate in the Province of Ontario. The transition process in the Ottawa-Carleton region was viewed by the Ottawa Transition Board as more than just a simple amalgamation exercise. Rather, it was an attempt to drastically alter the way of doing politics and business at the local level. Although the provincial government and the Board's objectives and desired outcomes in the amalgamation process were clearly articulated at the outset, the path to their attainment was not as straightforward. Factors of scope and ideology as well as the serious time constraints imposed upon the Board contributed to its difficulties. Furthermore, the Board faced a number of effective prevailing democratic constraints. This thesis studies the decision-making process of the Ottawa Transition Board in order to examine its initial project and the degree to which this project was reflected in the Board's final recommendations.

RÉSUMÉ

La question des fusions municipales est depuis un certain temps en Ontario chaudement débattue. L'interprétation de la tâche de fusionner les douze municipalités de la région d'Ottawa-Carleton par les membres du Conseil de transition d'Ottawa révèle une intention autre qu'une simple fusion. Plutôt, le processus de transition a été vu par le gouvernement provincial et par le Conseil comme l'occasion unique de créer une nouvelle organisation municipale. Alors que cet objectif ait été clairement articulé dès le départ, sa concrétisation dans les faits n'a pas été aussi évidente. Des facteurs d'idéologie et d'étendue ainsi que des contraintes temporelles à l'intérieur desquelles le Conseil devait rendre son mandat ont créé des obstacles. De plus, le Conseil a dû transiger avec une opposition de facteurs democratiques efficaces. Cette thèse étudie le processus décisionnel du Conseil de transition d'Ottawa de façon à pouvoir examiner son projet initial et le degré auquel ce projet a été reflété dans les recommendations finales du Conseil.

TABLE OF CONTENTS

ABOUT THE THESIS 1

1. INTRODUCTION 3

1.1 SETTING THE CONTEXT 3
1.2 OTTAWA TRANSITION BOARD 9
1.2.1 The Board's Vision and Scope 13
1.2.2 Narrow Ideological Perspective 17
1.3 DEMOCRATIC PROCESS 18
1.3.1 Public Opinion and Media Scrutiny 19
1.3.2 Concerned Citizens and Community Organizations 21
1.3.3 Elected Representatives and Others 24
1.4 HYPOTHESIS 26

2. REVIEW OF THE LITERATURE 27

2.1 HISTORICAL OVERVIEW 28
2.2 THE CONTEMPORARY DEBATE 33
2.2.1 The Market Model 36
2.2.2 The Participatory State 40
2.3 DISCUSSION 45

3. METHODOLOGY AND RESEARCH PLAN 53

3.1 METHODOLOGY 53
3.2 RESEARCH COMPONENTS 54
3.2.1 Component 1: Define the market model criteria 55
3.2.2 Component 2: Define the participatory model criteria 55
3.2.3 Component 3: Compare Board decisions to definitional elements 56
3.2.4 Component 4: Review of hypothesis 57
3.3 RESEARCH LIMITATIONS 58

4. MAPPING OF FINAL BOARD DETERMINATIONS.....59

4.1 ADMINISTRATIVE STRUCTURE.....60
4.1.1 Recommendations compatible with the market model.....65
4.2 POLITICAL INFRASTRUCTURE.....69
4.2.1 Recommendations compatible with the market model.....72
4.2.2 Recommendations non-compatible with the market model.....73
4.3 GOVERNANCE ISSUES.....74
4.4 CLOSING REMARKS.....84

5. REVIEW OF HYPOTHESIS.....86

BIBLIOGRAPHY.....96

APPENDIX A: Local Government Reform in Ottawa-Carleton

APPENDIX B: Terms of Reference for the Transition Boards

TABLES

TABLE A: Guiding principles issued by province (p. 7-8)

TABLE B: Project Teams (p.16)

TABLE C: Definitional elements of the two models of governance (p.50-51)

ABOUT THE THESIS

This thesis is a case study of the municipal amalgamation exercise undertaken in the Ottawa-Carleton region in the year 2000. On January 26, 2000, Tony Clement, the then Minister of Municipal Affairs and Housing, announced the creation of the Ottawa Transition Board (hereafter the **Board**). It was mandated to ensure the transition from 12 municipalities[1] in the Ottawa-Carleton region to the creation of a new single-tier municipality.[2] This new entity, called the "City of Ottawa", came into existence on January 1, 2001. Shortly after the Board's creation, it announced that the new City organization would embrace an enterprise culture as a means of achieving the desired goals of making local government more efficient and competitive by downsizing the organization and reducing costs. The transition process in the Ottawa-Carleton region was more than just a simple amalgamation exercise of consolidating 11 separate municipal entities and one regional municipality. Rather, it was an attempt to drastically alter the way of doing politics and business at the local level. The transition process was viewed by the Board as well as by the provincial government as a unique opportunity to create a new municipal organization based on their vision of what a municipal government should do and be. The underlying guiding principle of this vision was the adoption of an enterprise culture for the new organization. The intellectual

---

[1]City of Cumberland, City of Gloucester, City of Kanata, City of Nepean, City of Ottawa, City of Vanier, Township of Goulbourn, Township of Osgoode, Township of Rideau, Village of Rockliffe Park, Township of West Carleton and the Regional Municipality of Ottawa-Carleton.

[2]For further information on the historical evolution of municipal reforms in the Ottawa-Carleton region in the past 30 years refer to Appendix A.

basis of the enterprise culture is a market oriented model which basically celebrates the virtues of competitiveness and efficiency.

Although the provincial government and the Board's objectives and desired outcomes in the amalgamation process were clearly articulated at the outset, the path to their attainment was not as straightforward. The Board's task was made more complicated by the interpretation of its mandate, which was ambitious in its scope on the one hand, and ideologically narrow on the other. These factors of scope and ideology as well as the serious time constraints imposed upon the Board (not even one full year) contributed to the creation of tensions with those who were advocating a broader ideological interpretation and those who argued for a far less ambitious transition project. Furthermore, during the implementation of its mandate, the Board faced prevailing democratic constraints in the form of public opinion and media scrutiny, opposing lobbying efforts organized by concerned citizens and community organizations, as well as, diverging views expressed by elected representatives of the existing municipalities and by declared and undeclared candidates of the upcoming municipal elections.

The specific question addressed in this thesis is to what extent the decisions and recommendations made by the Board in the creation of the new City of Ottawa reflect the vision of the entrepreneurial role of municipal government that was defined at the outset of the transition process by the Board.

1. INTRODUCTION

1.1 SETTING THE CONTEXT

In the past fifty years, provincial governments across Canada have introduced a wide variety of reform initiatives aimed at improving the local system of government. The challenges of managing urban growth have been at the root of these various reform efforts across the country. To many observers, by the middle of the 20th century, the local system of government had become increasingly inadequate, largely because of the pressures of industrialization and urbanization. (Tindal and Tindal, 2000) In order to meet these challenges, various governments have introduced changes to local government. Reforms relating to the structure of municipal government have been central. Indeed, Andrew Sancton (1994) argues in *Governing Canada's City Regions: Adapting Form to Function*, that "With few exceptions, the instrument of choice in the past to cope with urban problems in Canada has been structural reform". (p. iii)

This emphasis on structural reform needs to be seen in the context of the two very important and basic roles that local governments play: a representative role and an administrative role, commonly seen in terms of the objectives of access or democracy and that of service or efficiency. Traditionally, the role of local government has been seen mainly as serving the needs of a particular group of people: property owners. In fact, most of the responsibilities of local government can be characterized

as services to property. Municipalities provide an extensive range of services to property and they also provide much of the money to pay for these services from taxes on property. However, municipalities are more than just providers of services. As noted by Andrew Sancton (2000), "They are the democratic mechanisms through which territorially based communities of people govern themselves at a local level". (p.167) Within this political context, inhabitants are able to express, debate and resolve local issues that directly concern them. It then follows that municipal governments exist to serve two primary purposes: "to act as a political mechanism through which a local community can express its collective objectives; and, to provide various services and programs to local residents". (Tindal and Tindal, 2000: 4) In short, the key purpose of local government is to provide people with a wide range of services all the while being democratic in its operations. In doing so, it fulfills its two basic mandates. It is important to emphasize, however, that there can be tension between these two roles. They are not necessarily contradictory but they are not necessarily compatible either. The key feature of municipal government has been, and continues to be, the interaction of its representative and administrative roles. (Tindal and Tindal, 2000: 4) The question has always been how to sustain local democracy and remain responsive to community needs while providing for the efficient delivery of services. Since the end of the Second World War, reform initiatives have been preoccupied, for the most part, with the service delivery role of local government and have paid scant attention to their representative and political roles. (Tindal and Tindal, 2000)

Most of the major attempts at municipal reform in Canada can be narrowed down to two types of reforms: 1-municipal consolidation; and, 2- the use of intermunicipal boards and joint servicing agreements. The first, municipal consolidation, includes amalgamation, annexation and the creation of regional governments. It includes any kind of municipal restructuring which involves both boundary changes and the creation of larger units for at least some local functions. On the other hand, the second broad initiative, the use of intermunicipal boards and joint servicing agreements, occurs without boundary changes or the creation of new second tier governments. Allan O'Brien (1993), author of *Municipal Consolidation in Canada and its Alternatives*, defines these efforts in the following terms: "This usually involves some intermunicipal agreement under which two or more municipal units decide to provide planning or some service on a regional basis (...) In some cases the joint activity may involve more than one service". (p. 4)

In Ontario, since the mid-1990s, the structural reform of choice has been amalgamation. "Amalgamation" is the process by which the total number of municipalities is reduced by merging two or more municipalities to create a new municipality. (Sancton, 1993: 11 and 30) A review of the literature on municipal reform reveals that the rationale behind this type of reform initiative rests primarily on the following arguments: the virtues of area-wide planning, the need for equity in access to taxation revenues, and potential economies of scale. (Sancton, 1994; also see Colton, 1980; Sancton, 1993; and, Vojnovic, 2000) The underlying belief behind these arguments is that bigger is better. In direct opposition to this argument, public choice

theorists argue that smaller and many is better, that the fragmentation of municipalities is a good thing. They believe that local governments do not need to be producers of public services, but only arrangers. They also believe that more municipal competition is the only way to ensure choice, accountability and efficiency. (Vojnovic, 2000: 4) In the middle of this debate are those who believe that there is such a thing as an "optimal municipal size". According to Vojnovic (2000), in the context of local government, an "optimal municipal size" is "one which covers a large enough area, and produces enough service output, to minimize the average cost of production". (p. 2) Despite studies raising serious questions about size, one thing is clear, there still is a push for bigger is better.

Another rationale for amalgamation relates more specifically to the ideology driving the municipal reform initiatives pursued by the Conservative government of Mike Harris in the province of Ontario since its election in June 1995. The reasons for these policies are clearly ideological and short term. They can be best summarized by the belief that less government is better and cheaper. (Graham, Maslove and Phillips, 2000: 24) Underlying this narrow view of government is the century-old idea that government must operate more like the private sector. This idea has been aired in Canada in one version or another since the turn of the century. Claims that people are over-governed and desire less government and fewer politicians are back in vogue. As Tindal and Tindal (2000) observe, the Conservatives in Ontario "premise their restructuring on the notion that local politicians and staff are 'the problem', that there are too many of them, they are wasteful in their practices, their operations are

inefficient, they tax too readily, and they spend irresponsibly". (Tindal and Tindal, 2000: 176) In short, the Neo-Conservative agenda can best be summarized by less government, fewer politicians and less taxes. In moving second reading of the *Fewer Politicians Act, 1999*, the legislation that allowed for amalgamation in four Ontario regions (Ottawa-Carleton, Hamilton-Wentworth, Sudbury, and Haldimand-Norfolk), the Minister of Municipal Affairs and Housing emphasized the virtues of reducing the number of municipal politicians and of saving money. (Sancton, 2000: 157) Indeed, the most common justification for consolidation is that it will reduce costs, thereby making the municipal system as a whole more efficient.

In line with this rationale, the Ministry of Municipal Affairs and Housing issued some guiding principles (under section 25.4 of the *Municipal Act)* to be considered by municipalities when developing restructuring proposals. Table A highlights these guiding principles.[3]

TABLE A: Guiding principles issued by province

Less Government	• fewer municipalities • reduced municipal spending • fewer elected representatives

[3]The following information was taken from: Government of Ontario. 1996. *A Guide to Municipal Restructuring*. Toronto: Ministry of Municipal Affairs and Housing.

Effective Representation System	• accessible • accountable • representative of population served • size that permits efficient priority-setting
Best Value for Taxpayer's Dollar	• efficient service delivery • reduced duplication and overlap • ability to capture the costs and benefits of municipal services within the same jurisdiction • clear delineation of responsibilities between local government bodies
Ability to Provide Municipal Services From Municipal Resources	• local self reliance to finance municipal services • ability to retain and attract highly qualified staff
Supportive Environment for Job Creation, Investment and Economic Growth	• streamlined, simplified government • high quality services at the lowest possible cost

The government's rationale with regards to the amalgamation exercise in the Ottawa-Carleton region was made clear by the Minister of Municipal Affairs and Housing when he announced the identity of Board members at a news conference in Ottawa on January 26, 2000: "The government established the new City of Ottawa to reduce the number of politicians and improve local government, making it simpler, more efficient, and more accountable. The goal is fewer politicians and lower taxes."[4]

---

[4]Government of Ontario. 2000. *News Release Communiqué - January 26, 2000.* Toronto: Ministry of Municipal Affairs and Housing.

1.2 THE OTTAWA TRANSITION BOARD

The task of developing, adopting and implementing the plan for a new amalgamated city based on the province's guiding principles outlined in Table A was given to a seven-person Board appointed by the province. The Board's role, as specified in the Terms of Reference, was to ensure a "smooth, seamless transition" to the new City and to recommend for the new City and its local boards, a year 2001 budget, and to forecast year 2002 and 2003 expenditures and revenues.[5] Board recommendations were to be made in the context of Special Advisor Glen Shortliffe's findings about potential tax savings.[6]

The Board was made up of five men and two women. Amongst the Board members, five were English-speaking and two were French-speaking. The Board was chaired by former provincial Cabinet minister Claude Bennett. A principal in The Strategies Group, Claude Bennett has had many years of experience in provincial and municipal government. He was first elected to Ottawa city council in 1961, and was a member of the first Ottawa-Carleton Regional government between 1968 and 1972 and

---

[5]For a copy of the Terms of Reference established by the provincial government for Transition Boards refer to Appendix B.

[6]Glen Shortliffe was appointed by the Government of Ontario. His mandate was to recommend reforms in municipal government in the Ottawa-Carleton region which would lower taxes, improve services, reduce bureaucracy, clarify lines of responsibility, and foster greater municipal accountability. His report - *Report of the Special Advisor: Local Government Reform in the Regional Municipality of Ottawa-Carleton* (Shortliffe Report) recommended a one city - one tier government with community satellite offices. For further information refer to Appendix A.

was on the regional executive committee from 1970 to 1972. He was elected to the Ottawa board of control in 1970 and served as senior controller and acting mayor. He was elected to the provincial parliament in 1971, and served as MPP for Ottawa South until 1987. He held several provincial Cabinet posts, including Minister of Municipal Affairs and Housing. From 1990 to 1995, Mr. Bennett served as chairman of the board of directors of the Canada Mortgage and Housing Corporation.

The other Board members included former Osgoode Township mayor Albert Bouwers, former member of the Citizen's Panel on Restructuring in Ottawa-Carleton Camille Guilbault, chartered accountant David Muir, former Ottawa councillor Edward Mulkins, Nepean Hydro Commission chair Kathy Greiner and Gloucester city manager Pierre Tessier.

The Board members met for the first time in late January and held their first public meeting in early February. On March 10 and 11, 2000, the Board held a 2-day planning session. This session served to develop a shared understanding of the Board's mandate and of the work that had to be accomplished to create the new City of Ottawa. A document entitled *Project Charter - City of Ottawa Transition* summarizes the consensus reached by Board members during this initial planning session.

The document states that the Board's mission is to create the foundation for a new Ottawa through the restructuring of twelve existing municipalities into one single government. It also states that the Board will have been successful at the end of its

mandate if it has ensured the following: the implementation of a downsized organization with the required capacity (i.e., the appropriate number of resources with the skills/competencies) to function efficiently and effectively, and the achievement of a seamless transition with no service disruption. Furthermore, the document states that the Board will operate under the following guiding principles and values: equal opportunity and respectful treatment of employees, transparency, rationalization and implementation of best practices and an innovative service delivery model rather than a simple amalgamation, and, respect of differences in requirements of various communities (particularly rural vs. urban).

This document clearly confirms the Board's intent of achieving its goals by restructuring the twelve municipal organizations in the Ottawa-Carleton region, rather than by simply amalgamating them. It is important to note that at the time of amalgamation, the Regional Municipality of Ottawa-Carleton was delivering approximately 80% of all the local services in the Ottawa-Carleton region. (Shortliffe, 1999) This document further elucidates the Board's intent of resting this restructuring exercise on downsizing the organization and making it leaner in order to reduce costs.

To achieve these goals, the Board announced in April 2000 that the new City organization would embrace an enterprise culture for its management and its operation. According to an April 10, 2000, press communiqué released by the Board, the concept builds on a number of accountability models to formalize a culture of citizen centred, innovative, cost effective, performance based service delivery, in each and every

functional area of city operations and administration. The communiqué goes on to state: "The Enterprise Culture is not privatization of government service delivery, nor is it an assurance of continued in-house delivery. Rather it is a concept that requires that the New City administration apply principles similar to those used by private enterprise in looking at service delivery."[7]

In adopting an operational model based on business principles the Board made its intent clear: "the goal is to go beyond establishing a new structure, to establishing a thinking, questioning and creative organization that is able to identify desired community outcomes and apply innovative best value solutions."[8] The project teams responsible for the service restructuring and integration were clearly mandated by the Board with the responsibility of ensuring that the change from existing organizational cultures be initiated and that a solid foundation be put into place "to permit a new enterprise culture to become a permanent part of the new City of Ottawa, in all areas and at all levels."[9]

Mr. Claude Bennett, chairman of the Board, was instrumental in setting the Board's agenda and articulating its new entrepreneurial vision for the new City. His local political experience, his long-standing involvement in Ontario politics and his

---

[7]Ottawa Transition Board. 2000. *Press Communiqué - April 10.* Ottawa.

[8]*Ibid.*

[9]*Ibid.*

intimate association to the conservative movement all contributed to his considerable influence on the process, on the debates and on the final decisions and recommendations made by the Board. His influence was most apparent in the interpretation of the Board's mandate, which, as previously stated, was ambitious in its scope on the one hand, and ideologically narrow on the other.

1.2.1 The Board's Vision and Scope

At the very beginning of the transition process, Mr. Bennett clearly articulated his vision of his role and that of the Board. He believed that the transition project was truly a significant and historical moment in the history of the City of Ottawa. Consequently, he recognized the importance of the legacy that he and all those involved in the transition process were leaving to the people of Ottawa. On many occasions, Mr. Bennett likened himself to a builder, a creator of a new entity. In fact, in his opening remarks at the Board's first public meeting on February 14, 2000, he invoked the spirits of such pioneers and local heroes as Philemon Wright and Colonel John By, the founding fathers of the cities of Hull and Ottawa. He declared: "Our community has prospered beyond the dreams of those first residents, and today we are standing at yet another watershed. As we begin our job, the Ottawa Transition Board recognizes the significance of this moment in time and the potential it holds for this community's next 200 years."[10]

---

[10]Ottawa Transition Board. 2000. *Chairman's Opening Address - February 14.* Ottawa.

This desire of viewing the transition as a historical occasion and opportunity to create a new municipal entity was vigourously promoted by the Board. For example, a project team was charged with the task of developing new visual identifiers such as a logo, a slogan, a flag, a coat of arms and a motto for the new City. The Board believed that the outcome of the work undertaken by this project team would be essential to the development of the new enterprise that would be the new City of Ottawa. Furthermore, in its quest to distinguish the new from the old, the Board insisted on changing the address of the proposed location of the new City Hall, the then current headquarters of the Regional Municipality of Ottawa-Carleton, from 111 Elgin Street to 111 Laurier Street East.

The vision and scope of this restructuring exercise, as conceived by the Board, was so expansive that it required the creation and support of a very complex internal administrative structure which at times was difficult to handle due to the short time frame of the transition process. Initially, the Board established approximately 50 transition project teams to assist in the task of amalgamating and restructuring the 12 existing municipal administrations into one entity by January 2001. Each project team was given specific objectives and asked to define the steps toward their achievement. The duration of transition projects varied according to the scope and size of their mandate.

The transition projects fell into three broad categories: service restructuring, integration and stand-alone projects. Service restructuring projects aimed at

amalgamating, rationalizing, and restructuring the way services would be delivered in the new City. These projects covered the full range of services provided by a municipal government including corporate services. There were 24 service restructuring projects established at the outset, one per service. In general, integration projects mainly touched horizontal preoccupations that run across service areas. All other projects fell into the third category, as highlighted in Table B on the following page.

TABLE B: Project Teams[11]

Service Restructuring	Integration	Stand-alone
• Recreation, Parks, Culture • Social Services • Libraries • Planning • Fire Services • Police Services • Land Ambulance and Integrated EMS • Emergency Measures • Cross Utility Issues • Public Works • Transit • Corporate Secretariat Services • Financial and Material Management • Supply Management • Human Resources • Informatics Services • Legal Services • Real Property Assets Management • Strategic Planning, Audit and Performance Measurement Services • Communications / Public Affairs • By-law, Licensing and Enforcement • Records, Information Management • Economic Development and Tourism (incl. Industrial Parks) • Integration of Information Technology	• Point of Service (enterprise model, service levels, satellite offices, e-business, etc.) • Rural Issues • Bilingualism Policy • Voluntarism Policy • Structure and Staffing • Labour Relations • Workplace Adjustment / Transition Support • Communication Plan and Implementation • Facilities Plan and Implementation • Integration of Information Technology Environments • Business Plans and Budgets for 2001/02/03	• Political Infrastructure • Elections Process • Ongoing Financial Monitoring • Visual Identity • Taxi Industry • Street Numbering and Naming • Assets / liabilities • Insurance • Banking, Treasury and Investment • Ottawa Development Issues (high tech growth) • Superannuation Fund

[11]The following information was taken from: Marc Gervais. 2001. *Public Record of the Ottawa Transition Board*, Publication of the Ottawa Transition Board in collaboration with the University of Ottawa, Ottawa, Ontario.

As time went by, the number of project teams grew from 50 to 80. This was largely due to the number of IM/IT system integration initiatives that were required. Also, many more stand-alone project teams were created (e.g., Environment, Youth, Access and Diversity). All but 4 project teams were headed by municipal employees from the former Ottawa-Carleton municipalities. The vast majority of the teams were also made up of municipal employees, and on occasion, members of the public and consultants. In all, over 1,000 people participated in the amalgamation exercise. The number and scope of the teams faithfully reflected the Board's interpretation of its mandate. By adopting such a broad interpretation, the Board ultimately made its task more complicated. This also contributed to the creation of tensions with those in the community who would have preferred a far less ambitious transition project.

1.2.2 Narrow Ideological Perspective

Throughout the process, Mr. Bennett challenged municipal employees involved in the transition to go beyond the traditional ways of thinking and acting and to propose alternative service delivery models that would reduce costs and make the municipal government more efficient. He encouraged them to "Think outside of the box". During a presentation to a community group, Mr. Bennett commented that municipal governments should re-think their involvement in the delivery of services that can be found advertised in the yellow pages of the local telephone directory. In other words, according to Mr. Bennett, it was legitimate for the Board to be addressing the following question: Should a municipal government be offering services that are already being

offered by the private sector? Underlying this questioning of the role of municipal governments was the desire to see the adoption of business approaches coming from the private sector. As has been previously stated, the Board's answer to this question was the adoption of an enterprise culture for the new City of Ottawa. By adopting such an ideologically narrow operational model for the new City, the Board ultimately rendered its task more complicated as support for a new city culture based on business principles was far from unanimous amongst the citizenry of Ottawa-Carleton. In fact, it proved to be a major point of contention with those advocating a broader ideological interpretation.

1.3 DEMOCRATIC PROCESS

Given this lengthy and focussed attempt to emphasize the significance and imperative necessity of implementing an enterprise culture within the new City, one would expect the Board's recommendations and decisions to clearly reflect this desired outcome. However, one is struck by the number of examples and counter indications of the ultimate success of the effort. The understanding of the causes underlying the presence of outcomes that are incompatible with the Board's initial vision lies in the assessment of a variety of factors not only related to time and mandate (particularly scope and ideology) as just mentioned, but also to intervening democratic processes.

People need and want to be involved in the decisions that directly affect their

lives, whether in the public or private realms. (Phillips, 1991) At the local level, decisions made by municipal governments impact in an immediate and significant way on the lives of their citizens. The decisions and recommendations made by the Board will undoubtedly influence the future direction of the new City of Ottawa. The outcomes of the transition process may potentially determine the level and quality of services offered by the new municipality as well as the relationship between the citizens of the new City and their elected officials. In short, at stake is the quality of life enjoyed by the citizens of the former Ottawa-Carleton region.

Given this, it is not surprising that the media, community organizations, as well as members of the general public took a keen interest in the transition process. Our study will reveal that the relationship between the Board and these groups was problematic. It is apparent there existed tensions between the Board and those in the community who did not necessarily share the entrepreneurial vision of municipal government promoted by the Board. The process of any given political change lies in the interplay of opposing forces. As a result, the transition exercise can be studied in the context of a clash of opposing visions of municipal governance and of the instruments of expression and actualization of these visions.

1.3.1 Public Opinion and Media Scrutiny

Throughout its mandate, the Board's legitimacy was put into question. In fact, its very existence was viewed by many as undemocratic. This perception stemmed from

the fact that the Board members were appointed by the provincial government and not elected by the people. Mr. Bennett himself acknowledged as much when he was quoted as saying to a journalist that "many of the complaints about the Transition Board arise from the fact the members were appointed, not elected".[12] Also, many people perceived the Board's decision making procedures as being secretive and dictatorial, and, in the end, unaccountable. A review of the local newspaper articles leading up to, and, at the time of, the announcement of the creation of the Board, as well as throughout the Board's mandate, highlights a general malaise felt by many people at this intrusion on the democratic process. However, it should be noted that this feeling was not unanimous. For example, the debate of whether or not this process would ultimately lead to better results was arguably not as clear cut. Some people believed that the controversial business of restructuring should rightly be at arm's length from elected representatives perceived as being too invested in the promotion of their self-interest. Furthermore, some people accepted the notion that the Board's legitimacy came from the provincial government, a duly elected and accountable body. Nevertheless, the fact that democracy had been suspended at the regional and municipal levels for almost one year, until the creation of the new unified City of Ottawa, did not sit well with the vast majority of citizens of the Ottawa-Carleton region. This ultimately lead to the creation of tensions between the Board and those who did not recognize its legitimacy or agree with its vision for the new City.

---

[12]Jacki Leroux, "Critics fear moves too far, too fast to be good", *The Ottawa Sun*, July 16, 2000, p.19.

This opposition to the Board on the grounds of its perceived undemocratic and illegitimate nature was never articulated in a formal theoretical way by its critics. Rather, it was expressed through public discourse in the media via journalists, political commentators, editorialists, and letters-to-the-editor written by concerned citizens. Here are but a few examples of the headings of some of the articles and editorials which appeared in the local newspapers on this topic: "New city doesn't need czars" (*The Ottawa Citizen*, November 30, 1999); "No, your vote doesn't count" *(The Ottawa Citizen*, January 28, 2000); "Board has little time for public" (*The Ottawa Sun*, March 8, 2000); "Democratic advice for our dictators" (*The Ottawa Citizen*, February 16, 2000); "So this is how a dictatorship reaches out" (*The Ottawa Citizen*, March 15, 2000); "*Un comité de vigilance accuse le Conseil de transition de manque de transparence*" (*Le Droit*, April 11, 2000); "Transition board undemocratic, critics charge" (webposted on *ottawa.cbc.ca*, May 9, 2000); and, "Board 'boys club' blasted" *(The Ottawa Sun*, July 23, 2000). These newspaper headings highlight the uneasiness felt by many at the perceived undemocratic nature of the Board. The dispelling of such negative perceptions proved to be a constant challenge for the Board throughout its mandate.

1.3.2 Concerned Citizens and Community Organizations

As stated, opposition to the Board was never articulated in a formal theoretical way but, certain interest groups did mobilize to formally express their concerns. They did so through the various avenues available to them, for example, public Board

meetings and public consultation sessions, as well as through other modes of communication, such as writing letters and sending e-mails and faxes.

Throughout its mandate, the Board met as a full Board on the second Monday of each month, and, when circumstances warranted, at the call of the Chair. These meetings were open to the public. One hour of every full public meeting was dedicated to public input. Individuals were asked to file their requests to appear before the Board in advance with the Clerk (of the Board). Each person was limited to one appearance before the full Board and allotted five minutes to make his/her oral presentation. To supplement oral presentations, individuals and delegations were encouraged to put their concerns in writing and to file materials with the Board. People could submit as many briefs, letters, and presentations as they wished.

As for each individual project team, the type and scope of public consultation mechanisms used varied depending on its particular mandate. Some project teams established challenge teams to review their draft reports. These challenge teams consisted of outside experts and representatives of various stakeholder groups. Other project teams established advisory committees made up of volunteers from the private and public sectors to assist them with their work (e.g., Taxi Industry, Youth, Visual Identity, etc.). In addition to this, many project teams held public information and consultation sessions (e.g., Street Naming, Voluntary Sector, Budget Plans, etc.). In these sessions, project members would inform the public of their team's work and citizens were given the opportunity to express their concerns, ideas, and suggestions

directly to the project team members.

The most intensive period of public consultations organized by the Board occurred on August 1, 2 and 3, 2000. These public consultation meetings were held to seek public input with regards to the twenty-four high level service delivery models that were presented to the Board by the various service restructuring teams during the July 25, 26 and 27 public information sessions. These public consultation meetings were viewed by the Board as an opportunity for community groups, business associations, employees and concerned citizens to provide the Board with their views. Those interested in making a presentation to the Board were asked to call and register in advance (on a first come, first served basis). Presenters were given five minutes to make their presentation. There was a brief period reserved for questions from Board members. The Board heard from approximately 110 presenters during the three-day public consultation meetings. According to Board documents, roughly 30% of these presenters were concerned citizens and 70% were representatives of various organizations and associations.[13] Amongst the most frequent topics addressed in these submissions we find the following key themes. The number in parentheses represents the number of presentations made per topic: environment (16), health and social services (15), library (8), planning and development issues (8), transit and transportation (8), services for the disabled (6), diversity/multiculturalism (5), and, volunteerism (4). The other topics included citizen participation and access, use of

[13]Ottawa Transition Board. *Public Consultation Sessions - August 1-3, 2000.* Ottawa.

pesticides, service levels, labour issues, bilingualism, affordable housing, alternative service delivery (ASD), child care, heritage, youth issues, and finally, fire and police services.

One cannot help but notice a certain incompatibility between the priorities articulated by the Board and the province at the outset as well as throughout the transition process (i.e., fewer politicians, lower taxes, less government, in short, a business agenda) and the issues raised by the residents of Ottawa-Carleton (i.e., the environment, health and social issues, quality of life issues, in short, a people's agenda). As we will see, concerned citizens and community organizations were successful in bringing these issues to the forefront and ensuring that they be given consideration by the Board during the transition exercise.

1.3.3 Elected Representatives and Others

Finally, many elected representatives (i.e., mayors and councillors) of the existing municipalities as well as individuals considering a bid for a seat on the new council also weighed into the debate from time to time. These interjections were made more often then not with the intent of denouncing a Board pronouncement or decision and rarely done in support of them. Amongst the most articulate and vocal critics of the Board were high profile RMOC councillors Alex Munter and Diane Holmes, City of Ottawa councillors Stéphane Émard-Chabot and Diane Deans, and former local and provincial politician Alex Cullen.

However, the most direct and obvious example of an organized front against the Board was mounted by an eleven-member watchdog group (Boardwatch) formed to combat what members called the "secretive and unaccountable" actions of the Board. A newspaper article pertaining to Boardwatch stated: "Boardwatch has echoed the same complaints as other interest groups - that Chair Claude Bennett and his crew are undemocratic, that they don't have enough time to do the job right, that they're exceeding their mandate".[14] Boardwatch's website stated it was committed to keeping the public better informed of the transition team's activities, exposing undemocratic actions and analysing the effects of the board's decisions. Boardwatch's membership included current politicians, former politicians (including a former mayor of Ottawa) as well as social and community activists. Consequently, these aforementioned elected representatives and would-be councillors along with Boardwatch played an important role in ensuring that the Board remain responsive to public opinion and needs.

In the end, it is apparent that the activities and pronouncements of all of these concerned citizens, community organizations and elected representatives, as well as the media scrutiny provided by local journalists, influenced public opinion, which in turn, impacted upon the Board's final decisions and recommendations.

---

[14]Jacki Leroux, *op.cit.*

1.4 HYPOTHESIS

Given these issues of democratic process, the serious time constraints imposed upon the Board, and the Board's ambitious and ideologically narrow interpretation of its mandate, we believe we will find variations in the Board's final determinations[15] which may reflect lesser than ideal manifestations of the entrepreneurial role of municipal government envisioned by the Board and the province for the new City of Ottawa.

[15]General term encompassing the Board's decisions, recommendations, activities, and pronouncements.

2. REVIEW OF THE LITERATURE

The municipal amalgamation exercise undertaken in the Ottawa-Carleton region raises some fundamental issues, such as: the redefinition of the role of local government; the practice and the function of democracy at the local level; and, the redefinition of the citizen's role and of the relative influence of various constituencies within a defined political entity.

This chapter attempts to define these issues in order to better understand the historical and theoretical context in which the amalgamation exercise occurred. The current debate and proposed changes to municipal governance in Ontario are couched in a tradition of change in governance issues at the municipal level. By providing an historical overview of the development of local government in Canada with an emphasis on the practice and the function of democracy at the local level, and, by presenting the essential elements of the contemporary debate surrounding the role of government we hope to give some perspective to the current process. The first part, the brief historical overview, will serve as an introduction to the second part, the presentation of the key components of the contemporary debate.

Central to the understanding of our presentation is the belief that democracy matters as much in the state as it matters, if not more, elsewhere. By definition, democracy necessitates participation from people, including citizens, elected

representatives, and stakeholders (i.e., employees). Although the issues of democracy may differ between these groups or may even be contradictory in theoretical terms, they are the same in practical terms, as argued by Ann Phillips' (1991) *Engendering Democracy*. "When the issues of democracy are conceived in terms of a general rubric of participation (how much is possible or desirable? How often should we participate and where?) this blurs the contrast [between them]." (Phillips, 1991: 18) For our discussion, these contradictions are not important as participation is the unifying factor, thus allowing us to study them simultaneously.

The chapter will therefore end with a discussion on two opposing governance models which incorporate these various elements of participation involving citizens, elected representatives and employees. This framework will also allow us to assess in Chapter 4 the extent to which the decisions and recommendations made by the Board reflect its entrepreneurial vision of local government.

2.1 HISTORICAL OVERVIEW

As stated in the introduction, the key feature of municipal government has been, and continues to be, the interaction of its representative and administrative roles. The historical starting point for dealing with this "democracy-efficiency trade-off" is the Baldwin or Municipal Act of 1849.

The Baldwin or Municipal Act of 1849 which was first introduced in Upper Canada (Ontario) in 1843, but was not passed until 1849, has had a seminal influence on the development of local government to the present. The Baldwin Act is significant for many reasons, chief amongst them being that it was the first piece of legislation to create a uniform system of municipal government over an entire province which was later used as a model by other provincial governments in the organization and development of their municipal institutions. (Graham, Phillips, and Maslove, 1998: 46)

More importantly for our discussion, however, is the fact that the Baldwin Act restricted the local right to vote to property owners. According to Graham, Phillips, and Maslove (1998), "this reinforced the orientation of municipal government to providing services to property and has had a lingering impact on participation in local life, even after the franchise was widened." (p. 46) Furthermore, the Baldwin Act embedded the principle that the municipal councils should receive delegated authority from the provincial legislature and are subject to its control. This led to the popular description of municipal governments as mere "creatures of the provinces". This notion was further reinforced and institutionalized with the passing of the British North America Act in 1867, where the local governments lack any formal legal recognition of their own. In short, local government only exists to the extent that the provincial government sees fit to provide for it. According to Isin (1992), this was to have lasting effects. This meant, still according to Isin, that the modern municipal corporation would have two essential characteristics:

> First, it is created at the pleasure of the legislature, and while the province may obtain the consent of the people of the affected locality, it need not. The act of incorporation is not a contract between the legislature and the inhabitants. The province can erect, change, divide and abolish a corporation at its pleasure and as it deems appropriate. Second, the authority conferred on the modern city is not local in nature but derives from the province. (p. 2)

Such an arrangement is obviously not compatible with the concept of municipal government as the expression of the political will of the community. Notwithstanding this fact, many prominent political personalities and thinkers have emphasized the democratic features of municipal government, starting with Robert Baldwin himself.

According to J. H. Aitchison (1949), the twin objectives of Robert Baldwin's political career were responsible government at the centre and "home rule" for municipal affairs. With regards to the latter, Baldwin was thoroughly convinced that "the people should manage their own affairs". As noted by Aitchison (1949), "In the sphere of local government this meant for him not merely elective representative institutions, but also freedom for such institutions to function without hampering statutory restrictions or central administrative control." (p.107)

Baldwin also argued that Township Councils would provide a school for practical statesmen and teach the importance of civil institutions. (Aitchison, 1949: 121) This notion was shared by others. For example, Tindal and Tindal (2000) describe J. S. Mill's position that

> municipal government constituted a training ground for democracy, wherein elected representatives would 'learn the ropes' before going on to

> service at a more senior level, and local citizens would learn about exercising their democratic rights in the context of issues which were relatively simple and understandable. (p. 5)

For his part, Lord Durham was struck by the lack of municipal institutions in the colony in the 1830s. He expressed concern that "the people receive no training in those habits of self-government which are indispensable to enable them rightly to exercise the power of choosing representatives in parliament".[16]

In *Democracy in America*, Alexis de Tocqueville stated that "municipal institutions constitute the strength of free nations".[17] Tocqueville apparently saw municipal government's democratic role in a direct, more fundamental light.

K. G. Crawford (1954), author of the first comprehensive modern study of municipal government in Canada, "saw municipal government as far from just a training ground, but as the level at which the democratic ideal was most likely to be fulfilled". (Tindal and Tindal, 2000: 6)

However, not everyone shared this positive view of municipalities. Crawford's contemporary, Georges Langrod (1953), for example, viewed municipal governments as "but a technical arrangement within the mechanisms of the administrative system, a

---

[16]Quoted in Isin, 1992: 132.

[17]Quoted in Tindal and Tindal, 2000: 6.

structural and functional detail..."[18] According to Tindal and Tindal (2000: 6), Langrod not only rejected the assumption that municipal governments are vital to democracy, he also contended that they could be contrary to the democratic process.

A major setback for municipal governments as instruments of local democracy actually occurred 100 years ago during the turn of the century reform movement. These reformers believed that city politics were corrupt, "subject to the evils of political gamesmanship, and increasingly pulled away from the essential mission, namely, the efficient administration of services to property and public protection". (Weaver, 1984: 56) Their reform agenda focussed on reform of the political and administrative structures of urban governments to make them apolitical. "Efficiency was seen as the foundation of a sound local economy, the engine of good urban life." (Weaver, 1984: 56)

This emphasis on efficiency and local government's administrative role has persisted throughout most of the 20th century and is very much central to the amalgamation projects recently undertaken by the government of Ontario. In fact, as stated previously in the introduction, a review of the literature on major municipal reforms since the Second World War reveals that for the most part, the adopted reforms have dealt with the service delivery role of local government and have paid scant attention to its representative and political roles. (Tindal and Tindal, 2000)

---

[18] *Ibid.*

Amalgamation projects are seen as a continuation of this tendency.

As a result, the argument that people are over-governed and desire less government and fewer politicians along with the claim that the private sector is the appropriate "efficient" model that governing administrations should emulate are once more at the forefront of the political debate.

2.2 THE CONTEMPORARY DEBATE

Our research is situated in the broader context of the current debate on the contemporary visions of state and governance and the role of the citizen therein. At the heart of the debate is what government does, why it does it, and how. It comprises the objects and purposes of government, the functions of government, and the instruments of governance. (Johnson, 1993)

By the late 1970s, virtually all governments were under stress. (Campbell,1983) Since then, the very nature of government has been subjected to severe questioning both in Canada and elsewhere. The classical view of governance based on the traditional model of the civil service in government with the Keynesian policy regime as its socio-economic underpinning have been judged by many to be obsolete, dysfunctional, and out of step with contemporary reality. (Peters, 1993) In an effort to modernize governments and make them more effective and efficient, many new models have emerged. What has occurred is nothing short of a paradigmatic shift.

In *Government By Market*, Peter Self (1993) summarizes the new paradigm that has become dominant in Western democracies in the following terms:

> This paradigm holds that governments should in general do less; that they should reduce or relinquish their previous responsibilities for maintaining full employment and a comprehensive system of state welfare; that they should privatise public services or their delivery wherever practicable; and, that they should reform their own operations in accordance with market concepts of competition and efficiency. These beliefs in 'government by the market' rest upon propositions that the market system is inherently a better method for satisfying human wants and aspirations than recourse to government, and that the political process is subject to numerous imperfections and distortions. (p. ix)

This shift in emphasis is primarily attributed to the fundamental changes affecting the world economy and the perceived role of government therein. Terms like post-fordism, information technology, internationalization of economic activity, new global economy, international economic forces and globalization have been used to explain this new phenomenon. Initiatives to restructure government parallel and reflect developments which have occurred at the senior levels of government, in Canada, and even more so in other jurisdictions such as Britain, Australia, New Zealand, and the United States. As noted by Aucoin (1995):

> In Canada, as elsewhere, the common wisdom is that the public sector should be reduced, both in size and cost. Governments are facing the need to make choices between the programs they consider essential to administer themselves and those that may be transferred to other organizations or discontinued. A new vision of how government should function has recently emerged, notably with greater emphasis on performance and results. (p. i)

At the municipal level, these changes brought about a new wave of amalgamations in the 1990s, promoted in large part by provincial governments who felt that they needed larger municipalities to which they could download responsibilities and costs. That downloading, in turn, has increased the fiscal squeeze on municipalities, prompting them to embrace a variety of techniques designed to make them more business-like. (Tindal and Tindal, 2000: 1) Consequently, the role of municipal government as a business corporation rather than as an instrument of local democracy has dominated debate in recent years.

In this light, B. Guy Peters (1993), in *The Public Service, The Changing State and Governance,* examined different models that have emerged in response to the perceived shortcomings of the traditional models of governance which have been judged to be no longer compatible with the new socio-economic world order. Amongst the four models examined by Peters (1993), the two main visions of governance which appear to be the most popular alternatives to the traditional model of administration are the market model and the participatory model. The first deals primarily with service delivery and can be associated with the administrative role of local government, commonly known as "service or efficiency". The second deals primarily with the role of the citizen in society and can be linked to the representative role of local government, commonly known as "access or democracy". As with the two basic roles of local government, these two ideologically different models are not necessarily contradictory or opposed, but they are not necessarily compatible either.

A further study of the intellectual origins and principle tenets of these two models will help us to situate the current amalgamation projects within a theoretical model.

2.2.1 The Market Model

The intellectual basis for the market model which basically celebrates or defends the virtues of competitiveness comes from 1- public choice thought, 2- the new public management, and, 3- neo-conservative ideology.

Peters (1993) illustrates how early public choice theorists such as Niskanen, Tullock, Moe, Ostrom, and Bender analysed the failings of conventional bureaucracies.

> They argued that because of the self-interest of the members of the organizations, especially the "bureau chiefs" at the apex, public bureaucracies tended to expand at an unjustifiable rate and to charge their sponsors (read legislatures) too much for the services produced. It was argued that the permanence of bureaucrats and their monopoly of information put them at a competitive advantage when dealing with the legislature. The root of any failings in the public sector, as seen from this perspective, is the self-interest of bureaucrats. (Peters, 1993: 7)

Public choice thought concentrates on the problems and limitations of the democratic political process. It has been defined as "the economic study of nonmarket decision making". (Mueller, 1979: 1) It is a theoretical perspective based on the application of economic ideas to political structures and processes. According to Michael Keating (1995):

> Its central tenets are based upon individualist premises and utilitarian philosophy. That is, it holds that the unit of analysis is the self-interested individual and that the public good is no more that the aggregate of individuals' aspirations. Individuals define their own self-interest and pursue it. Democracy is seen less as a system for taking collective decisions than as a mechanism for allowing individuals maximum scope for choice. So public choice theorists support local government structures which approximate as closely as possible to markets, allowing individuals to make choices about services, taxes and other policies. Efficiency is seen as best promoted by competition, among individuals and among service providing units. Since bureaucrats are also seen as self-interested utility-maximizers (Niskanen, 1973), it is important to subject them to competitive discipline by allowing individuals and communities to shop around for the best services. Development is best promoted by encouraging competition among places and allowing capital to find its most profitable location undistorted by government regulation. (p.123)

Within the public choice perspective, local government is analogous to firms and citizens, to consumers. Ironically, public choice theorists argue that smaller and more numerous are better, that the fragmentation of municipalities is desirable. As stated previously, they believe that local governments do not need to be producers of public services, but only arrangers. They also believe that more municipal competition is the only way to ensure choice, accountability and efficiency. (Vojnovic, 2000: 4)

Second, the term "new public management" was first coined by Christopher Hood in 1991. Proponents of the new public management maintain that adequate managerial structures provide high quality services that citizens value; increase the autonomy of public managers, especially from central control agencies; measure and reward organizations and individuals on the basis of whether or not they meet demanding targets; and, appreciate the virtues of competition and of keeping an open

mind on whether public purposes should be performed by the private sector rather than the public sector.[19] According to Tindal and Tindal (2000):

> Its philosophy is rooted in the conviction that private sector management is superior to public administration. The solution, therefore, is to transfer government activities to the private sector through privatization and contracting out. Since that obviously can't be done for every government activity, the next best thing is to transfer business practices to government operations. (p. 285)

This idea has been around for some time. In fact, the claim that the private sector is the appropriate "efficient" model was presented, defended, and popularized more recently in David Osborne and Ted Gaebler's 1992 book, *Reinventing Government: How the Entrepreneurial Spirit is Transforming the Public Sector.* Osborne and Gaebler (1992) used the phrase "entrepreneurial government" to describe the new model they saw emerging across America throughout the eighties. They define an entrepreneur as a person who "uses resources in new ways to maximize productivity and effectiveness". (p. xix) They describe entrepreneurial public organizations as organizations that

> steer more than they row; empower communities rather than simply deliver services; encourage competition rather than monopoly; are driven by their missions, not their rules; fund outcomes rather than inputs; meet the needs of the customer, not the bureaucracy; concentrate on earning, not just spending; invest in prevention rather than cure; decentralize authority; and, solve problems by leveraging the market-place, rather than simply creating public programs. (p. xix)

As noted by Tindal and Tindal (2000: 285), whatever the merits of these respective arguments, the fact is that the new public management has brought significant change

---

[19]Summarized from Borins, 1995.

in a number of parliamentary democracies, especially in countries such as Britain, Australia, and New Zealand.

Third, the underlying ideology of the market model can be summarized by the neo-conservative market-oriented politico-economic agenda adopted by the Conservative government of Ontario. At the heart of this concept is the belief that government intervention usually does not work and that markets usually do. The extension of this belief can be summarized by the following principles: less government spending, fewer politicians, lower taxes, leaner infrastructure, removing barriers to growth, in short, less government. The neo-conservative agenda is to privatize, deregulate, restructure and "do better with less"[20].

In her recent critique of this agenda, author Brooke Jeffrey (1999) states:

> The neo-conservative agenda promotes a negative option, not a positive one. Its goal is dismantling the state and removing it as much as possible from the marketplace. Not surprisingly, then, the «Common Sense Revolution» proposed to reduce the deficit and lower taxes solely by cutting back dramatically on the expenditures - the programs, services and activities - of the provincial government. (p.198)

However, not everybody shares this view of government dominated by the market place. Many see a larger role for the public sector. Also, many attribute shortcomings to the market model and signal potentially negative implications for citizens and communities. For example, Jon Pierre (1995) studied the increasing

---

[20]Progressive Conservative Party of Ontario. 1994. *The Common Sense Revolution.*

emphasis on the definition of the public as "consumers" of government services. He examined the impact of the differences among characterizations such as "citizens", "clients", and "consumers" on political culture. In particular, he found that the definition of the public as consumers may undermine important political and constitutional values associated with citizenship. (Pierre, 1995)

Also, Self (1993) shows the inability of public choice theory to offer any acceptable view of the public interest and argues the need to return to a more positive concept of active citizenship based upon a wider range of social values.

Furthermore, in a critic levelled at new public management, it is claimed that "private sector management practices are not easily transferred to government. If public management is lethargic, cautious, expensive, unresponsive, or any of the other criticisms levelled against it, these shortcomings have more to do with parliament and politicians than with public servants." (Tindal and Tindal, 2000: 285)

2.2.2 The Participatory State

According to Peters (1993), the participatory state model is almost the antithesis of the market approach in terms of the political ideologies of most of its advocates. It is generally associated with the political Left, and has been discussed under a number of different titles (e.g., empowerment state). In fact, the participatory state model refers to participation from both without (citizens) and within (workers) the state.

> The fundamental assumption in this approach is that there is a great deal of energy and talent being under-utilized at the lower echelons of hierarchies, and that the workers and clients closest to the actual production of goods and services in the public sector have the greatest amount of information about the programs. It is assumed that if those ideas and talents are harnessed, government will work better. The general prescription, therefore, is for greater participation and involvement on the part of those groups within government who are commonly excluded from decision making. Somewhat predictably, the advocates of this approach tend to come more from the political left, although some from the right - those interested in empowerment and self-management by clients - also advocate versions of this approach. (Peters, 1993: 13)

Though this citation refers primarily to participation from within the state, the model also calls for involvement from the citizens.

The participatory state's intellectual roots can be traced back to a body of literature that argues that a process of involvement and participation is the best way to motivate individuals. A second body of literature which argues that the lower echelons of public organizations are central to the effective functioning of those organizations, and as a simple reality, the role of "street level bureaucrats" needs to be recognized. Also, there are various other contributions in the literature on "discursive democracy" and other similar concepts that argue for enhanced participation by citizens in the decisions that directly affect their lives.[21] Finally, as mentioned previously, the participatory state has been discussed using a number of different labels, such as empowerment state, citizen engagement, or just simply, public participation.

---

[21]Summarized from Peters, 1993.

Kernaghan (1992) defines the empowerment central to this view of governance in the following terms:

> At its broadest level, it can be viewed as a growing phenomenon involving demands by people all over the world to be recognized, consulted, and valued. It is also used more narrowly to describe a wide range of efforts to enhance the power and the efficacy of individuals, groups, and organizations in society. (p.194)

In his 1992 article titled "Empowerment and public administration: revolutionary advance or passing fancy", Kernaghan describes the external and internal dimensions of empowerment in the spheres of organization and management. "The external dimension", explains Kernaghan, "involves an organization's efforts to empower its clients or customers by involving them in its decision-making process". (p. 195) In the public sector, this involvement can be pursued through such means as partnership arrangements and various forms of community and client involvement. As noted by Kernaghan (1992), "This external aspect of empowerment is similar to higher-level forms of citizen participation, in which citizens exercise real power rather than being manipulated or being involved in merely token participation". (p. 195)

The internal dimension of the concept of empowerment is described as a synthesis of several theories and practices in organizational behaviour and human resource management. "It has been influenced by theories and techniques in such areas as participative management, quality circles, job enrichment, training, organizational design, and leadership, and it is closely related to the organizational development (OD) movement." (Kernaghan, 1992: 196)

In her latest book, *Reinventing Government or Reinventing Ourselves*, Hindy Lauer Schachter (1997) challenges the assumptions of the "reinventing-government" movement now in vogue. Her book unites a call for active citizenship with the current concern for improving public-agency performance. She argues that citizens should not be viewed as customers of government, as reinventing-government advocates assume, but rather as government's owners. By analysing a turn-of-the-century model of urban reform that depicts this relationship between citizens and government, Schachter shows how reinvigorating an active public is essential to increasing agency efficiency and responsiveness. She offers two strategies for moving toward active citizenship: better citizenship education, including service learning, and public agencies' provision of better-focussed information for their owners. By doing so, Schachter links two of the three dimensions of democratic participation previously discussed, citizens and employees.

In *Citizen Engagement: Lessons in Participation from Local Government*, Graham and Phillips (1998) argue that more information from the public should be obtained in order to produce better decisions and they argue that real power in decision making should be shared with citizens. In their book, they provide the following historical overview of the evolution of public participation in the sphere of municipal politics. First, according to their interpretation of its evolution, municipal government initiatives to involve citizens were parallelled by community activists' efforts to involve the public in community development. According to the authors, "The thinking was that community-development initiatives that involve citizens in decision-making build

healthier communities and nourish democracy by enhancing civic education and by producing effective citizens who are less alienated from political institutions." (Graham and Phillips, 1998: 6) This notion was first argued by John Stuart Mill. Second, Saul Alinsky in the 1930s through 1950s stressed and developed techniques for community empowerment.[22] Third, during the late 1960s and early 1970s, the Canadian federal government promoted "social animation" activities by providing grants to local organizations and by supporting activists. Fourth, urban-reform movements of the 1970s were fuelled by the notion of neighbourhood control, as groups of citizens tried - with considerable success - to stop major urban-renewal projects. And finally, the notion of "social capital", a term used by Jane Jacobs[23] in 1961 and made popular in the 1990s by the work of Robert Putnam[24], also argues for civic participation. It refers to trust relations but, for Putnam, one of the most powerful ways of developing social capital was through horizontal networks of volunteer community-based organizations. Putnam argues that "social capital builds social trust and mutual cooperation among citizens, bolsters performance of the polity, and contributes to more efficient government and a stronger economy". (Graham and Phillips, 1998: 6)

Dave Broad and Wayne Anthony (eds.) (1999), in *Citizen or Consumers? Social*

---

[22]See Saul Alinsky. 1971. *Rules for Radicals*. New York: Random House.

[23]See Jane Jacobs. 1961. *The Death and Life of Great American Cities*. New York: Vintage Books.

[24]See Robert Putnam. 1993. *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton: Princeton University Press.

Policy in a Market Society, argue that social policy is about citizens choosing the kind of society they want to live in. They believe that the mid-20th century Keynesian welfare state can be seen as a citizenship package which included acceptance of intervention by the state to maintain economic growth and social stability. In their minds, this meant the inclusion of many previously excluded groups in the social policy process and the institutionalization of a collective responsibility for individual welfare. But they also argued that, with the ascendancy of neo-liberalism, the politics of citizenship are being replaced by a notion of citizens as consumers, whose medium of social interaction and source of economic and social security is the capitalist market.

They conclude that we must move beyond the early post-World War II conceptions of social welfare which subsumed citizenship within social rights. These rights must be expanded to incorporate areas such as gender issues and the status of immigrant and stateless populations, perhaps with a new conception of community.

2.3 DISCUSSION

It is apparent that the second model, the participatory state, in all its manifestations, is not as easily defined as its counterpart, the market model. This can first be explained by the fact that the participatory state model has various distinct intellectual roots all belonging to the larger spectrum of democratic ideals. Second, market approaches are based on economic principles. They have very clear

parameters and vocabulary. Economic principles have been developed into formal, practical models using specific definitional elements and vocabulary. However, participatory approaches have not yet been formalized and therefore their arguments are strongly anchored in philosophy and social theory. As a result, though it is possible to debate market approaches using very formal, operationalized language, participatory approaches cannot, at this point and time, be discussed with such precision.

Nevertheless, using Peters (1993) market and participatory models as a basis, we have attempted to define these two models using six key aspects of local governance: bureaucracy, political infrastructure, policy making, links to the community, the conception of the *public interest* and an overall conceptualization of what constitutes "good government", and the perceived role of the citizen.

The first key aspect of local governance which has been identified is the bureaucracy which includes both the organizational structure and personnel management. This section basically attempts to answer two fundamental questions. First, how should the municipal structure be organized? The market model calls for the splitting up of large departments into smaller "agencies", through assigning functions to lower levels of government, or through using private or quasi-private organizations to deliver public services. This approach is particularly applicable when the goods or services in question are marketable, and thus, can be easily contracted out or privatized. As for the participatory model, it appears to be more concerned with the process of bureaucracy than with its structure. However, this model does call for a

"flatter" organizational structure with fewer tiers between the top and bottom levels in order to ensure multiple entry points. Such a structure, it is believed, greatly facilitates and even enhances citizen participation.

Second, how should municipal employees be recruited, motivated and managed? Although, both models strive to recruit the best possible employees and retain them, they do differ in their managerial styles. The market model tends to adopt such measures as a merit principle pay scale, performance measurements, reward schemes and performance bonuses in order to motivate employees and ensure efficiency. On the other hand, the participatory model is primarily concerned with including the lower levels of the organization more directly in managerial decisions as a means of achieving the same performance and efficiency goals just outlined. Furthermore, the participatory model strives to include the clients of the organization (i.e., the citizens) more directly in managerial decisions.

The second key aspect of local governance is the political infrastructure. At issue is what should the primary role and responsibilities of municipal councillors be? The market model clearly calls for a reduced role for politicians, a reduced number of politicians and a reduced support for politicians; on the other hand, the participatory model advocates a stronger role for politicians. Also, this model envisages citizens playing important roles in collaboration with politicians via consultative bodies, neighbourhood councils and advisory committees.

The third key aspect of local governance, policy making, involves the role of the senior management team, other senior career public servants and lower echelon workers in the policy process. Both models advocate a decentralized decision making and implementation process, however, they differ in terms of who should be called upon to provide input and formulate policy. The market model expects multiple, "entrepreneurial" agencies within the organization or at arms-length of the organization to make autonomous decisions. Often, this approach relies on outside expert opinion to help it in its decision making process. The participatory model on the other hand, relies more heavily on input from the citizens, lower echelons workers and politicians. This can be described as a "bottom-up" approach to policy making.

The fourth key aspect of local governance is the local government's links to the community. More specifically, how should local government interact with the voluntary sector - including neighbourhood associations, issue-oriented groups and identity-based social movements, and local service clubs? The market model, by its very nature, tends to encourage links to the business community. Advocates of this model point to the economic benefits of initiating and fostering relations with business organizations. The participatory model, by its inherent nature, tends to encourage greater interaction with not only the business community, but also the voluntary sector. In fact, the participatory model emphasizes the societal benefits of nurturing such links.

The fifth key aspect of local governance is described as the conception of the *public interest* and an overall conceptualization of what constitutes "good government".

Here, the ideological differences between these two models become quite apparent. Advocates of the market model believe that government should be judged on the basis of how cheaply it can deliver public services. Also, that citizens be allowed to make consumer choices is of prime importance. Rather than viewing the clients of the organization as consumers, advocates of the participatory model view them as citizens. This model assumes that the public interest can be best served by allowing citizens and employees the maximum possible involvement in decision making. Also, this vision attempts to involve societal interests in governance more explicitly.

Finally, we can conclude this comparative analysis by stating that the market model's conception of citizens as consumers/taxpayers greatly reduces the role of the citizen. Conversely, the participatory model's conception of citizens greatly enhances their role in the community.

These six aspects of local governance are summarized in Table C on the following page.

TABLE C: Definitional elements of the two models of governance

ASPECT OF LOCAL GOVERNANCE	MARKET MODEL	PARTICIPATORY MODEL
BUREAUCRACY		
i) Organizational structure:	splitting up of large department into smaller "agencies", through assigning functions to lower levels of government, or through using private or quasi-private organizations to deliver public services	process more important than structure structural reforms that facilitate citizen participation "flat" organizational structure: fewer tiers between the top and bottom levels
ii) Personnel management:	merit principle pay scale performance measurements reward schemes performance bonuses	include citizens in decisions that directly affect their lives include lower levels of the organization more directly in managerial decisions
POLITICAL INFRASTRUCTURE	reduced role of politicians reduced support for politicians reduced number of politicians	stronger role for politicians use of consultative bodies, neighbourhood councils and advisory committees

POLICY MAKING	decentralization of policy and implementation of decisions decentralizing bureaucratic functions to multiple, "entrepreneurial" agencies that would be expected to make autonomous decisions greater influence of expert opinion	decentralized decision making input of citizens involvement of lower echelon workers ("bottom-up") greater influence of politician input
LINKS TO THE COMMUNITY	business community no links to voluntary sector	encourages "dialogical" process between citizens and policy makers
THE CONCEPTION OF THE *PUBLIC INTEREST* AND AN OVERALL CONCEPTUALIZA-TION OF WHAT CONSTITUTES "*GOOD GOVERNMENT*"	government should be judged on the basis of how cheaply it can deliver public services citizens allowed to make consumer choices focus on consumers	assumes that the public interest can be best served by allowing citizens and employees the maximum possible involvement in decisions attempts to involve societal interests in governance more explicitly focus on citizens
ROLE OF THE CITIZEN	consumers / taxpayers reduces the role of the citizen	citizens enhances the role of the citizen

We started our review with a reference to the two traditional roles of local government: access/democracy and service/efficiency. Our discussion leads us to the observation that the two opposing models outlined in the table above can be further understood in terms of the relative priority they attach to each of these two traditional roles. The market model focusses primarily on the service/efficiency aspect of local government whereas the participatory state is more preoccupied with the access/democracy aspect. We can therefore situate the municipal amalgamation exercise undertaken in the Ottawa-Carleton region within the "democracy-efficiency trade-off" which has been, and continues to be, as evidenced here, the key feature of municipal government in Canada.

In adopting an operational model based on business principles, the Board's intent to focus primarily on issues relating to the efficient delivery of services was made clear. Our task, as outlined in the introduction, is to assess the extent to which the Board was successful in implementing its entrepreneurial vision of local government. In Chapters 1 and 2, we have provided an historical and theoretical overview of the contemporary debate surrounding this restructuring exercise. Our next step is to outline in the following chapter the procedure and method that will enable us to carry out our inquiry.

3. METHODOLOGY AND RESEARCH PLAN

3.1 METHODOLOGY

The initial research for this thesis was done in the context of a project that set out to organize the public documents of the Ottawa Transition Board. From its creation in January 2000, the Board recognized the importance of having a record of the processes and decisions involved in the historical restructuring exercise of amalgamating the 12 municipalities in the Ottawa-Carleton region. With this goal in mind, the Board initiated and established a documentation project in collaboration with the University of Ottawa. It was decided that the objective of the project would be threefold: to provide a record of the sequencing of decisions; to provide brief descriptions of the main elements considered in major decisions; and, to provide descriptions of the consultation and decision-making processes involved in these decisions. The result of this initiative is the document entitled *Public Record of the Ottawa Transition Board*.[25] This document was made public on March 1, 2001.

To understand and organize the public documents of the Board, a rigourous exercise of observation, documentation and compilation was undertaken. In order to capture the essence of the transition process, the following steps were taken: the

---

[25]Marc Gervais. 2001. *Public Record of the Ottawa Transition Board*, Publication of the Ottawa Transition Board in collaboration with the University of Ottawa, Ottawa, Ontario.

Board's scheduled bi-weekly public meetings as well as the Board's Communications Team's weekly meetings were attended; the work of the key project teams was closely monitored; the minutes of all Board meetings as well as the project team reports were gathered, read and summarized; and, all articles from the local newspapers on Board business were compiled and organized.

The work undertaken to produce the aforementioned document served as the foundation for our reflection on the apparent discontinuity between what the Board set out to do and the end result of its efforts. Our analysis relies primarily on the interpretation of the public documents produced by the Board in the fulfilment of its mandate.

3.2 RESEARCH COMPONENTS

As already stated, the process of political change lies in the interplay of opposing forces. Our research will examine this interplay between two sets of political actors: those who wish to control the process (e.g., the Board, the provincial government, business interests, etc.) and those who wish to influence the process (community organizations, special interests groups, the media, politicians, etc.). However, although our research studies the interplay of two series of political actors, it will focus exclusively on the actions and decisions of the Board. The transition exercise will therefore be

assessed in the context of a clash of opposing visions of municipal governance. To accomplish this, two models have been chosen to serve as the theoretical framework of our research: the market model and the participatory model.

Our research is therefore organized into four components. Their purpose is to determine the extent to which the decisions and recommendations made by the Board in the creation of the new City of Ottawa reflect the vision of the entrepreneurial role of municipal government (market model) that was defined at the outset of the transition process by the provincial government and the Board.

3.2.1 Component 1: Define the market model criteria

Using Peters' (1993) assumptions as a basis, an attempt has been made to define the market model by outlining its content in light of six important aspects of local governance. This was accomplished in the previous chapter (Chapter 2).

3.2.2 Component 2: Define the participatory model criteria

Also in the previous chapter (Chapter 2), this model was defined according to the same six aspects of local governance.

3.2.3 Component 3: Compare Board decisions to definitional elements

The Board set up over 80 project teams to address all of the transition issues. Each transition project was given the task of studying a particular issue and providing the Board with recommendations. From these recommendations, the Board made its final determinations regarding the amalgamation and restructuring of the current 12 municipal bodies into the new City of Ottawa. Amongst these, the key transition projects were: administrative structure, service restructuring, point of service (service delivery strategy), structure and staffing, political infrastructure, language services policy, rural issues, voluntary sector and youth.

The final Board determinations within each of these projects will be mapped onto the elements of Table C in an attempt to quantitatively determine their compatibility with the market model. Those Board determinations that reveal aspects which are contradictory to the market model will be analysed in an attempt to determine the extent to which they can be accounted for by the involvement or interplay of factors generally associated with the participatory model of governance. When the recommendation maps equally onto both models, a qualitative analysis of the decision making process will be applied in an attempt to elucidate the Board's intent.

These projects were chosen for the following reasons: 1- they attracted and generated a lot of public interest during the course of the transition process; 2- they were given prominent coverage by the various media outlets; 3- they illustrate the

decision-making process that led to the Board's final recommendations; and, 4- the approaches and recommendations made by these project teams reflect the variety of models possible, ranging from the market model at one end of the spectrum to the participatory model at the other.

This section of the research draws from the following primary sources: Ottawa Transition Board project team reports, Ottawa Transition Board publications, and Minutes of Ottawa Transition Board meetings.

3.2.4 Component 4: Review of hypothesis

A final discussion will attempt to explain the observed variations in the outcomes of the desired goals set by the Board and the province in terms of factors related to time, mandate (particularly scope and ideology) and intervening democratic processes. The inevitable intervention of democratic principles will be revealed as having played a determining role in the overall process via public opinion, media scrutiny, opposing lobbying efforts from concerned citizens and community organizations, and diverging views from both within and without the municipal organizations involved.

3.3 RESEARCH LIMITATIONS

This research will allow us to situate the changes brought to local governance in Ottawa as proposed or implemented by the Board within their theoretical methodological framework, thereby providing a future basis for analysis, commentary, and criticism.

However, the following research limitations have been identified: first, our research will not be an impact study of the decisions and recommendations made by the Board; second, the sources for our research will be restricted to public documents only; third, although our research studies the interplay of two series of political actors, it will focus exclusively on the actions and decisions of the Board; and, fourth, all decisions and recommendations made by the Board to the newly elected Council can be revisited by Council at anytime in the future, and therefore, are subject to change.

4. MAPPING OF FINAL BOARD DETERMINATIONS

The purpose of this chapter is twofold. First, to identify amongst the major Board decisions and recommendations which ones are compatible with the market model principles discussed in Chapter 2, and which ones are not. This will be accomplished by mapping the Board's final determinations onto the definitional elements of the two models of governance, the market model and the participatory model, as summarized in Table C. Second, to highlight the Board's activities and pronouncements (i.e., objectives, approaches and rationalizations) in an attempt to elucidate the Board's intent and underlying vision as they relate to its expressed goal of making the municipal government more business-like.

For reasons previously stated, our discussion focusses on the recommendations resulting from the work of certain project teams considered key to the transition process. Also, for the sake of clarity and convenience, these project teams have been regrouped under three main headings: **Administrative Structure** (includes the work of the Service Restructuring Teams, the Point of Service Project Team and the Structure and Staffing Project Team), **Political Infrastructure**, and **Governance Issues** (i.e., Language Services Policy, Rural Issues, Voluntary Sector, Youth, etc).

4.1 ADMINISTRATIVE STRUCTURE

The details of the departmental organizational structures came about as a result of the combined efforts of several projects, most notably: Service Restructuring, Point of Service, and Structure and Staffing. Therefore, the recommendations made by the Board with regards to the overall organizational structure as well as issues relating to personnel management have been regrouped and will be dealt with together.

The recommendations highlighted in this section refer to the first and third aspects of local governance as shown in Table C: the bureaucracy and policy making.

As stated previously in Chapter 2, the bureaucracy includes both the organizational structure and personnel management. With regards to the organizational structure, the market model calls for the splitting up of large departments into smaller "agencies", through assigning functions to lower levels of government, or through using private or quasi-private organizations to deliver public services. As for the participatory model, it appears to be more concerned with the process of bureaucracy than with its structure. However, this model does call for a "flatter" organizational structure with fewer tiers between the top and bottom levels in order to ensure multiple entry points. As for the issue of personnel management, the market model tends to adopt such measures as a merit principle pay scale, performance measurements, reward schemes and performance bonuses in order to motivate employees and ensure efficiency. On the other hand, the participatory model is

primarily concerned with including the lower levels of the organization more directly in managerial decisions as a means of achieving the same performance and efficiency goals just outlined. Furthermore, the participatory model strives to include the clients of the organization (i.e., the citizens) more directly in managerial decisions.

With regards to policy making, as previously stated in Chapter 2, both models advocate a decentralized decision making and implementation process. On the one hand, the market model expects multiple, "entrepreneurial" agencies within the organization or at arms-length of the organization to make autonomous decisions. Often, this approach relies on outside expert opinion to help it in its decision making process. On the other hand, the participatory model relies more heavily on input from the citizens, lower echelons workers and politicians.

Although the work done by the Service restructuring, Point of Service, and Structure and Staffing project teams was closely related, they did have somewhat different objectives. For example, the main objective of the Service Restructuring Teams (SRTs) was to restructure and not simply amalgamate. They were mandated to design and implement a new service model that: reflects best practices; resonates with the vision adopted by the Board for the new City of Ottawa: "*To be the municipal centre of excellence, meeting citizens' expectations for value for money, and accessible, convenient quality services.*"[26]; is citizen focussed; addresses cross-service issues such

---

[26]Ottawa Transition Board. 2000. *Project Charter for Service Restructuring Teams.* Ottawa.

as service standards, accountability framework, enterprise model, etc.; realizes savings in the order of approximately 75 million by 2003; and, reflects a leaner organizational structure.

For its part, the Point of Service Project Team's main objective was to provide an overall service delivery vision for the new City of Ottawa and the framework to bring it about. This project team was mandated to design and implement the overall strategy and the supporting framework to ensure a citizen-focussed, results-based, performance-driven service delivery organization for the new City. The Point of Service Project Team was also mandated to manage the design and implementation of cross-functional service delivery components in order to bring accessible services to the community through a one window concept. This includes the design and implementation of service centre(s) and electronic service delivery.

Finally, the objective of the Structure and Staffing Team was to create an organization structure which reflects the new City vision, to define and fill its key positions, and to undertake the staffing of positions so that the organization can become operational by January, 2001.[27]

It is important to keep in mind that all integration and restructuring project teams

---

[27]The Board only had a mandate to hire at the executive level and could not begin staffing for unionized positions with the exception of exempt positions such as administrative assistants and executive assistants.

shared the primary goals of collectively realizing savings in the order of approximately 75 million by 2003 and implementing a leaner organizational structure by reducing personnel.

With these two main goals in mind, the Board decided that the new City of Ottawa's administration would be comprised of six departments: Protective Emergency Services; People Services; Transportation, Utilities & Public Works (TUPW); Development Services, Corporate Services; and, Human Services.

The Protective and Emergency Services Department regroups 911, ambulance, fire and police, as well as emergency measures planning and by-law enforcement services. The People Services Department regroups services previously known as: Recreation, Culture, Arts and Heritage, Social Services, Public Health, Homes for the Aged, Long Term Care, Social Housing, Child Care and Library Services. TUPW includes all of the basic municipal services that residents have come to expect. TUPW ensures the roads are plowed, garbage is collected, streets and parks are maintained, water is available, and buses will run. The Development Services Department includes building services, development and infrastructure approvals, planning, environment and infrastructure policy, and business development. The Corporate Services Department includes Communications and Public Affairs, Financial, Fleet, Informatics, Legal, Point of Service, Real Property and Facilities Management, Records Management, Secretariat, and Strategic Planning and Audit. Finally, the Human Resources Department has four directors reporting to the General Manager of Human Resources.

The four "directorates" are: Labour Relations, Compensation, Payrolls and Benefits, Organizational Effectiveness, and Human Resources Services.

The overall organizational structure of the new City, as approved by the Board, is comprised of 4 levels of management which accounts for 138 managers: 1 City Manager, 6 General Managers, 31 Directors and 100 Managers. Upon making public the new organizational structure as it related to the first three levels of management, the Board announced that it had approved a leaner management structure for the new City of Ottawa with about 75% fewer management positions than in the current combined 12 Ottawa-Carleton municipalities. "This lean new management structure will lead to a more efficient municipal administration and provide for millions of dollars in salary savings and overhead costs," said Transition Board Chair Claude Bennett. "The new structure is designed to encourage flexible and innovative management practices that will help make the new City of Ottawa the most dynamic and competitive organization it can be."[28]

As highlighted by the Board in its final report, there are two cornerstones to the administrative model introduced through the transition process - Points of Service Strategy and Centres of Expertise. The first initiative provides the public with the means of accessing municipal services; and the second helps to re-organize the municipality to deliver those services. From the Board's perspective, these two

---

[28]Ottawa Transition Board. 2000. *Press Release - August 14.* Ottawa.

initiatives are seen as fundamental to achieving the cost savings targets of the new City.

4.1.1 Recommendations compatible with the market model:

Amongst the recommendations adopted by the Board with regards to the new administrative structure, the following reflect market model priorities:

- "Council review the current range and levels of municipal services and consider **lowering or eliminating** certain municipal services that are not core to the need of a healthy City, (e.g. operations of a golf course or equestrian park)."[29]
- "Given that centres of expertise will strive to find the best methods of delivering services, the Board urges Council to express the political will to explore the most efficient and effective way of doing the business of government, including exploring **alternative service delivery (ASD)** solutions."[30]

According to Board documents, ASD is a general term that describes a number of resourceful options in service delivery. These options include improved delivery by a department, special operating agencies, privatization, tri-sector partnerships (i.e., public-private-volunteer), public partnerships within the community, inter-governmental partnerships, volunteerism, and use of technological advances, such as the internet or kiosks.

---

[29]Ottawa Transition Board. 2001. *The New City of Ottawa: Fulfilling Its Potential.* Final Report and Recommendations. Ottawa. p.15.

[30]*Ibid.*

Due to the terms of reference of its mandate, time constraints, and overriding labour issues, the Board could not alter service levels or even eliminate certain services such as operating municipal daycare or homes for the aged; revisit past service policies and make any decisions regarding the harmonization of services throughout the City; or, fully explore alternative service delivery (ASD) options, as it would have liked to. However, in its final report, the Board did offer some starting points with regards to ASD such as community centre programming, fleet management, property management and maintenance, landfill management and operations, call centre management, 911, ambulance, printing, homes for the aged, joint metre reading for water, natural gas and hydro.[31]

These first two recommendations are compatible with the market model principle of eliminating or lowering certain municipal services or contracting them out to private or quasi-private organizations. Furthermore, to infer that leisure activities, such as golf, and recreational services, such as community centre programming, are non essential to a healthy city is indicative of a philosophical standpoint compatible with the entrepreneurial vision. In the end, the result is the same: a transfer of responsibility from the public domain to the private sector.

Another recommendation adopted by the Board with regards to the new administrative structure reflecting market model priorities is:

---

[31] *Ibid.*

- "The City should proceed as quickly as possible with the staffing process, maintaining an open and transparent hiring process and keeping as an objective the target of **1,100 fewer positions** over three years (including hydro)."[32]

This recommendation is compatible the market model principle that calls for a leaner bureaucratic infrastructure.

As previously stated in Chapter 1 with regards to personnel management issues, the Board announced in April 2000, that the new City organization would embrace an enterprise culture. On April 10, 2000, in a press release titled *"The New City of Ottawa's Organizational Framework"*, the Board announced it would direct the new City Manager and senior management team to apply an accountability model in the delivery of services to the public. The press release goes on to state:

> Depending on the activity, there are three vehicles to provide the necessary accountability throughout the municipal structure: performance agreements, measured outcomes, and a business template: Performance Agreements: a process of performance reviews that will ensure everyone is aware of what is expected of them and how they contribute to the goals of the organization (i.e. to be applied with corporate services, people services); Measured Outcomes: the setting of competitive benchmarks to manage outcomes and service to the public, and, ultimately, to achieve best practices within the organization (road maintenance, printing, etc.); Business Template: activities that are measured with bottom-line results and are managed to deliver high-quality service at the lowest possible cost (hydro, outsourced activities)"[33]

---

[32]*Ibid.* p.22.

[33]Ottawa Transition Board. 2000. *Press Communiqué - April 10.* Ottawa.

This accountability model in the delivery of services to the public as articulated and promoted by the Board is also compatible with market model priorities as they relate to personnel management.

Finally, the Board's Terms of Reference (see Appendix B) did not allow it to directly address policy. However, the hierarchical top-down structure which was implemented by the Board does allow us to speculate on the potential effect of such a structure on policy making. It would appear that such a structure would lend itself well to the implementation of a highly centralized policy and decision making process. Such a model is incompatible with both the market and participatory models. However, the adoption of an entrepreneurial model as the guiding principle of the new City's day-to-day operations and administration would imply the implementation of market oriented principles compatible with those highlighted in Table C (i.e., decentralization of policy and implementation of decisions; decentralizing bureaucratic functions to multiple, "entrepreneurial" agencies that would be expected to make autonomous decisions; and, greater influence of expert opinion). In other words, though the Board did not establish policy directly, it implanted a framework that would, if implemented as recommended, influence the nature and direction of the future City's policy making.

We therefore conclude that nothing in the Board's output regarding the new administrative structure can be interpreted as a reflection of participatory model principles. The Board did not put into place an organizational structure that facilitates citizen participation or any mechanisms that would include citizens in decisions that

directly affect their lives. Furthermore, there was no mention of any mechanisms that would include lower levels of the organization more directly in managerial decisions.

4.2 POLITICAL INFRASTRUCTURE

This section examines the decisions and recommendations made by the Board as well as its stated objectives with regards to the new proposed governance infrastructure for the new City of Ottawa. This new political infrastructure is referred to in the second aspect of local governance as shown in Table C. As stated previously, the market model clearly calls for a reduced role for politicians, a reduced number of politicians and a reduced support for politicians; on the other hand, the participatory model advocates a stronger role for politicians. Also, this model envisages citizens playing important roles in collaboration with politicians via consultative bodies, neighbourhood councils and advisory committees.

The Political Infrastructure Project Team was appointed by the Board to examine and make recommendations in the following areas: roles and responsibility of elected officials; roles and responsibility and competency profiles of political staff; compensation levels of elected officials and their staff; other costs (operating budgets) and requirements (facilities, equipment, etc.); structure of committees; interface mechanisms with municipal staff and support requirements; and, operating guidelines.

On June 12, 2000, the Ottawa Transition Board approved the following recommendations:[34]

SALARY

- That the preference of the Ottawa Transition Board is that the salaries of the mayor and councillors be fully taxable.
- That the mayor be paid $104,000 with 1/3 tax exempt in accordance with existing legislation. The Ottawa Transition Board prefers the mayor be paid $140,000 annual taxable gross income.
- That the councillors be paid $56,000 with 1/3 tax exempt in accordance with existing legislation. The Ottawa Transition Board prefers the councillors be paid $70,000 annual taxable gross income.
- That the mayor and councillors not receive severance pay.
- That the mayor and councillors receive no other stipends for chair or memberships on committees, boards, or agencies.
- That the mayor's staff and councillors' staff be hired for a term not exceeding the term of council and that they not be entitled to receive severance pay in accordance with the Employment Standards Act.

BENEFITS

- That the elected officials be able to access the medical / vision / dental / life insurance available to city management.
- That the mayor's staff and councillors' staff be able to access the medical / vision / dental / life insurance available to city staff.

OFFICE LOCATION

- That the mayor's office be at City Hall.
- That the councillors' main office (and that of their staff) be in a municipal building, other than City Hall, in their ward. Shared workstations for councillors will be made available in a common work area at City Hall.

PERSONNEL BUDGET

- That a car and driver be provided by the city for the mayor's use.
- That the mayor be provided with a staff budget of $250,000 annually. Other staff required to support the mayor's office (e.g. reception, filing, other administrative staff) be city employees.

---

[34]Ottawa Transition Board. 2000. *Board Decisions - June 12*. Ottawa.

- That each councillor be provided with a staff budget of $100,000 annually.
- That support for the common office work area at City Hall (e.g. reception) be city employees.

On August 28, 2000, the Ottawa Transition Board approved the following recommendations:[35]

ROLES
The Board views the fundamental role of elected representatives as being legislators. The mayor and councillors are the legislative arm of municipal government to develop policy and create the laws of the City.

OFFICE BUDGETS
The budget for the office of the mayor be $250,000, and each councillor's office budget be $25,000.

COMMITTEES

Standing Committees
- There will be four Standing Committees of Council: Operations Committee, Development and Planning Committee, Community Services Committee (including protective emergency services department), and Corporate Services Committee.

Ad Hoc Committees and Task Forces
- Council may, from time to time, establish such ad hoc committees and task forces as it finds necessary. Any ad hoc committee or task force shall have a specific mandate, set membership, a finite term and a defined reporting relationship to council through one of its standing committees.

Advisory Committees
- That Council should draw upon expertise of members of the public through advisory committees. Advisory committees should report to the standing committee through a council member (sponsor) and should be supported financially by the City. For example: Council may wish to establish advisory committees to review, research, comment, and advise with respect to rural issues or environmental issues, or concerns expressed by youth.

---

[35]Ottawa Transition Board. 2000. *Board Decisions - August 28*. Ottawa.

4.2.1 Recommendations compatible with the market model:

The market model calls for a reduced number of elected representatives. Provincial legislation reduced the number of municipal politicians in the Ottawa-Carleton region from 84 to 22 (21 councillors and 1 mayor).

In early March 2000, Mr. Claude Bennett, Chairman of the Board, publicly stated that the Board was considering making council positions part-time jobs. This suggestion is compatible with the market model principle of reducing the role of elected representatives. However, it was not retained by the Political Infrastructure Project Team nor did it appear in the Board's final recommendations.

The Board did announce that it viewed the fundamental role of elected representatives as legislative in nature and that the mayor and councillors are the legislative arm of municipal government responsible for developing policy and creating laws. This decision appears quite inoffensive in itself save for when it is understood in the context of a prior decision made by the Board with regards to the location of councillors' offices.

On June 12, 2000, the Board approved a recommendation stipulating that the mayor's office be at City Hall and that the councillors' main office (and that of their staff) be in a municipal building, other than City Hall, in their ward. It was also decided that shared workstations for councillors would be made available in a common work area at

City Hall. According to the Board, the councillors' main office (and that of their staff) should be located in their ward in order to ensure their accessibility and their accountability to constituents. The Board further stated that this recommendation is in accordance with the citizen-based enterprise culture and service delivery model adopted by the Board for the new City. It would, however, appear that this recommendation was an attempt by the Board to ensure that the day-to-day operations of the City's affairs be in the hands of non-elected officials and at arms length from the elected representatives. As encouraged by market model principles, it was an attempt to shift the concentration of power away from elected officials thereby reducing their role and influence on city governance.

The Board also attempted to save money as well as reduce the role and support of politicians by reducing the mayor's and councillors' office and personnel budgets, by reducing the number of standing committees to four, and by eliminating the usual compensatory benefits (i.e., stipend) allotted to councillors for serving as chair or as a member of a committee, Board, or agency.

4.2.2 Recommendations non-compatible with the market model:

Amongst the recommendations adopted by the Board with regards to the new proposed political infrastructure, three recommendations clearly do not reflect market model priorities, but rather participatory model priorities: 1- the allotment of staff support for Council; 2- the existence of advisory committees; and, 3- the allotment of staff

support for committees.

These recommendations are very much in line with the participatory model principles of providing elected representations with the proper resources and support in the fulfilment of their duties as well as involving citizens in the decision making process.

However, it is apparent that the vast majority of recommendations made by the Board to the new council with regards to the political infrastructure reflect the entrepreneurial role of municipal government promoted by the province and the Board. On a whole, these recommendations are compatible with market model principles which advocate a reduced number of politicians, a reduced role for politicians, and a reduced support for politicians.

4.3 GOVERNANCE ISSUES

This grouping of Board considerations incorporates a number of concerns which, at the outset, appear less clearly compatible with the Board's expressed ideological positions. Once the Board settled on an expanded mandate of city building rather than a simple amalgamation, it was met with demands which fell outside the immediate purview of a market or entrepreneurial vision of governance. These demands (environmental, youth and other interest groups) imposed themselves as relevant and integral parts of governance responsibilities requiring policy recommendations and statements not clearly concerned with structural or purely administrative considerations.

In fact, the Shortliffe Report did not make reference to these issues, nor did it recommend that the Board address them.

The recommendations made by the Board in response to the work of these project teams speak primarily to the fourth aspect of local governance, links to the community. To a lesser degree, they also relate to the fifth and sixth aspects of local governance as shown in Table C and discussed in Chapter 2: the conception of the *public interest* and an overall conceptualization of what constitutes "good government"; and, the role of the citizen in the new City.

First, as summarized in Table C, the market model, by its very nature, tends to encourage links to the business community, whereas the participatory model, by its inherent nature, tends to encourage greater interaction with not only the business community, but also labour, women, racial and ethnic minorities, as well as the voluntary sector. In fact, the participatory model emphasizes the societal benefits of nurturing such links.

Second, advocates of the market model believe that government should be judged on the basis of how cheaply it can deliver public services. Also, that citizens be allowed to make consumer choices is of prime importance. Rather than viewing the clients of the organization as consumers, advocates of the participatory model view them as citizens. This model assumes that the public interest can be best served by

allowing citizens and employees the maximum possible involvement in decision making. Also, this vision attempts to involve societal interests in governance more explicitly.

Finally, the market model's conception of citizens as consumers/taxpayers greatly reduces the role of the citizen. Conversely, the participatory model's conception of citizens greatly enhances their role in the community.

Given that the governance issues discussed here are of a more social character, they would not normally be labelled as essential in a pure market model analysis of a healthy new city. It could be argued however that they represented a certain challenge to the Board and, in the end, may have imposed a shift away from purely bottom line considerations.

It is clear, however, that in some cases the Board did attempt to resist a broadening of social perspectives within its deliberations. The first example of this is the timing and sequencing of the establishment of the project teams dealing with these issues. As previously stated, the Board initially unveiled approximately 50 transition projects in April 2000 (i.e., service restructuring, integration and stand-alone projects).[36] Amongst these was the creation of two distinct project teams to deal with rural issues and voluntarism. However, a project team to deal with environmental issues was not created until June 2000. As for youth issues, the Youth Transition Committee (YTC)

---

[36]See Table B.

was not established until July 24, 2000. This date marks the midway point of the Board's mandate. Finally, in September 2000, in the latter stages of its mandate, the Board created project teams to address diversity and community access issues.

As the Board itself oversold the importance and significance of the transition exercise, it was faced with demands by various interests groups who did not wish to be left out, as illustrated by these examples. At first, it would appear the creation of project teams dealing specifically with these issues (i.e., environment, youth and other interest groups) was a mere afterthought or a gesture of appeasement. If the Board sincerely believed in the importance and necessity of giving youth and other interest groups a voice in the creation of the new City, why were these initiatives not undertaken earlier? In the end, however, the Board pointed to these projects as examples of the openness and inclusiveness of the process it put into place in the creation of the new City.

A second example that shows that the Board did attempt to resist a broadening of social perspectives within its deliberations is revealed by the language services policy recommendations which reflect a more conservative view of services which is intended to prioritize pragmatic service considerations over the more politically contentious principle of bilingualism.

A third example is the prioritization of voluntary sector involvement in public service delivery as a means of obtaining low cost labour. This position has often been

identified as an example of the promotion of a corporate agenda.

However, overall, the nature of many of these issues did not lend themselves well to the entrepreneurial views promoted by the Board. As a result, these issues introduced into the Board's recommendations a broader perspective reflecting principles more compatible with a participatory model.

For example, in the voluntary sector recommendations there is a clear restriction against the use of volunteers as replacements for paid employees or for public participants of advisory roles. Furthermore, the recommendations support public financing for the voluntary sector.

A further example of the adoption of considerations contrary to a purely market model vision is the important number of recommendations regarding the establishment, implementation, and financing of equity, diversity and access programs and services equal to those of other city priorities.

The adoption of an environment charter and a report on quality of life indicators reflect, on the one hand, concerns that are not traditionally associated with market initiatives, and, on the other hand, potential costs that are not compatible with the Board's expressed budget priorities.

Perhaps the most striking indicator of the influence of non-market considerations

is the number and scope of initiatives recommended by the Board to ensure youth participation in the new City. For this reason, the recommendations made by the YTC to the Board, and subsequently adopted by the Board, deserve a more in depth analysis.

The YTC was comprised of approximately 25 youth who drew their membership from area high schools and universities as well as from several youth agencies. The YTC worked closely with a 15-member Partners Committee made up of prominent individuals, such as Max Keeping of CJOH-TV and Judy Bullis of the Ottawa Sun, to name but two, representing business, educational and not-for-profit communities. The goals of the YTC were:

> To educate, inform, mobilize and involve young people from the Ottawa-Carleton Region in the new municipal government; to gauge the level of knowledge of the younger generation on municipal issues and appreciate the current and future needs of this important segment of the population; to make the new city better known to all citizens through the dynamism and involvement of a younger generation of citizens; to foster a sense of belonging, ownership and pride in the new city among young people; to establish permanent links between young people and the leadership of the new city; and to give young people in the new city a sense of hope and optimism about the future in Ottawa.[37]

In its quest to achieve these goals, the YTC presented a series of workshops on municipal government to a select number of high schools across the region and conceived and distributed a booklet on municipal government entitled "*Ottawa 101*" to

---

[37]Ottawa Transition Board. 2000. Final Report of the Youth Transition Committee. "*Youth: A Place at the Table*". Ottawa.

area high school students.

Furthermore, the YTC presented the Board with a set of recommendations aimed at creating a framework for significant youth involvement and input in the new City of Ottawa. A report entitled "*Youth: A Place at the Table*" containing the YTC's recommendations was tabled and accepted by the Board on December 5, 2000.

On December 19, 2000, the Board passed a motion recommending that the Council of the new City of Ottawa consider giving youth a formal voice in the new City of Ottawa using the YTC's report as a guide as well as consider this motion as a priority item in early 2001.[38]

The YTC's report "*Youth: A Place at the Table*" contains four main recommendations: 1- that the new City of Ottawa have an Ottawa Youth Cabinet; 2- that Council elect two of its members to be Youth Advocates; 3- that Council establish the Youth Initiatives Program; and, 4- that an annual year-end report card on the status of youth be published.[39]

With regards to the first recommendation, the report outlines that the Ottawa Youth Cabinet's principal tasks be:

---

[38]Ottawa Transition Board. 2000. *Board Decisions - December 19*. Ottawa.

[39]The following information was taken from the Youth Transition Committee's final report entitled: "*Youth: A Place at the Table*". December 4, 2000.

> To advise, review for comment and make recommendations on any matter that affects the youth of the city; to identify and remove barriers that hinder youth from making full use of city services; to hold a yearly event to showcase youth and youth participation (beginning 2002); to hold an annual Youth Roundtable in partnership with the Mayor's Office (beginning 2002); to have an on-going, year-to-year initiative that informs and educates youth about the political process and civics in general; to ensure the voting process is "youth friendly". (p. 22-23)

The report goes on to provide the following direction with regards to the recommendation calling for two councillors be elected as Youth Advocates to the Ottawa Youth Cabinet. First, "when Council is electing the Youth Advocates, they should take into account experience working with youth and dedication to youth and youth-oriented issues". Second, "that the Youth Advocates act as a link between the Ottawa Youth Cabinet and City Council". Third, "that the Youth Advocates and Ottawa Youth Cabinet work in partnership to advise and/or address issues with senior staff". Finally, "the Youth Advocates are expected to attend all meetings of the Ottawa Youth Cabinet and to participate as full voting members of the Committee". (p. 23-24)

The most ambitious recommendation made by the YTC in terms of resources is the established of the Youth Initiatives Program (YIP). YIP's goals are:

> To provide the financial resources to encourage the grassroots development of youth initiatives; to incorporate into the new city's alternative service-delivery and enterprise culture ideals by allowing outside not-for-profit organizations and individuals to deliver programs and carry out youth-specific projects with minimum financial assistance from the city; to foster a sense of civic pride amongst young people, by giving them both the organizational and financial tools to initiate a positive change in their communities. (p. 14)

The report makes the following YIP proposals:

- That the Ottawa Transition Board authorize the expenditure of $75,000 for the initial start-up of the Youth Initiatives Program.
- That the $75,000 be put into an account to be administered by the Corporate Services Department, with the caveat that it may not be spent until July 31, 2001.
- During the initial months of 2001, the Ottawa Youth Cabinet must define a terms of reference for the project and organize how to communicate and promote the program to the community. Additionally the structure of the YIP subcommittee of the Ottawa Youth Cabinet must also be developed.
- Post July 31, 2000, the Ottawa Youth Cabinet may spend up to $50,000 of the fund. Subsequent to this, they must raise one-half dollar for every "city dollar" spent in order to encourage community-ownership of the program.
- In the following years, the City of Ottawa will contribute $25,000 on an annual basis to the financial well-being of this program, again with the condition that one-half dollar of corporate sponsorship must be identified for every dollar spent. (p. 15)

The report also provides the following examples of "YIP-sponsored" projects: a community-oriented youth employment centre, youth entrepreneurship initiatives, the creation of youth sports/leadership programs at community facilities, a young person's leadership conference, and creative initiatives such as a drama production, an art exhibition, and a celebration of diversity event. (p. 16)

Finally, the YTC's report recommends that an annual year-end report card on the status of youth be published to measure the success of various youth-focussed initiatives taken on both by the Ottawa Youth Cabinet and the City of Ottawa, based on indicators set at the beginning of the year as well as indicators set to measure changes on an annual basis. (p. 24)

The estimated total financial implications (not including staff support) of the

recommendations made by the YTC to the Board for 2001 is $93,000. For post 2001, the estimated total annual financial implications (not including staff support or youth event costs that are to be determined by the inaugural Ottawa Youth Cabinet) is $43,300.

It is apparent that the recommendations contained in the YTC's final report "*Youth: A Place at the Table*" reflect principles and ideals which are wholly compatible and representative of a participatory model of governance. By recommending that the new Council consider giving youth a formal voice in the new City of Ottawa using the YTC's report as guide, the Board is encouraging a "dialogical" process between youth and policy makers, allowing youth the maximum possible involvement in decisions and giving them an active role to play in the overall governance of the new City. This recognition and focus on youth and youth concerns enhances the role and standing of young citizens within the new City. Such ideas are not compatible with market priorities which view citizens as consumers and taxpayers and strive to deliver services as economically as possible.

Furthermore, the YTC's recommendations appear to be incompatible with the administrative structure adopted by the Board, as it relates to policy making. As discussed earlier on in this chapter, the Board put into place an administrative structure which is top-down and relies heavily on expert council. This, coupled with the fact that the spirit and intent reflected in the YTC's final report is also found in the final reports of other project teams such as the Diversity and Community Access Project Team and the

Accessibility for the Disabled Project Team, which were also received by the Board and presented to the new Council for consideration, illustrates well the Board's ambiguity and difficulty in achieving its goal of implementing an entrepreneurial culture for the new City.

4.4 CLOSING REMARKS

Our analysis in this chapter has focussed on an attempt to outline the presence within the Board of a strong desire and intent to contain the restructuring and establishment of the new City of Ottawa within constraints reflecting market model priorities. By mapping the Board's final determinations onto the definitional elements of the two models of governance, the market model and the participatory model, as summarized in Table C, we identified amongst the major Board decisions and recommendations which ones are compatible with the market model principles discussed in Chapter 2, and which ones are not.

Our analysis has revealed that the Board was relatively successful at implementing administrative and political structures based on market priorities while having much less success when dealing with governance issues. With regards to these issues, a certain broadening of the Board's ideological perspective was noted. We also found that not all of the Board's pronouncements and intentions are reflected in its final decisions and recommendations (e.g., part-time councillors). The Board's lack of complete success in its endeavour to implement an entrepreneurial vision for the new

City of Ottawa will be the focus of our following chapter. In this final chapter, we will discuss this broadening of the Board's ideological perspective in the light of factors which may have influenced its deliberations.

5. REVIEW OF HYPOTHESIS

As stated in the introduction, although the provincial government and the Board's objectives and desired outcomes in the amalgamation process were clearly articulated at the outset, the path to their attainment was not as straightforward. As highlighted throughout the thesis, the Board's task was made more complicated by the interpretation of its mandate, which was ambitious in its scope on the one hand, and ideologically narrow on the other. These factors of scope and ideology as well as the serious time constraints imposed upon the Board (approximately one year) contributed to the creation of tensions with those who were advocating a broader ideological interpretation and a far less ambitious transition project. Furthermore, during the implementation of its mandate, the Board faced prevailing democratic constraints in the form of public opinion and media scrutiny, opposing lobbying efforts organized by concerned citizens and community organizations, as well as, diverging views expressed by elected representatives of the existing municipalities and by new council-elect hopefuls, whether officially declared as candidates, or not. As a result, all of these factors led to a variation in the Board's final determinations which reflect lesser than ideal manifestations of the entrepreneurial role of municipal government envisioned by the Board and the province for the new City. This chapter discusses the factors which influenced and eventually lead to variations in the initially projected expectations of the Board in this ambitious transition exercise.

First, the vision and scope of this restructuring exercise, as conceived by the Board, was so expansive that it required the creation and support of a very complex internal administrative structure. Also, as discussed in the previous chapter, it brought to the discussion table social issues and considerations which were not directly related to the restructuring of the local administration. Apparently caught up in its own rhetoric, the Board felt itself obliged to create project teams which, in the end, had very little to contribute to **this** particular stage of restructuring and were not necessarily compatible with their entrepreneurial vision of the role of municipal government (i.e., Governance Issues). However, once the Board settled on an expanded mandate of city building rather than simple amalgamation, it became difficult for the Board to refuse the demands of interest groups whose agendas fell outside the immediate purview of a market or entrepreneurial vision of governance. As discussed in Chapter 4, these demands (environmental, youth and other interest groups) imposed themselves as relevant and integral parts of governance responsibilities requiring policy recommendations and statements not clearly concerned with structural or purely administrative considerations.

In creating project teams to address social issues (e.g., diversity and community access, accessibility, volunteerism, and youth issues), the Board permitted and encouraged the free expression of a vision that was not compatible with its own. By definition, these teams did not share the objectives or the values promoted by an entrepreneurial vision. This undoubtedly made it more difficult to reach a consensus around the principles of the enterprise model. For the most part, the final

recommendations of these teams reflected a more expansive view of government based on community, and, therefore, did not lend themselves well to a restructuring plan based on market principles.

Within the narrow ideological perspective adopted by the Board, the aim was reducing government spending, the number of politicians, taxes, the infrastructures, personnel, in short, reducing the government's role in the public sphere.

It is clear that the application of such a profound revision of municipal governance will elicit reactions amongst the citizenry, especially amongst citizens who do not necessarily share the particular vision being advanced, but rather support a more expansive view of government based on community rather than entrepreneurial principles. Furthermore, the process of political change rarely occurs in the absence of an interplay of opposing forces. As such, the municipal amalgamation in the Ottawa-Carleton region was marked by a diversity of conflicting visions of municipal governance. Consequently, the Board was only partially successful in implementing its entrepreneurial vision.

As much as the time constraints did not allow for the application of such an ambitious restructuring project, the final result of the process was influenced by various groups and individuals who expressed their resistance. Many citizens, politicians, community groups, and members of the media did not share the entrepreneurial vision adopted by the Board and did not support its intention of completely restructuring local

government in the Ottawa-Carleton region. There are many examples that clearly illustrate this resistance. Here are but a few.

First, as previously stated in the introduction, we can trace the origins of this resistance back to the fact that the amalgamation was forced upon the citizens of Ottawa-Carleton by the provincial government. Moreover, the discontentment of this imposition was exacerbated by the fact that the seven Board members were appointed by the province and were not elected representatives. Even if many citizens supported the one-city model concept and a transition board to see its realization through to its end, the fact remains that these two assaults on democratic principles contributed to casting doubt on the legitimacy of the Board, the process, and, ultimately, on the Board's final decisions and recommendations. The expression of this doubt was most prevalent in the media via journalists, political commentators, editorialists, and letters-to-the-editor written by concerned citizens. Also, the watchdog group Boardwatch served as a vehicle of opposition to the Board's activities.

Throughout the entire process, this question of legitimacy proved to be a public relations challenge for the members of the Board. The Board was very sensitive to this criticism and went to great lengths to appear democratic, inclusive, open and transparent in their procedures and in their decision making processes. For example, when the Board unveiled its 10 Point Plan in early March 2000, it included the following two points: "The Board will ensure the citizens are informed and have ample opportunities to get involved, contribute perspectives and ideas, and participate in

shaping the new City of Ottawa" and "There will be an open and transparent process with Board operations and its work within the community. Communications on important transition issues will be regular and reliable".[40] A year later, in its final report, the Board attempted to silence its critics by pointing out the following:

> The Board implemented an extensive communications programs to ensure the public was well informed and able to participate in the transition process. By January 31st, 2001, the Board called over 30 public meetings and received both individual and group input from every sector and municipality. Public interest in the transition process was significant. Media releases and background information were regularly delivered via e-mail, fax and correspondence throughout the year to thousands of residents and community groups. Print materials were distributed to municipal offices, local libraries, and at every public meeting. At the height of the Board's activity, its website had over 1,000,000 hits per month. Its outreach database had over 10,000 names of citizens, elected representatives and businesses.[41]

The clearest indication that the Board was influenced by this criticism is revealed by the change in discourse held by the Board from the beginning to the end of the process. In the beginning, the Board spoke clearly and directly of the application of an entrepreneurial model of governance for the new City. Mid-way through their mandate, the Board spoke more of alternative service delivery (ASD) options. In the Board's final report, there is no mention of the entrepreneurial vision that had previously been espoused so vigourously. Rather, one finds references to the improvement of the

---

[40]Ottawa Transition Board. 2000. *Press Release - March 3*. Ottawa.

[41]Ottawa Transition Board. 2001. *The New City of Ottawa: Fulfilling Its Potential*. Final Report and Recommendations. Ottawa. p.8.

quality and level of services offered to the clients. This example clearly demonstrates that the Board was forced to soften its entrepreneurial rhetoric in order to win over confidence and gain acceptance of a much broader segment of the population.

Furthermore, councillors of the "old" municipalities as well as the newly elected councillors of the new City of Ottawa, elected on November 13, 2000, played an important role in shaping public opinion with regards to the work of the Board. It is obvious, and not all that surprising, that the majority of councillors did not share the Board's enthusiasm for this expanded interpretation of its task, nor did they recognize its historical significance. In fact, in their public pronouncements, many councillors distanced themselves from the Board's work and declared they would simply allow themselves to ignore its contribution when necessary. For example, the newly elected Council started its very own budget process in January 2001 before receiving the Board's municipal budget for 2001 and budget forecasts for 2002 and 2003 as required by the Terms of Reference of its mandate. (see Appendix B)

A more revealing indication of the newly elected Council's opposition to the Board's enterprise culture was the swift and unequivocal reversal of certain Board recommendations pertaining to the new political infrastructure. For example, shortly after their election, many councillors began to move their belongings into offices at City Hall. The remainder of the councillors did so early in the new year even before their first public meeting. Second, judging that their office and personnel budgets were inadequate, the newly elected Council moved swiftly to increase them to what they

believed to be more reasonable levels. Finally, the newly elected Council passed a motion increasing the number of standing committees from 4 to 6. By implementing these changes, it appears that the newly elected Council of the new City of Ottawa has no intention of running the city like a business, as was promoted by the Board.

It is thus apparent that democratic factors intervened during and after the transition process in the creation of the new City of Ottawa.

At the outset, we set out to determine to what extent the decisions and recommendations made by the Ottawa Transition Board in the creation of the new City of Ottawa reflect the vision of the entrepreneurial role of municipal government held by the Board and obviously shared by the provincial government. We mapped the Board's final decisions and recommendations as well as its activities and pronouncements onto the definitional elements of two governance models, the market model and the participatory model. This analysis was done with regards to the recommendations resulting from the work of certain key project teams which were regrouped under three main headings: Administrative Structure, Political Infrastructure and Governance Issues. Let us then summarize our findings.

First, the Board was relatively successful at implementing an administrative structure based on market priorities. However, their restructuring efforts were impeded by many factors, such as the brevity of the mandate and the limitations of the terms of reference. In addition, media scrutiny and voices of opposition amongst the general

population also played against the Board. Many citizens voiced their concerns and misgivings about Mr. Bennett and the Board's entrepreneurial agenda. It would be safe to say that the vast majority of citizens in the Ottawa-Carleton region do not support the idea of privatizing municipal services, especially social services. As a result, the Board was forced to tone down its entrepreneurial rhetoric. In the end, it could not go as far as it would have liked in this area.

Second, the Board's success rate was also fairly high with regards to implementing a political infrastructure compatible with market priorities. But here again, media scrutiny and voices of opposition coming from many different quarters played a determining role in influencing public opinion. This opposition explains Mr. Bennett's stillborn suggestion of having part-time councillors as well as the unsuccessful recommendation of having councillors set up their offices in their wards as opposed to City Hall. Furthermore, the newly elected Council went a long way in watering down the entrepreneurial character of the new political infrastructure adopted by the Board by reversing many of its main recommendations.

Third, the Board's success in creating a municipal organization based on market principles is perhaps lowest with regards to the governance issues. Dealing with these issues at this time in the restructuring process was very much in line with the Board's interpretation of its mandate - one which was ambitious in its scope and vision. In some cases, dealing with these issues at this time was inevitable (e.g., language issue). However, in many instances it was a direct result of the Board's own doing. The Board

got caught up in its own rhetoric (e.g., youth, diversity and access). For the most part, these issues were not compatible with the enterprise model that was being promoted so vigourously by the Board.

Therefore, our initial hypothesis has been confirmed that factors related to time, mandate and democratic process lead to a variation in the Board's final determinations which reflect lesser than ideal manifestations of the entrepreneurial role of municipal government envisioned by the Board and the province for the new City.

The municipal amalgamation exercise undertaken in the Ottawa-Carleton region in the year 2000 illustrates an on-going debate in democratic societies today. This debate has to do with level of participation: How much is possible and desirable? How often should we participate and where? As argued by Ann Phillips (1991), democracy necessitates participation from people, including citizens, elected representatives, as well as stakeholders (i.e., employees). As was evident in our case study of the transition process in the former Ottawa-Carleton region, these many threads were inter-related. The issues at hand as well as the very nature of the dynamics of these groups within a democracy are, in practical terms, the same. People need and want to be involved in the decisions that directly affect their lives. This is true whether it be in the public or private realms. Our study focussed exclusively on the actions and decisions of the Ottawa Transition Board. Future research and studies would need to look more closely at the question of how a society understands democracy. In other words, how to balance the principles of representative democracy and the interests of participatory

democracy.

BIBLIOGRAPHY

Aitchison, J. H. 1949. "The Municipal Corporations Act of 1849", *Canadian Historical Review*, 30 (June).

Aucoin, Peter. 1995. *The New Public Management: Canada in Comparative Perspective*. Montreal: Institute for Research on Public Policy (IRPP).

Bookchin, Murray. 1992. *Urbanization Without Cities: The Rise and Decline of Citizenship*. Montreal: Black Rose Books.

Borins, Sandford. 1995. "The new public management is here to stay," *Canadian Public Administration*, Spring, pp. 122-132.

Broad, Dave and Wayne Antony (eds.) 1999. *Citizen or Consumers? Social Policy in a Market Society*. Halifax: Fernwood Publishing.

Campbell, Colin. 1983. *Governments Under Stress*. Toronto: University of Toronto Press.

Canadian Broadcasting Corporation. 2000. "Transition board undemocratic, critics charge", *CBC Radio*, webposted May 9, 2000.

Colton, Timothy J. 1980. *Big Daddy*. Toronto: University of Toronto Press.

Crawford, Kenneth Grant. 1954. *Canadian Municipal Government*. Toronto: University of Toronto Press.

Denley, Randall. 2000. "Democratic advice for our dictators", *The Ottawa Citizen*, February 16, 2000.

Denley, Randall. 2000. "So this is how a dictatorship reaches out", *The Ottawa Citizen*, March 15, 2000.

Drucker, Peter J. 1985. *Innovation and Entrepreneurship: Practice and Principles*. New York: Harper & Row, Publishers.

Gervais, Marc. 2001. *Public Record of the Ottawa Transition Board*, Publication of the Ottawa Transition Board in collaboration with the University of Ottawa, Ottawa, Ontario.

Government of Ontario. 2000. *News Release Communiqué - January 26, 2000.* Toronto: Ministry of Municipal Affairs and Housing.

Government of Ontario. 1999. *Fewer Politicians Act.* Toronto: Printed by the Legislative Assembly of Ontario.

Government of Ontario. 1996. *A Guide to Municipal Restructuring.* Toronto: Ministry of Municipal Affairs and Housing.

Graham, Katherine A., Allan M. Maslove and Susan D. Phillips. 2000. *Learning from Experience? Ottawa as a Cautionary Tale of Reforming Ontario's City-Regions.* To be published soon.

Graham, Katherine A. and Susan D. Phillips. 1998. *Citizen Engagement: Lessons in Participation from Local Government.* Toronto: The Institute of Public Administration of Canada.

Graham, Katherine A. and Susan D. Phillips with Allan M. Maslove. 1998. *Urban Governance in Canada.* Toronto: Harcourt Brace & Company.

Hood, Christopher. 1991. "A Public Management for All Seasons?", *Public Admistration*, Vol. 69, no. 1 (Spring), pp. 3-19.

Isin, Engin F. 1995. "The Origins of Canadian Municipal Government", in James Lightbody, ed., *Canadian Metropolitics.* Toronto: Copp Clark Ltd., 51-91.

Isin, Engin F. 1992. *Cities Without Citizens. The Modernity of the City as a Corporation.* Montreal: Black Rose Books.

Jeffrey, Brooke. 1999. *Hard Right Turn: The New Face of Neo-Conservatism in Canada.* Toronto: Harper Collins Publishers Ltd.

Johnson, A. W. 1993. *What is Public Management? An Autobiographical View.* Ottawa: Canadian Centre for Management Development.

Keating, Michael. 1995. "Size, Efficiency and Democracy: Consolidation, Fragmentation and Public Choice, pp.117-134, in David Judge, Gerry Stoder and Harold Wolman (eds.), *Theories of Urban Politics*, London: Sage Publications.

Kernaghan, Kenneth. 1992. "Empowerment and public administration: revolutionary advance or passing fancy?", *Canadian Public Administration*, Volume 35, No. 2 (Summer), pp.194-214.

Langrod, Georges. 1953. "Local Government and Democracy", in *Public Administration*, Vol. XXXI, Spring, pp. 25-33.

Lanthier, Andrea. 2000. "Board 'boys club' blasted", *The Ottawa Sun*, July 23, 2000.

Leroux, Jacki. 2000. "Board has little time for public", *The Ottawa Sun*, March 8, 2000.

Leroux, Jacki. 2000. "Critics fear moves too far, too fast to be good", *The Ottawa Sun*, July 16, 2000.

McLean, Iain. 1987. *Public Choice: An Introduction.* Oxford: Basil Blackwell Ltd.

Mueller, Dennis C. 1979. *Public Choice: An Introduction.* New York: Basil Blackwell Inc.

O'Brien, Allan. 1993. *Municipal Consolidation in Canada and its Alternatives.* Toronto: ICURR Publications.

Osborne, David and Ted Gaebler. 1992. *Reinventing Government: How the Entrepreneurial Spirit is Transforming The Public Sector.* Don Mills, Ontario: Addison-Wesley Publishing Company, Inc.

Ottawa Transition Board. 2001. *The New City of Ottawa: Fulfilling Its Potential.* Final Report and Recommendations. Ottawa.

Ottawa Transition Board. 2000. *Board Decisions - June 12.* Ottawa.

Ottawa Transition Board. 2000. *Board Decisions - August 28.* Ottawa.

Ottawa Transition Board. 2000. *Board Decisions - December 19.* Ottawa.

Ottawa Transition Board. 2000. *Chairman's Opening Address - February 14.* Ottawa.

Ottawa Transition Board. 2000. Final Report of the Youth Transition Committee. "*Youth: A Place at the Table*". Ottawa.

Ottawa Transition Board. 2000. *Press Communiqué - April 10.* Ottawa.

Ottawa Transition Board. 2000. *Press Release - March 3.* Ottawa.

Ottawa Transition Board. 2000. *Press Release - August 14.* Ottawa.

Ottawa Transition Board. 2000. *Project Charter - City of Ottawa Transition.* Ottawa.

Ottawa Transition Board. 2000. *Project Charter for Service Restructuring Teams.* Ottawa.

Ottawa Transition Board. 2000. *Public Consultation Sessions - August 1-3, 2000.* Ottawa.

Ottawa Transition Board. 2000. *Service Delivery Model. High Level Business Case Analysis: A Workbook for Service Restructuring Teams.* Ottawa.

Paquet, Gilles. 1999. *Governance Through Social Learning.* Ottawa: University of Ottawa Press.

Paquette-Legault, Diane. 2000. "Un comité de vigilance accuse le Conseil de transition de manque de transparence", *Le Droit*, April 11, 2000.

Peters, B. Guy. 1993. *The Public Service, The Changing State and Governance.* Ottawa: Canadian Centre for Management Development.

Peters, B. Guy and Donald J. Savoie (eds.). 1995. *Governance in a Changing Environment.* Montreal & Kingston: McGill-Queen's University Press.

Phillips, Anne. 1991. *Engendering Democracy.* University Park, Pennsylvania: The Pennsylvania State University Press.

Pierre, Jon. 1995. "The Marketization of the State: Citizens, Consumers, and the Public Market", pp. 55-81, in Peters, B. Guy and Donald J. Savoie (eds.). *Governance in a Changing Environment.* Montreal & Kingston: McGill-Queen's University Press.

Progressive Conservative Party of Ontario. 1998. *Blueprint: Mike Harris' Plan to Keep Ontario on the Right Track.*

Progressive Conservative Party of Ontario. 1994. *The Common Sense Revolution.*

Sancton, Andrew. 2000. *Merger Mania: The Assault on Local Government.* Montreal and Kingston: McGill-Queen's University Press.

Sancton, Andrew. 1994. *Governing Canada's City-Regions: Adapting Form to Function.* Montreal: The Institute for Research on Public Policy.

Sancton, Andrew. 1993. *Local Government Reorganization in Canada Since 1975.* Toronto: Intergovernmental Committee on Urban and Regional Research Press.

Savoie, Donald J. 1995. "What is wrong with the new public management?" , *Canadian Public Administration*, Spring, pp. 112-121.

Schachter, Hindy Lauer. 1997. *Reinventing Government or Reinventing Ourselves. The Role of Citizen Owners in Making a Better Government.* New York: State University of New York Press.

Seidle, Leslie. 1995. *Rethinking the Delivery of Public Services to Citizens.* Montréal: Institute for Research on Public Policy.

Self, Peter. 1993. *Government by the Market? The Politics of Public Choice.* Boulder & San Francisco: Westview Press.

Shortliffe, Glenn S. 1999. *Report of the Special Advisor: Local Government Reform in the Regional Municipality of Ottawa-Carleton* (Shortliffe Report).

The Ottawa Citizen Editorial Staff. 2000. "No, your vote doesn't count", *The Ottawa Citizen*, January 28, 2000.

The Ottawa Citizen Editorial Staff. 1999. "New city doesn't need czars", *The Ottawa Citizen*, November 30, 1999.

Tindal, C.R. and S. Nobes Tindal. 2000. *Local Government in Canada*, 5rd ed. Scarborough: Nelson - Thomson Learning.

Vojnovic, Igor. 2000. *The Transitional Impacts of Municipal Consolidations.* To be published soon.

Weaver, John C. 1984. "Tomorrow's Metropolis' Revisited: A Critical Assessment of Urban Reform in Canada, 1890-1920", in Gilbert A. Stelter and Alan F. J. Artibise, eds., *The Canadian City: Essays in Urban and Social History.* Ottawa: Carleton University Press, pp. 456-477.

Websites:

ottawaboardwatch.org

ottawa.cbc.ca

ottawatransition.on.ca

APPENDIX A

Local Government Reform in Ottawa-Carleton[1]

The current system of local government in the Ottawa-Carleton region originated in 1969 with the establishment of the Regional Municipality of Ottawa-Carleton (RMOC). Over the last thirty years, the Ontario government as well as the Regional and municipal councils have set up commissions, committees, and a citizen-based panel to study the system in place and offer recommendations on ways of improving the system in place. Some recommendations have been accepted and implemented while others have not. The result has been a progressive consolidation of responsibilities in the region. At the time of amalgamation, the RMOC was responsible for approximately 80% of the total municipal tax expenditures of the Ottawa-Carleton region. The following is a brief historical overview of these major reform initiatives undertaken in Ottawa-Carleton in the past thirty years.[2]

January 1, 1969	Creation of a regional government composed of 16 lower tier municipalities and the Regional Municipality of Ottawa-Carleton (RMOC) The 16 municipalities include the Cities of Eastview (Vanier) and Ottawa, the Villages of Richmond, Rockcliffe Park and Stittsville, as well as the Townships of Torbolton, Osgoode, North Gower, Nepean, Marlborough, March, Huntley, Goulbourn, Gloucester, Fitzroy and Cumberland.

---

[1]Taken from Marc Gervais. 2001. *Public Record of the Ottawa Transition Board*, Publication of the Ottawa Transition Board in collaboration with the University of Ottawa, Ottawa, Ontario.

[2]For a detailed assessment of the political evolution of the Regional Municipality of Ottawa-Carleton, up to 1990, consult: J. Douglas Cameron, Q. C., *A Quarter Century of Political Evolution, 1965-1990*, Ottawa-Carleton Legal Department, 1990; for a historical perspective of the major governance reviews in the Ottawa-Carleton region consult: Region of Ottawa-Carleton, "History of Regional Government and History of Governance Review in Ottawa-Carleton", pp. 23-30, in *Review of Governance in Ottawa-Carleton*, November 1996; Ontario, Municipal Affairs and Housing, "Appendix G: A Historical Perspective of Governance Reviews for the Ottawa-Carleton Region", in *Report to the Minister: Local Government Reform in the Regional Municipality of Ottawa-Carleton*, 25 November 1999.

December 4, 1973	Original 16 lower tier municipalities reduced to 11 as follows: • The Township of West Carleton was created by the amalgamation of the former Townships of Fitzroy, Huntley and Torbolton; • The Rideau Township was created from the former Townships of Marlborough and North Gower with annexation of small parts of Gloucester and Nepean and Osgoode Townships in the Manotick area; • The present Goulbourn Township was formed by the amalgamation of the Villages of Richmond and Stittsville with the former Goulbourn Township.
October 1976	*Report of the Ottawa-Carleton Review Commission* (Henry B. Mayo)[3] Mandate: "*to examine, evaluate, and make recommendations on the structure, organization and operations of local government in the Ottawa-Carleton area.*"[4] Main recommendation: that the two-tier system of government be retained with some change in boundaries
1987-88	*Report of the Ottawa-Carleton Regional Review Commission, Vol. 1 and Vol. 2* (David Bartlett)[5] Mandate: to examine political representation in the region and the allocation of functions and finances between the two tiers Main recommendations: -that the regional level be given responsibility for economic development

---

[3]See H. B. Mayo, *Report of the Ottawa-Carleton Review Commission*, Ottawa, 1976.

[4]*Ibid.*, p. xi.

[5]See D. W. Bartlett, *Ottawa-Carleton Regional Review Phase I Report: Accountability and Representation*, Ottawa, 1987, and D. W. Bartlett, *Ottawa-Carleton Regional Review Phase II Report: Functions and Finances*, Ottawa, 1988.

-that the regional chair and the majority of regional councillors be directly elected

November 1990

Report of the Electoral Boundaries Commission for Ottawa-Carleton (Katherine A. Graham)[6]

Mandate: to develop a new system of election boundaries for the regional and local governments in the area

Main recommendation: that a number of the electoral districts for the regional council cross local municipal boundaries

November 1991

Direct election of Regional Chair

November 1992

Ottawa-Carleton Regional Review Commission: Final Report (Graeme Kirby)[7]

Mandate: "*To consult with municipalities and the public on the degree of interest and support for structural reform to municipal government in Ottawa-Carleton and for the direct election of members to regional council.*"[8]

Main recommendations:
-that the two-tier structure of government be retained
-that the responsibilities of Regional government be increased
-that elections be based on Graham's model of cross border boundaries
-that the regional councillors be directly elected by the citizens of the Region

---

[6]See K. A. Graham, *Report of the Ottawa-Carleton Electoral Boundaries Commission*, Ottawa, 1990.

[7]See Graeme M. Kirby, *Ottawa-Carleton Regional Review Commission: Final Report*, Ottawa, 1992.

[8]*Ibid.*, p. 1.

1994	Direct election of regional councillors Exclusion of the mayors of the local municipalities from the regional council
January 1, 1995	Establishment of a single Ottawa-Carleton Regional Police Force
August 1997	*Citizen's Panel on Local Governance in Ottawa-Carleton* established at the request of the councils of the Regional Municipality of Ottawa-Carleton and the 11 local municipalities. Its mandate was to recommend, in consultation with the citizens of Ottawa-Carleton, a new structure for local government.
March 1998	*Citizen's Panel* disbanded before completing its mandate.[9] In April 1998, co-chairs Grete Hale and Diane Desaulniers wrote in an open letter to the residents of Ottawa-Carleton: "*We felt that the opportunity for an objective and open-minded process has been lost due to the promotion by others of entrenched positions and particular governance models.*"
1998	Although many different options for local governance in Ottawa-Carleton have been discussed through the years, the following models received the most attention by the citizens of Ottawa-Carleton and their elected officials in recent years:

[9]See Citizens' Panel on Local Governance in Ottawa-Carleton, *Challenges and Choice: Options for Local Governance in Ottawa-Carleton: A Discussion Paper*, February 1998.

- **One City - One Tier:**[10] This model proposed the creation of a new single tier city composed of the 12 municipalities in the Ottawa-Carleton region.

- **The Tri-City Model:**[11] This model proposed the abolition of the Regional Municipality of Ottawa-Carleton and the creation of three new cities composed of Ottawa, Vanier and Rockcliffe Park; Cumberland and Gloucester; Kanata and Nepean. Under this model, common services would be provided by an Authority composed of the new Mayors and some elected Councillors.

- **The Four Cities Model:**[12] The only difference in this model from the Tri-City model is that Nepean and Kanata would continue to be separate cities.

- **The Rural Alliance:**[13] In the Rural Alliance model, services would be provided by four local municipalities (The Townships of West Carleton, Goulbourn, Rideau, and Osgoode) and would be retained at the same level as is currently provided by the Regional Municipality of Ottawa-Carleton.

September 24, 1999 — Appointment of the Special Advisor Glen Shortliffe by the Government of Ontario.

Mandate: To recommend reforms in municipal government in the Ottawa-Carleton region which would lower taxes, improve services, reduce bureaucracy, clarify lines of

---

[10]See City of Ottawa, *One City, One Tier, One Taxpayer: Proposed Model of Governance*, April 28, 1998.

[11]See Cities of Gloucester, Kanata and Nepean, *Right-Sizing Government in Ottawa-Carleton: The Tri-City Model*, June 1998.

[12]See City of Kanata, *The Right Size of Governance in Ottawa-Carleton: The Four City Model*, October 22, 1999.

[13]See BDO Dunwoody, *Rural Alliance Scenarios: Townships of Goulbourn, Osgoode, Rideau, West Carleton*, August 1997; and BDO Dunwoody, *Rural Alliance Scenarios: Townships of Goulbourn, Osgoode, Rideau, West Carleton*, October 1999.

responsibility, and foster greater municipal accountability.

November 25, 1999 — *Report of the Special Advisor: Local Government Reform in the Regional Municipality of Ottawa-Carleton* (Shortliffe Report)[14] recommending a One City - One Tier government with community satellite offices.

As summarized by the Special Advisor on page 8 of his report, highlights of his recommendations are:

- all existing municipalities and the Regional Municipality of Ottawa-Carleton to be dissolved and one new city called "Ottawa" to be legislatively established as a single tier municipality,
- the City of Ottawa to be legislatively designated a bilingual city with services to be provided in both official languages where warranted,
- the existing 18 regional wards, their names and boundaries to be continued as the electoral wards of the new City,
- Ottawa City Hall to be located at 111 Lisgar St.,
- the Mayor to be elected at large and the 18 Councillors by ward. All elected officials to be full-time,
- for the first two terms of the new Council the representatives from the Western Townships and Cumberland / Osgoode to have 2 votes each on all matters before Council, and on the proposed Standing Committees, in order to ensure enhanced representation of rural interests,
- for the first term of Council, six Standing Committees to be established to meet publicly. This will allow citizens greater accessibility to local decision-making and would contribute to greater accountability by Councillors. The Committees will be:
    - Corporate Services / Planning & Development,
    - Community Services,
    - Emergency Services,
    - Library Services,
    - Rural Affairs (responsibilities to include agricultural lifestyle, drainage and other issues of particular interests to rural citizens), and
    - Transportation and Works.

The Standing Committees will meet in different geographic locations in the new City.

---

[14]See Ontario, Municipal Affairs and Housing, *Report to the Minister: Local Government Reform in the Regional Municipality of Ottawa-Carleton*, 25 November 1999.

Also,

- a leaner City administration to be created as set out in the Special Advisor's Report,
- eight Community Satellite Offices (location guidelines included in the Report) to be created to provide front-line services to ensure accessibility and citizen engagement. These offices will provide some services, information, small community meeting rooms and other facilities. These Satellite locations will also provide office space for the ward Councillors to meet with constituents and to liaise on their behalf with City Hall,
- the ward Councillors to be, subject to the new City operating policies, responsible for such matters as parks, recreation facilities, community services, animal control, etc.,
- the five municipal electrical utilities to be amalgamated into one utility, and
- the former Village of Rockcliffe Park to remain a heritage area.

The Shortliffe Report also made many other recommendations for cost savings and increased efficiencies and fairness such as average levelling of salaries rather than raising them to the highest rate, pooling reserves and debts, amalgamating programs, and area rating of fire and transit services. The total savings of a One City - One Tier model were estimated at between $70 and $80 million.

December 22, 1999	Bill 25, *An Act to provide for the restructuring of four regional municipalities and to amend the Municipal Act and various other Acts in connection with municipal restructuring and with municipal electricity services*, receives Royal Assent, thereby enabling the amalgamation of the 11 municipalities that make up the Regional Municipality of Ottawa-Carleton and the RMOC. (The short title of this Act is the *Fewer Municipal Politicians Act, 1999.*)
January 26, 2000	Creation of the Ottawa Transition Board
November 13, 2000	Municipal elections - Election of the new City of Ottawa's first council and mayor
January 1, 2001	Birth of a «new» city: The City of Ottawa

January 7, 2001	New council and mayor officially sworn in
January 31, 2001	Ottawa Transition Board's mandate ends

APPENDIX B

February 8, 2000

Terms of Reference for the Transition Boards

Background:

The reform of municipal governments in Haldimand-Norfolk, Hamilton-Wentworth, Ottawa-Carleton and Sudbury will mean tax savings and fewer municipal politicians.

These reforms have been preceded by over 100 other restructuring initiatives across the province. The Government and the municipal sector have learned many lessons from these experiences. Key among them is the importance of effective transition planning and management. In this case, the Minister's authority to establish a Transition Board and the Board's purpose were set out in legislation:

"The primary function of the transition board is to facilitate the transition from the old municipalities and their local boards to the city and its local boards:

- by controlling the decisions of the old municipalities and their local boards that could have significant financial consequences for the city and its local boards; and
- by developing business plans for the city and its local boards in order to maximize the efficiency and costs savings of this new municipal structure." (Section 18 in the Greater Sudbury and Hamilton schedules; Section 21 in the Haldimand and Norfolk schedules; Section 19 in the Ottawa schedule.)

The January 26th "Backgrounder" accompanying the press release announcing the Transition Board members sets out the purpose of the Boards and outlines their executive, business planning and controllership powers. This is attached.

Regulations setting out the details of the Board's powers and duties have now been prepared and are attached. These regulations, along with the legislation and these terms of reference should enable the Transition Board to do its job of ensuring a smooth, effective transition to the new municipal structure(s).

Transition Board's Operations:

The Board is subject to rules about conflict of interest, freedom of information and protection of privacy and open meetings, and is required to have a procedural by-law. The Board is also to establish operational policies, including policies about:

- how and when to obtain public input on its business planning activities
- how to report to the public on its recommendations and decisions throughout the

year
- the process to be used for hiring and contracting
- the process to be used for paying expenses incurred by the Board
- the safekeeping of the Board's files and records.

Controllership Powers of the Transition Board

The Board is to issue guidelines with respect to the year 2000 budgets, overseeing the decisions of the existing municipalities and their local boards that could have significant financial consequences for the city and its local boards. More specifically, it is to approve the types of transactions set out in the regulations - "Powers and Duties of Transition Board".

Business Planning Powers of the Transition Board

The Board's role is to recommend, for the new municipality and its local boards, a year 2001 budget and suggested tax levy and to forecast year 2002 and 2003 expenditures and revenues. These recommendations are to be made in the context of the Special Advisor's findings about potential tax savings. In building the budget and forecasts, the Transition Board is to identify its assumptions about the following matters:

- the core business of the municipality and its local boards
- the scope and level of services to be provided, including service level harmonization and where services should be provided in French
- the expected performance outcomes (in measurable terms) in each of the key service areas*
- the appropriate methods of service delivery and the inputs required to deliver the expected level of performance (including: the staff complement and volunteers; information technology; accommodation; equipment; service management contracts) and the projected costs of these inputs
- the human resource implications resulting from 1-4, including those service areas where there are projected surpluses or deficits of staff, where re-skilling will be required, what monetary entitlements or other benefits should be provided for surplus employees, and how to enable the new municipality to be an employer of choice for the staff that it does need
- the redundant assets that should be disposed of.

*Note that this is an iterative process so that the expected performance outcomes (item 3) may have to be adjusted after full exploration of the inputs (item 4) required to deliver the expected level of performance.

In its business planning role, the Board is to be guided by the following principles:

- Best value - The benefits of its recommended method of service delivery should clearly outweigh the benefits of other available options. Significant productivity

gains for taxpayers should be an outcome of the restructuring.

- Sustainability - Sustainable municipal infrastructure and services financed at competitive tax and utility rates.
- Materiality - The focus of service improvements (economy, efficiency and effectiveness) should be on the functions of significant expenditure, not simply the amalgamating functions.
- Community governance - Municipalities play an important role, not just in delivering services, but in representing the community before other governments, community organizations and businesses.

Financing Methods:

The Board is also to recommend the financing methods for the municipal services in the new municipality, including:

- which services should be area rated, the "urban service" area for the rate and the minimum dollar thresholds required to implement and area rating scheme
- whether differences in service levels should be area rated, the "urban service" area for this purpose and the minimum dollar thresholds required to implement a service level area rating scheme
- how assets and liabilities of the former municipalities should be distributed, and which (whole or partial) assets and liabilities should not be shared by all of the new municipality's ratepayers. This should include the matter of whether the new municipality should retain reserves and reserve funds, other than the working capital reserve, of the former municipalities for the purpose for which they were established. If not, the Board would be expected to recommend how reserves and reserve funds should be treated, i.e., area rated (in whole or in part) back to the former municipality or used to offset the general levy of the new municipality
- whether to phase in tax changes arising from the restructuring and, if so, which costs to phase in, the time period for the phase-in and the floors and ceilings for the phase-ins.

Governance and decision-making systems

The Transition Board is also to make recommendations about the governance and decision-making systems that should be put in place for the municipality to plan and implement its business. Recommendations are required on the following:

- the role of the municipal councillor, council and staff
- the salary and compensation of municipal councillors and the appropriate type and level of administrative support, if any
- the council committee structure
- the accountability framework for local boards (those matters beyond statutory requirements, e.g., frequency, timeliness of financial planning & reporting; HR,

IT, purchasing protocols, types of performance measures & program reviews required)
- the strategic & business planning processes
- the alternative service delivery framework
- the decision-making processes and protocols for the municipal administration.

In developing its advice about governance and decision making systems, the Board is to consider the following principles:

- community participation and the need to design the councillor role to make it easier for leading citizens to consider serving the community in civic office.
- accountability - the need to focus political accountability on policy and community governance; allows professional management and staff accountability for service management and customer service.
- cohesive governance - council decision making structures and processes should encourage the council to act as a team and a corporate body.

Executive Powers of the Transition Board

The Board's executive powers are as set out in legislation and the regulations. These include:

- hiring the new municipality's officers required by statute and any other employees of executive rank that the Transition Board considers necessary for the good management of the municipality
- establishing the key elements of the municipality's organizational structure to implement its business plan
- determining the disposition of municipal electric utility assets, e.g., incorporation, sale, lease back
- negotiating with unions (and where unsuccessful, apply to the Ontario Labour Relations Board) to determine the bargaining units and agents for the new municipality
- preparing a procedural by-law for the new Council and determine, in consultation with the incoming council, the date and time of the first meeting of the new council
- establishing and testing electronic or manual information systems, records and books of accounts for the new municipality and its local boards
- appointing the Returning Officer for the year 2000 municipal election.

In hiring executive staff, designing the organization and determining the bargaining units, the Board is to:

- pay particular attention to the opportunity to create a new restructured, customer focussed organization, not just an amalgamated organization. Experience in the design and implementation of customer focussed approaches should be sought

in recruiting executive staff.

- build service delivery and organizational arrangements with the customer as the starting point.
- design organizational structures and systems to promote organizational learning, i.e., be designed to permit a high capacity for the organization to learn about and fit with its customers' changing needs.

MQ

66045

UMI

MICROFILMED 2002

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

Université d'Ottawa • University of Ottawa

PERMISSION DE REPRODUIRE ET DE DISTRIBUER LA THÈSE

PERMISSION TO REPRODUCE AND DISTRIBUTE THE THESIS

NOM DE L'AUTEUR / NAME OF AUTHOR: GORSE, Nicolas
ADRESSE POSTALE / MAILING ADDRESS: 50, rue Charles, Ottawa (Ontario), K1M 2H2
GRADE / DEGREE: M.C.S. — **ANNÉE D'OBTENTION / YEAR GRANTED:** 2001
TITRE DE LA THÈSE / TITLE OF THESIS: The Feature Interaction Problem: Automatic Filtering and Generation of Validation of Test Suites at the Design Stage

L'auteur permet, par la présente, la consultation et le prêt de cette thèse en conformité avec les règlements établis par le bibliothécaire en chef de l'Université d'Ottawa. L'auteur autorise aussi l'Université d'Ottawa, ses successeurs et cessionnaires, à reproduire cet exemplaire par photographie ou photocopie pour fins de prêt ou de vente au prix coûtant aux bibliothèques ou aux chercheurs qui en feront la demande.

Les droits de publication par tout autre moyen et pour vente au public demeureront la propriété de l'auteur de la thèse sous réserve des règlements de l'Université d'Ottawa en matière de publication de thèses.

The author hereby permits the consultation and the lending of this thesis pursuant to the regulations established by the Chief Librarian of the University of Ottawa. The author also authorizes the University of Ottawa, its successors and assignees, to make reproductions of this copy by photographic means or by photocopying and to lend or sell such reproductions at cost to libraries and to scholars requesting them.

The right to publish the thesis by other means and to sell it to the public is reserved to the author, subject to the regulations of the University of Ottawa governing the publication of theses.

N.B. LE MASCULIN COMPREND EGALEMENT LE FEMININ

08/02/2001

DATE (AUTEUR) SIGNATURE (AUTHOR)



Université d'Ottawa • University of Ottawa



Université d'Ottawa • University of Ottawa

FACULTÉ DES ÉTUDES SUPÉRIEURES
ET POSTDOCTORALES

FACULTY OF GRADUATE AND
POSTDOCTORAL STUDIES

GORSE, Nicolas
AUTEUR DE LA THÈSE - AUTHOR OF THESIS

Master of Computer Science
GRADE - DEGREE

SITE
FACULTE, ECOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

THE FEATURE INTERACTION PROBLEM: AUTOMATIC FILTERING OF INCOHERENCES & GENERATION OF VALIDATION TEST SUITES AT THE DESIGN STAGE

Luigi Logrippo
DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

B. Pagurek

P. Scott

J.-M. Weber

J.-M. De Koninck, Ph.D.
LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

SIGNATURE

DEAN OF THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

The Feature Interaction Problem: Automatic Filtering of Incoherences & Generation of Validation Test Suites at the Design Stage

Nicolas Gorse

Thesis submitted to the
School of Graduate Studies and Research
in partial fulfillment of
the requirements for the degree of

Master of Computer Science

under the auspices of the
Ottawa-Carleton Institute for Computer Science

University of Ottawa
Ottawa, Ontario, Canada
September, 2000

© Nicolas Gorse, Ottawa, Canada, 2000

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-66045-1

Canada

Abstract

This thesis addresses the feature interaction problem in the research area of telephony systems design. We start with a review of this problem and present our definition of a feature and of a feature interaction. We enumerate and discuss some known approaches with respect to their strengths and limitations.

We present our initial approach and propose some refinements in order to automate the detection of incoherences and derivation of validation scenarios at the design stage using predicate logic and Prolog. The identification of incoherences corresponding to potential feature interactions allows the designer to refine the requirements and to produce a better specification.

A UCM model is derived from requirements. This model allows the designers to have a general point of view of the system and to detect design defaults. The construction of the UCM model is followed by the generation of a LOTOS specification.

The derivation of validation test suites and their application against the specification permits to check whether or not the incoherences identified in the requirements lead to feature interactions.

The results obtained over two case studies show that the application of this approach improves the design and validation of a specification.

Acknowledgement

I would like to thank all the people who helped me in completing this thesis.

First, I would like to thank my supervisor, Dr. Luigi Logrippo, for his encouragement, guidance and assistance through my graduate studies.

Second, I would like to thank the members of the University of Ottawa LOTOS group for their advises and helpful comments. I would like to thank particularly Jacques Sincennes who suggested the idea of using predicate logic and Prolog to identify incoherences and who supported me through long discussions.

Third, I would like to thank for their financial support: the University of Ottawa, Mitel Corporation, and Communications and Information Technology Ontario (CITO).

Last but not least, I dedicate this thesis to my wife, Sandra Jacqueson, who always supported me and listened patiently to all my endless monologues.

Contents

1 Introduction **11**
- 1.1 Background . . . 11
- 1.2 Motivation . . . 11
- 1.3 The Proposed Approach . . . 12
  - 1.3.1 Filtering . . . 12
  - 1.3.2 Validation . . . 12
- 1.4 Organization of the Thesis . . . 13

2 Framework **15**
- 2.1 The Feature Interaction Problem . . . 15
  - 2.1.1 Definition of a Feature . . . 15
  - 2.1.2 Definition of a Feature Interaction . . . 16
  - 2.1.3 Prominence in Telecommunication Systems . . . 17
- 2.2 Existing Approaches . . . 18
  - 2.2.1 Avoidance & Detection Using Patterns . . . 18
  - 2.2.2 Avoidance & Detection Using Use Case Maps . . . 18
  - 2.2.3 Detection Using LOTOS . . . 19
  - 2.2.4 Detection Using Permutation Symmetry . . . 19
  - 2.2.5 Detection Using Temporal Logic . . . 20
  - 2.2.6 Static Detection Method Without State Enumeration . . . 20
  - 2.2.7 Related Problems . . . 20
- 2.3 Structure of the Project . . . 21
  - 2.3.1 Initial Approach . . . 22
  - 2.3.2 Refinements . . . 23
- 2.4 The Proposed Filtering and Validation Process . . . 24
- 2.5 In Summary . . . 26

3 Use Case Maps, LOTOS & Mitel's System **27**
- 3.1 Presentation of Use Case Maps . . . 27
  - 3.1.1 Basic Notation & Interpretation . . . 27
  - 3.1.2 Scenario Paths Notation . . . 28
  - 3.1.3 Components Notation . . . 31
  - 3.1.4 Additional Notations . . . 32
- 3.2 Mitel System . . . 34
  - 3.2.1 Use Case Maps Model . . . 34
  - 3.2.2 Features Considered . . . 35

3.3 From Use Case Maps to LOTOS . . . 36
3.3.1 LOTOS Concepts . . . 36
3.3.2 Analysis & Decisions . . . 39
3.3.3 Features Integration . . . 41
3.4 In Summary . . . 43

4 Identification of Incoherences – Theory **45**
4.1 Features Representation . . . 45
4.1.1 Basic Principles . . . 45
4.1.2 Formalism . . . 46
4.1.3 Example . . . 47
4.2 Incoherences . . . 49
4.2.1 Definitions . . . 49
4.2.2 Direct Incoherence Rules . . . 50
4.2.3 Transitive Incoherence Rules . . . 58
4.3 Priorities . . . 62
4.3.1 Representing Priorities Using Contradiction Pairs . . . 62
4.3.2 Limitation & Possible Solution . . . 62
4.4 In Summary . . . 63

5 A Tool for Feature Interaction Detection **65**
5.1 Scenario Based Validation . . . 65
5.1.1 Testing Strategy . . . 66
5.1.2 Testing Using LOLA . . . 66
5.1.3 Feature Interaction Testing . . . 67
5.2 Automatic Incoherences Filtering & FI Detection . . . 69
5.2.1 Implementation of Feature Descriptions & Rules into Prolog . . . 69
5.2.2 Incoherences Filtering . . . 72
5.2.3 Filtering Algorithm . . . 73
5.3 Automatic Test Suite Generation . . . 75
5.3.1 From Properties to Scenario Parts . . . 76
5.3.2 Test Suite Principles . . . 77
5.3.3 Scenarios Generation . . . 78
5.4 In Summary . . . 81

6 Application & Evaluation **83**
6.1 Application . . . 83
6.1.1 Case Study 1 – Mitel . . . 83
6.1.2 Case 2 – The Feature Interaction Workshop 2000 . . . 91
6.2 More Complicated Examples . . . 94
6.3 Performance Evaluation . . . 95
6.3.1 Human Aspects Considerations . . . 95
6.3.2 Algorithmic Performance . . . 95
6.4 Comparison . . . 98
6.5 In Summary . . . 100

7 Conclusion & Future Work **101**
7.1 Achievements 101
7.2 Contributions 102
7.2.1 Translation of a UCM Model into a LOTOS Specification 102
7.2.2 Incoherences Filtering Process 103
7.2.3 Automatic Derivation of Validation Test Suites 103
7.2.4 Scenario Based Validation of a Large LOTOS Specification 103
7.2.5 Use of LOTOS in an Industrial Context 103
7.2.6 Applications & Conclusive Results over two Case Studies 104
7.3 Further Research 104
7.3.1 Priorities 104
7.3.2 Multi-way Incoherences 104
7.3.3 Distribution over a Many Hosts Network 105
7.3.4 Considering other methods 105

A Filtering Results for Mitel **111**

B Prolog Implementation **123**

Chapter 1

Introduction

1.1 Background

Specification and testing of complex distributed systems have always been a challenge. Today's telephone systems are often decomposed into a base system *(the kernel)* and features [1, 2]. The base system implements the basic behavior, such as simple calls management, while features are optional units or increments of functionality that are available in the system and that can be used to implement more complicated behaviors such as *Call Waiting* or *Call Forward on Busy Line*. A user can then choose to subscribe to the use of some of these features depending on his/her needs.

The advantage of such *Feature-Oriented* systems is that they separate the principal behavior from other optional extended ones. This distinction facilitates the addition of new capabilities without having to redesign or re-implement the whole system. However, the combination of features can lead to conflicting or contradictory behaviors. This problem, the so-called *feature interaction* problem [3, 4, 1], is characterized by the fact that some features can be combined together in such a way that they result in non-desired behaviors.

Some interactions are desirable and planned. Others are to be avoided, and most of these can be avoided at the specification and validation stages. Specification and validation are two major phases taking place in the design of complex, and usually distributed, architectures such as telephony systems. The design usually starts in the early stages with the informal requirements followed by formal specifications [5]. Once the system is formally specified, a validation can be performed in order to verify, as far as possible, that the specification is free of errors. Such a validation also insures that the specification of the system fulfills the requirements.

1.2 Motivation

Various Formal Description Techniques [6, 7, 8] address the verification and validation of systems. These techniques can be directly applied to the feature interaction problem but their efficiency can be sometimes limited. Looking for possible feature interactions at the validation level often implies checking each and every possible combination of features, in all possible execution sequences. Even if only considering pair-wise combinations, the number of combinations to analyze can be considerably large. Hence, trying to find feature interactions

in a detailed specification of a system is not always a viable option.

A help to find interactions is to identify, at the requirements level, combinations presenting specific *incoherences* that can lead to potential interactions in the specification.

In the reminder of this thesis, the concept of incoherence will play a crucial role. This concept will be defined formally in chapter 4, section 4.2. For the time being, we provide the following informal definition: an incoherence refers to some incoherent behavior of features with respect to what would be coherent from the point of view of the intended system's behavior.

An example of such an incoherence is the fact that two features, let us say `Incoming Call Screening` and `Call Forward on Busy`, can be triggered by the same event and present different results (see chapter 4, section 4.2.2). Another example is the fact that a feature such as `Call Forward on Busy`, results in triggering another feature such as `Outgoing Call Screening` while their results present a contradiction (see chapter 4, section 4.2.3).

Features can be formally modeled using an appropriate formal representation and information provided by the requirements. The use of a formal representation allows the application of formal filtering rules to identify incoherences between features.

Then, a scenario based validation can be performed using automatic derivation of test suites: for each incoherence identified, a validation test suite can be built and applied to the specification to check whether or not the incoherence leads to an interaction. Using such a method allows to focus on the potential interactions and hence to reduce the testing phase by restricting tests referring to combinations that do present incoherences.

1.3 The Proposed Approach

Based on the ideas presented above, we propose an approach targeting the identification of incoherences corresponding to potential interactions and the validation of a specification with respect to such incoherences. Our work does not address run-time resolution of feature interactions [3, 4, 9, 10]. The identification of incoherences and detection of feature interactions is done statically during the specification phase of the system. The proposed approach is mainly composed of a filtering part and a validation part.

1.3.1 Filtering

Starting from requirements, the filtering part consists in identifying incoherences between feature specifications. Based on the information provided by requirements, features are modeled using a formal representation. The features modeled are analyzed by a tool that implements formal rules for the identification of incoherences between pairs of features. From the requirements and knowledge obtained from this filtering, a formal specification is derived. The knowledge provided by filtering allows to derive a better specification by making the designers aware of the incoherences present in requirements.

1.3.2 Validation

Based on the knowledge obtained from the filtering, the validation part consists in checking whether or not the incoherences previously identified lead to interactions in the specification.

The correspondence between the representation of features for the filtering and their representation in the specification is done via *mapping rules*. These mapping rules are used to translate the incoherences identified by the filtering tool into test suites. Such test suites are applied against the specification to perform scenario based validation, targeting the detection of feature interactions corresponding to the incoherences identified.

1.4 Organization of the Thesis

The thesis consists of an introduction, five main chapters and a conclusion. Each chapter is presented below with a brief summary of its contents.

Chapter 1 – Introduction

Chapter 1 contains the introduction of the thesis.

Chapter 2 – Framework

Chapter 2 describes the conceptual framework on which this thesis is based. It introduces the reader to the so-called feature interaction problem and some of the existing approaches and their limitations. The structure of our research project is discussed, along with our filtering and validation approach.

Chapter 3 – Use Case Maps, LOTOS & Mitel's System

Chapter 3 presents the specification part of our project. It introduces the reader to Use Case Maps, their purpose and notation and the telephony system (provided by Mitel Corp. as a set of UCM) on which this work is based. In addition, the LOTOS language and the translation principles of a UCM model into LOTOS are presented.

Chapter 4 – Identification of Incoherences – Theory

Chapter 4 presents the proposed filtering process. This process is based on a formal representation of features and on formal *incoherence rules* targeting the identification of incoherences corresponding to potential interactions between features. The basic principles of our feature representation, formalism and incoherence rules are presented.

Chapter 5 – Identification of Incoherences & FI Detection – Tool

Chapter 5 presents the scenario based validation of our LOTOS specification with respect to features and feature interactions testing. A tool for automatic filtering of incoherences and automatic derivation of validation test suites is presented.

Chapter 6 – Application & Evaluation

Chapter 6 presents the application of our approach on two different specifications. In addition, we present a performance evaluation of our tool and a comparison with a similar technique.

Chapter 7 – Conclusion & Future Work

Chapter 7 concludes the thesis. It reviews the achievements and contributions and presents the possible improvements of the approach along with further research.

Chapter 2

Framework

This chapter introduces the reader to the feature interaction problem and to some of the current approaches used to deal with it. It also describes the project to which our work is related and our initial approach, and refinements, for the detection and validation of feature interactions. The approach is developed and explained in detail along the thesis.

2.1 The Feature Interaction Problem

The feature interaction problem is often related to telephony system features. However, a feature interaction can occur in any complex system and the definition of such interactions need to be clearly stated. This section presents our definition of a feature as well as what we consider to be a feature interaction with respect to telephony systems. Having introduced the problem, we give reasons of its prominence in such systems and we review some existing approaches. Then, we propose an approach for the detection and validation of feature interactions in the early stages of development of a system.

2.1.1 Definition of a Feature

The concept of feature in telephony is best described pragmatically with respect to user requirements. A feature is a functionality offered by a system and has the purpose of fulfilling certain user intentions in the context of a call. A feature is seen as an optional unit or increment of functionality that is added to a system [1]. Such a system is seen as a base implementation around which features are *orbiting*. Let us call $\mathcal{B}$ the base of the system and $\mathcal{F}_x$ a feature where x identifies the name of the feature, and the composition operator $\oplus$. The composition of a feature $\mathcal{F}_x$ with the base system $\mathcal{B}$ is then denoted: $\mathcal{B} \oplus \mathcal{F}_x$. Hence, a system decomposed into a base $\mathcal{B}$ and n features $\mathcal{F}_x$ where $1 < x < n$ is denoted: $\mathcal{B} \oplus \mathcal{F}_1 \oplus \mathcal{F}_2 \oplus ... \oplus \mathcal{F}_n$ [11, 12].

This way of modeling separates the behavior of the base of the system from the behavior of the features. As well, it separates single feature behaviors from composed feature behaviors. We believe that one advantage of such modeling is that it makes the specification usually easier to extend, modify, and test, since these activities can be done incrementally. On the other hand, we observe that adding features makes the composition more complex and leads to the so called feature interaction problem.

A feature composed with a base system influences the behavior of the base system. The system $\mathcal{B} \oplus \mathcal{F}_x$ is different from the system $\mathcal{B}$. We then say that the feature $\mathcal{F}_x$ is interacting with the base system $\mathcal{B}$. Several features can be composed together with the system. In that case, features tend to influence the behavior of the system as well as the behavior of each other. We say that the features interact with the system and with each other. Since features interact together, we call this a feature interaction. Section 2.1.2 explains in detail the notion of feature interaction and the problems to which such interactions often lead.

2.1.2 Definition of a Feature Interaction

An interaction between features can be either a *cooperative interaction* or an *interfering interaction.* A *cooperative interaction* occurs when features interact together without causing undesired behaviors or other problems to the system (e.g. deadlocks). If features and their composition are well defined, only *cooperative* interactions should occur. An *interfering interaction* is the contrary of a *cooperative interaction*: it causes incompleteness and inconsistencies in the specification, possibly making it unimplementable. When occurring at run-time, the *interfering* interaction usually harms the system and results in undesired behavior from the user's point of view.

The term *feature interaction* usually refers to an *interfering feature interaction.* In order to simplify the reading, either the acronym *FI* or the expressions *feature interaction* or simply *interaction* are used thereafter to refer to an *interfering feature interaction.*

In a system, different feature interactions can occur for different reasons. We identify two main categories of feature interactions:

Direct interaction – There may be a direct interaction if two or more features are triggered at the same time, by the same event and lead to different, or contradictory results. In this case, the system may behave inconsistently, with non determinism or contradiction (which is a special case of non determinism) in the results.

For example, a user X who subscribes to Call Waiting is notified if someone calls him when he is busy. Suppose that X also subscribes to Voice Mail, a feature that is triggered by an incoming call when he is busy or does not answer after a certain time. If not studied and designed carefully, the composition of these two features can lead to a behavior not desired by the user. For instance, when busy, Voice Mail can be triggered as soon as there is an incoming call. Voice Mail then overrides Call Waiting, which might not be the behavior intended by the user.

Transitive interaction – The transitivity is denoted by the fact that the results of one of the features involved in the composition lead to the triggering of another feature also involved in the composition. An interaction may occur if such transitivity leads to contradictory results between the features or to a loop.

For example, let us suppose that user X subscribes to Call Forward Always and forwards all incoming calls to user Y. Suppose that user Y subscribes to Incoming Call Screening and blocks all incoming calls from user Z. If Z calls X, the Call Forward Always feature forwards Z to Y and expects a connection. This triggers the Incoming Call Screening feature of Y that expects to block the connection. Results are contradictory.

All feature interactions that can emerge from feature compositions are specific to the system specification. Hence, a particular feature interaction that occurs in a given system may not occur in another one. In addition, even if considering the same requirements, different results can be obtained depending on the level of abstraction used to model the features.

As mentioned, an interaction between features can be a *cooperative* or an *interfering* one. Accordingly to our definition of a feature interaction, only interfering interactions are called feature interactions. However, it is important to keep in mind that the difference between *interfering* and *cooperative* interactions can be fuzzy. The borderline is always subjective and depends on the designer or tester interpretation, except if the interaction clearly violates behavioral principles. Given a particular composition of features, the result can be interpreted differently depending on the definition of feature interaction that is being considered.

An example is the following: let us consider that some user X has Call Forward Always to user Y. User Y also subscribes to Call Forward Always but forwards all incoming calls to user Z. Let us imagine that user A calls user X. X forwards him to Y, and, Y forwards him to Z. Then, A ends up ringing Z, while X intended A to ring Y. This can be seen as a *cooperative* feature interaction, since a forward chain is created and the user is *cooperatively* forwarded through it. On the other hand, this can be seen as an *interfering* interaction since the Call Forward feature of Y interferes with the Call Forward feature of X and the final result is different from the the one intended by user X.

2.1.3 Prominence in Telecommunication Systems

Telecommunication systems are all facing feature interactions due to many factors [1]:

- The development of new features is usually stimulated by new technology. New technology often eliminates old obstacles and, by doing so, invalidates deeply established assumptions.

- Most telecommunication devices are designed to be as simple and as standardized as possible. In order to be applicable to these devices, the features developed must also be standardized, sharing a common vocabulary and signal, which is not always the case.

- Telecommunication systems over the world have to inter-operate. Unfortunately, the interfaces allowing them to inter-operate do not always follow standards. This often causes problems when a call is processed through different systems.

- Features in telecommunication systems tend do be exceptions with respect to each other. A certain feature can implement a general behavior while another implements a particular case of this general behavior. Features such as Call Forward Always (general case) and Call Forward on Busy (exception case) illustrate this case.

- The fact that telecommunication systems are aging but are still growing quickly with the addition of hundreds of features tends to be a risk. The more features a system possesses, the more interactions can occur.

- Telecommunication systems have a long tradition of developing projects, and even marketing, by features. Such an organization inevitably discourages thinking across feature boundaries.

- Features are often developed by different teams, especially in case of inter-operating systems such as systems from different vendors or countries. This leads to the fact that features have to deal with other unknown features that cannot always be trusted.

2.2 Existing Approaches

Detecting feature interactions in a product implies to build test cases and to apply them against the product. Removing interactions found in a final product implies internal code modifications, which is usually a high cost task [13].

Avoiding interactions at the design stage of a system does not remove the necessity of testing the final product but considerably reduces the number of tests to build since most of the interactions are already removed or avoided [3, 4]. Hence, approaches that focus on static avoidance and detection of feature interactions at the design level are generally a cost effective option. The approach we present in this thesis follows these principles.

A significant number of approaches, involving related techniques and tools, have been proposed to deal with the feature interaction problem. These approaches have their strengths and limitations and are usually complementary.

We believe that combining them is often better than trying to choose the best one. In this section, we give an overview of different approaches related to our own approach and briefly discuss their strengths and limitations.

Note that this thesis does not address the problem of feature interaction resolution. This is the problem of removing interactions after they are detected. It is a separate problem on which there is abundant literature [3, 4, 9, 10].

2.2.1 Avoidance & Detection Using Patterns

A pattern language is a set of patterns that are used together to solve a problem. This is a method that has been introduced recently. The use of a pattern language provides solutions for the design of features as well as for the resolution or handling of feature interactions.

G. Utas proposes such an approach [14] . This approach uses a pattern language for call processing and for feature interactions. The patterns developed define generic methods to deal with the development of a call model and with interactions (cooperative or interfering) that occur between features.

This approach allows the designer to model the basic system, the features and the interactions between features in a clean way, and thus, to avoid many of feature interactions. However, detection of interactions is not automated and depends on the designer.

2.2.2 Avoidance & Detection Using Use Case Maps

Another possible approach is to use a graphical notation to help designers to get a global view of the control flow of a system. Having a global and simple view of the system and

features allows the designer to have a better understanding of the model, and thus, to detect possible feature interactions faster.

Use Case Maps [15, 16] are a scenario based notation. Their purpose is to allow designers to describe the behavior of complex systems in a high-level fashion. Using this notation allows to express requirements using standard constructs.

The notation is intended to be useful for requirements specification, design, testing, maintenance, adaptation, and evolution [17]. Use Case Maps have been used in a number of areas such as requirements engineering and design [18], validation [19], detection and avoidance of undesirable feature interactions [20, 9].

The analysis of the concepts expressed by a UCM system is manual. The tester must visually follow the paths to detect inconsistencies. As for patterns, detection of interactions is not automated and depends on the designer. The possible translation of a UCM model into a formal specification [21] enables the use of an executable model on which automatic validation can be performed.

2.2.3 Detection Using LOTOS

LOTOS (Language of Temporal Ordering Specifications) is a Formal Description Technique (FDT) that has been developed within ISO [6, 8]. It is based on process algebraic methods. Tools allowing the *execution* of a specification allow the application of test scenarios (scenario based validation). Such tools allow the designer to use various techniques to detect feature interactions that can be present in the specification considered.

Goal Oriented Execution is a technique for feature interaction detection that is based on the use of LOTOS tools. LOTOS semantic is defined in terms of axioms and inference rules. Given a behavior, these rules are used to generate the next possible actions as well as the next behavior following the execution of one possible action. A sequence of actions is seen as a trace of actions. *Goal Oriented Execution* uses a type of inference rule that generates traces of actions leading to a preselected action in the specification: the goal.

Applied to the feature interaction problem [22], this technique consists in building a LOTOS specification of the system in consideration together with its features and their properties. Then, identifying goals that correspond to the violation of such properties allows to apply the inference rules of the *Goal Oriented Execution* to check whether such goals can be reached in the specification. If such goals are reachable, a feature interaction exists.

This method however presents some limitations. State explosion [23] can occur due to the number of possible paths required to reach an action. Some actions can possibly remain unreachable. Reaching some goals may imply the use of intermediary goals. In addition, even if all goals are reachable, detecting all feature interactions implies to identify all such sequences of actions corresponding to them and it is not possible to insure that all such sequences have been identified.

2.2.4 Detection Using Permutation Symmetry

Many approaches are based on Finite State Machine (FSM) [24, 25] for the description and specification of features. The methods used for detecting feature interactions in these approaches are often based on exhaustive searches. Such methods are complete and reliable

in principle but since they explore the whole space of states and transitions of the system, they usually need a huge computational time and often lead to the so called state explosion problem, even when considering small specifications [26].

Permutation symmetry [26] proposes to reduce the state space by using a state space reduction technique based on the identification and removal of symmetric relations in the specification. This reduction technique keeps the essential relations allowing the detection of the same interactions (in the same number) as in the complete specification with a gain of time.

This method allows a reduction of the number of states and transitions. However, in its current formulation, this method is only applicable to FSM or Petri Nets based specifications and the reduction of the number of states and transitions is not sufficient to eliminate the state explosion problem.

2.2.5 Detection Using Temporal Logic

Temporal logic [27] is a language that provides a set of special operators used to express formal properties in a natural way. This language defines predicates over infinite sequences of states, allowing, for instance, to express the fact that a sequence is *always true*, *sometimes false*, *always false*, etc. Hence, each formula of temporal logic is (in general) satisfied by some sequences of states and falsified by some others.

Such a language can be applied to the feature interaction problem. In paper [28], A.P. Felty and K.S. Namjoshi present an approach that consists in specifying features using temporal logic. They show that inconsistencies between features can be automatically identified using existing model checking tools [29].

After experiments on a set of feature specifications derived from Telcordia standards, the approach revealed itself as being *reasonably efficient and quite accurate* [28]. Most of the interactions given in Telcordia standards were detected, as well as new ones. However, as mentioned by the authors, the logics used in this work are limited to next-state descriptions, so no liveness properties can be expressed.

2.2.6 Static Detection Method Without State Enumeration

T. Ohta and T. Yoneda developed an approach based on a state transition representation of features and knowledge *elicitation* [30, 31]. Features are represented using a formal language. The authors developed an algorithm that *elicits* knowledge from the services specification. This knowledge is used in conjunction with rules to detect feature interactions.

This approach allows to avoid problems such as state explosion, high computational time and low coverage. However, the feature interaction detection remains manual. This approach is explained in greater detail in chapter 6, section 6.4.

2.2.7 Related Problems

In spite of the detection methods presented, some problems remain:

- Methods presented do not identify all interactions. Even with a strong and precise language to represent features and with tools to automate the detection of possible feature interactions at the design stage, many interactions can remain undetected.

- Adding new features to a system implies testing their interactions with the existing ones. It also implies considering more possible paths and states. This can lead to state explosion problems, making the integration testing difficult, if not impossible.

Today's environment in telecommunications is competitive. The number of features increases very quickly, leading to an important number of potential interactions: detecting feature interactions is necessary, but also time consuming. The project on which our work is based gave us the opportunity to propose a method that allows the detection of many types of feature interactions in a reasonable time.

2.3 Structure of the Project

Our work is related to a collaborative Research and Development project involving the University of Ottawa *(Telecommunications Software Engineering Research Group, School of Information Technology and Engineering)*, Mitel Corp. and Communication and Information Technology Ontario (CITO). The project has a wide scope and involves the use of many

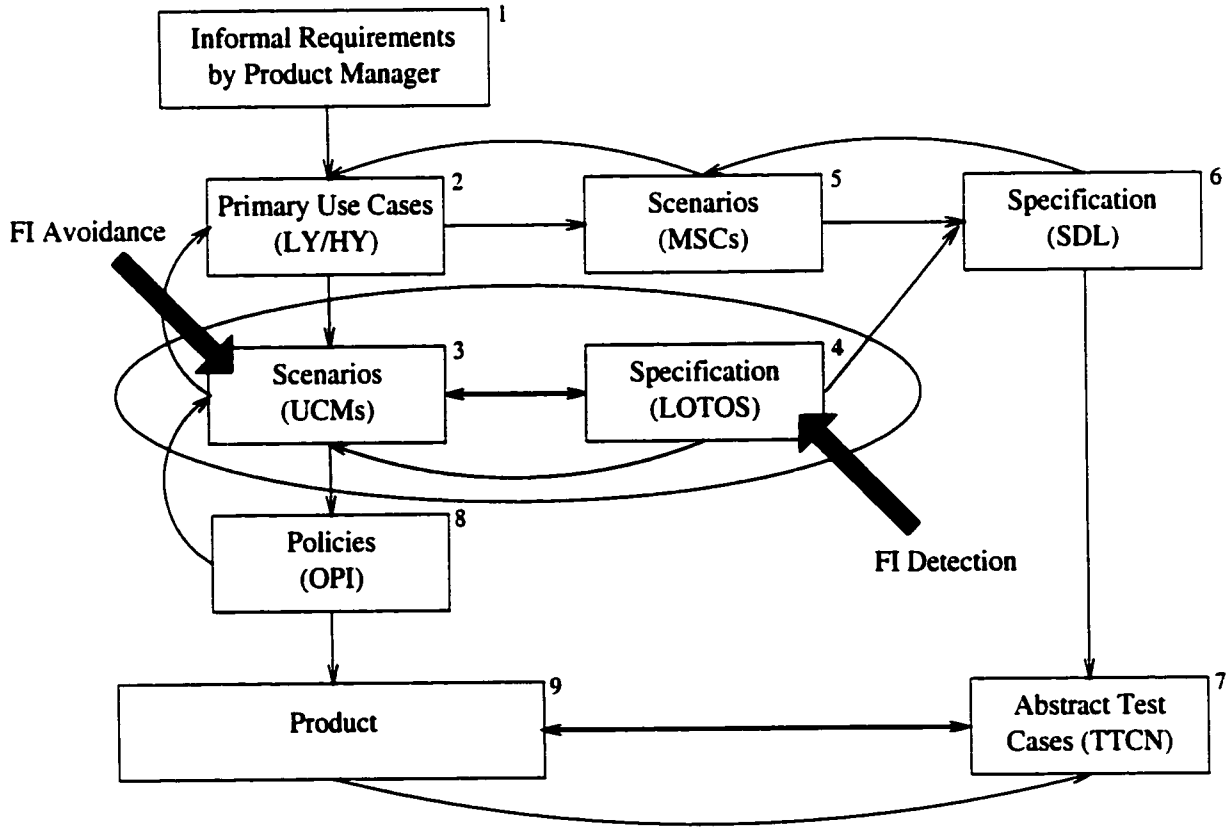


Figure 2.1: FI Detection in a Formal Product Development Model

techniques, some (e.g. UCM, LOTOS) discussed in this thesis, and others (e.g. SDL, OPI, TTCN) not discussed here. In order to give a general description of the project, we are obliged to mention below a number of concepts on which we do not elaborate because they are not used in the rest of the thesis. The reader who is not familiar with these concepts can ignore this section without harm.

Mitel Corp. provided us with a set of Use Case Maps (about 110 maps) representing an agent-based telephony system with 9 features. This system and the features it contains are presented in chapter 3, section 3.2. The aim of the project is to establish a Software Engineering process for the fast specification and testing of such a system.

As illustrated in fig. 2.1, the method starts with informal requirements (1) from which primary use cases (2) are derived. The project splits in two iterative processes:

- One (5) consists in building scenarios using Message Sequence Charts (MSC) [32]. These scenarios are used to produce and validate the SDL (6) [7, 8] specification. The SDL specification allows the derivation of Abstract Test Cases (TTCN) (7) [33] used to test the final product (9).

- The other one consists in building scenarios using Use Case Maps (3). These scenarios are used to provide OPI policies [34] (8), used in the implementation of the final product. In parallel, they are used to produce a LOTOS specification (4) which provides an executable semantic to UCM. The LOTOS specification is validated using scenarios obtained from UCM, providing testing information for the SDL model.

2.3.1 Initial Approach

As mentioned in section 2.2, our methodology focuses on the static detection and validation of feature interactions in the early stages of development of a system. Our initial approach is part of boxes (3) and (4) in fig. 2.1 and is composed of the following steps:

Use Case Maps Model

The methodology starts by employing Use Case Maps (UCM) to build a general description of the system (3). As explained in section 2.2.2, UCM are a scenario based notation. Their purpose is to allow designers to describe the behavior of complex systems in a high-level fashion. Using this notation allows to express requirements using standard constructs and provides designers, product managers, and even non-specialists with a graphical representation that offers a global view of a system.

The UCM validation targets *FI avoidance*. This graphical representation enables designers to focus on the *big picture* and, we believe, to avoid and detect design defects at early stages. Because it is high-level, this notation is implementation-independent. Therefore, validated UCM are reusable on different platforms and different products, and at different stages of product development and use.

We see UCM as being efficient as an intermediate representation between requirements and formal specification. The UCM concepts and notation are presented in detail in chapter 3, section 3.1. In order to validate the design and detect possible errors, the UCM

representation of the system is implemented using the Formal Description Technique (FDT) LOTOS and its tools.

LOTOS Specification

Introduced in section 2.2.3, LOTOS (Language of Temporal Ordering Specifications) is a Formal Description Technique (FDT). The use of LOTOS allows to build a formal specification of a system. Verification and validation are performed on this model using LOTOS tools. LOTOS concepts are presented in chapter 3, section 3.3.1.

The derivation of the LOTOS specification is currently done by hand, although a partial automation is being considered. Required knowledge here is a good understanding of the UCM system, the features being considered, expertise with UCM, LOTOS, and the techniques used to translate UCM into LOTOS.

To derive the LOTOS specification, the UCM model is analyzed to determine an appropriate specification structure. During this analysis, the designer makes choices concerning the mapping of the UCM model structure and constructs into LOTOS. Since Use Case Maps are used at a higher level of abstraction than LOTOS, this analysis gives the opportunity of inspecting the documentation and detecting missing parts, contradictions, and incoherences. The translation guidelines are presented in chapter 3, section 3.3.2.

Feature Interaction Detection

The LOTOS validation targets *FI detection* based on scenario-based testing. Scenarios are derived from UCM for validation and verification purposes against the LOTOS specification. Note that these scenarios can be re-used later on for similar activities against an SDL specification of the system, up to the generation of test cases in machine-readable format.

The derivation of scenarios from the UCM model is done by hand. Again, required knowledge here is expertise with UCM, LOTOS, and the features being considered. Once translated into LOTOS, test scenarios are used to validate the LOTOS specification, then to test the system. Scenarios may be categorized according to what they are meant to test: basic system properties, individual features properties or feature interactions.

The testing is performed using LOLA *(LOtos LAboratory, Universidad Politécnica de Madrid)* [35], a free-ware tool. Each test scenario, translated into LOTOS, is synchronized with the LOTOS specification (see chapter 3, section 3.3.1 for explanations on synchronization). LOLA reports whether visiting the composed behavior leads to success for all execution branches (MUST PASS), for only some branches (MAY PASS) or for none of them (REJECT). Scenario based validation is explained in chapter 5, section 5.1.

2.3.2 Refinements

In this approach, the derivation of test scenarios for testing feature interactions is done by hand. Producing scenarios manually is not a long task if the number of features is small. However, a large number of features makes it almost impossible, especially because the number of possible combinations increases considerably with the addition of new features due to the number of new possible combinations it creates.

A first refinement of this process is to automate the identification of possible feature interactions in requirements as well as their detection in a specification. This is the subject of this thesis. Briefly, such an approach requires two steps. The first step takes place after the users' requirements phase. It is based on feature definitions only and consists in filtering the incoherences corresponding to potential interactions. The next step takes place after the specification phase and consists of the automatic derivation of validation test suites for each and every incoherence identified in the first step. These test suites are composed of scenarios derived using information relative to the structure and syntax of the specification. They are applied to the specification to detect feature interactions.

This refinement increases the strength of the current approach. It allows the avoidance and detection of feature interactions on the basis of a simple and fast description of the features. Moreover, the automatic derivation of test suites for the validation of the specification considerably reduces the time of the testing process. This process is explained in section 2.4.

2.4 The Proposed Filtering and Validation Process

We should, at this point, provide a general overview of the filtering of incoherences and feature interactions detection process. We provide a few informal definitions in order to enable a first understanding of what follows.

Filtering considers the feature definitions given in requirements and analyzes these definitions in order to identify the presence of incoherences that can lead to potential interactions in the specification. The identification of such incoherences is seen as a filter since a number of potential interactions can be identified at the requirements level. These incoherences and the rules used for their identification are presented in chapter 4.

In this thesis, the word *incoherence* refers to incoherences that are present between features at the requirements level and that correspond to potential interactions possibly present in the specification. The words *interaction* and *feature interaction* refer to a feature interaction present (or not) in the specification.

As illustrated in fig. 2.2, the process starts with requirements and splits into two paths. The paths are in parallel but they depend on each other. The one on the left concerns the identification of incoherences and derivation of test suites to detect feature interactions in the specification. The one on the right illustrates the process of deriving of the Use Case Maps model, later translated into a LOTOS specification. The whole process is composed of the following steps:

Incoherences Filtering (1)

Even if informal, requirements can provide enough information for the identification of incoherences between two features.

The information about features is extracted from the requirements and mapped into a Prolog [36, 37] representation (see chapter 5, section 5.2 for explanations about Prolog). The descriptions of features are analyzed by our tool which combines them by pairs and identifies the incoherences they present. As previously mentioned in section 2.4, these incoherences, automatically identified, reveal interactions that can possibly occur in the specification.

At this time in the design process, the complete detailed specification of the features is not known, and thus, the analysis of incoherences is only based on the information that is extracted from the requirements. Due to this, it is not possible to assert that the incoherences identified lead to feature interactions in the specification, nor that all absence of incoherences implies the absence of feature interactions.

Even if not absolute, these results provide useful information regarding the features that, once combined together, can lead to interactions. This information can be used by designers

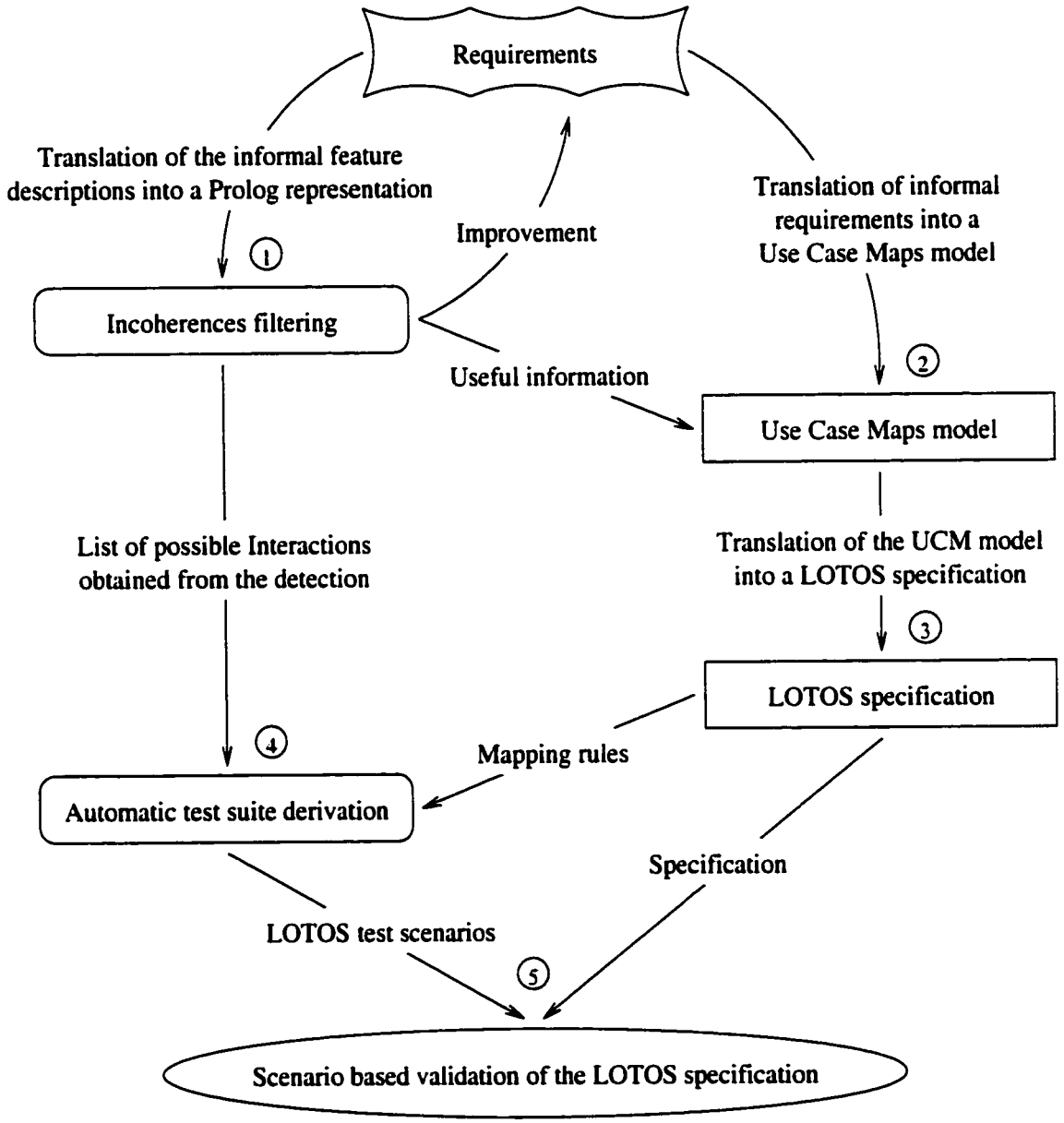


Figure 2.2: Feature Interaction Validation Process

to refine requirements and to avoid the presence of such interactions in the specification. The filtering theory and its automation are respectively presented in chapter 4 and chapter 5.

Use Case Maps Model (2)

Requirements are translated in a UCM model in the same way it is presented in section 2.3.1, and explained in detail in chapter 3, section 3.1. It is built in accordance with the knowledge about incoherences obtained from the filtering process. Thus, the identification of incoherences helps to avoid their presence in the UCM model.

LOTOS Specification (3)

The LOTOS specification is done by manually translating the Use Case Maps into LOTOS, as it is presented in section 2.3.1 and explained in detail in chapter 3, section 3.3.2.

Automatic Test Suites Derivation (4)

Since the LOTOS model is executable, test scenarios can be run against it. Deriving test scenarios corresponding to the incoherences identified allows to verify whether or not the interactions they correspond to exist in the specification.

The structure and syntax of scenarios rely on the structure and syntax of the specification to which they are applied. The incoherences are represented as scenarios containing actions. These actions are translated into LOTOS test scenarios via mapping rules. These rules are built with respect to the structure and syntax of the LOTOS specification. Their purpose is to provide a correspondence between the Prolog representation of a feature and its implementation in the LOTOS model.

Using LOTOS testing tools, the test suites are applied against the specification to check whether or not interactions are present. This process is not strictly dependent on the use of LOTOS; other specification languages could be used. The automatic derivation of test suites is presented in chapter 5, section 5.3.

Scenario Based Validation of the LOTOS Specification (5)

The testing of feature interactions in the LOTOS specification is done by applying scenarios against the specification using LOLA, as previously presented in section 2.3.1 and explained in detail in chapter 5, section 5.1.

2.5 In Summary

We have introduced the framework on which our thesis is based. We have presented our definitions of a feature and a feature interaction and enumerated some of the current approaches used to deal with feature interactions. We have proposed an approach combining a known approach and a filtering process for the identification of incoherences and detection of feature interactions in a model. This approach is presented in detail in the rest of the thesis.

Chapter 3

Use Case Maps, LOTOS & Mitel's System

This chapter introduces two techniques used in the project, UCM and LOTOS, and introduces the system designed using these two techniques. We present Use Case Maps, their purpose and the way they can be used to specify systems. We present the system provided by Mitel Corp. as a set of UCM and explain how the translation of this model into a LOTOS specification is done and which considerations and decisions must be made.

3.1 Presentation of Use Case Maps

The task of designing and describing complex systems can be very laborious, even for experienced analysts. It is a good option to describe such systems with a high level of abstraction and in an efficient way such that they are fast to design and to understand.

As mentioned in chapter 2, section 2.2.2, Use Case Maps (UCM) are an adequate technique for representing complex systems. They are an informal design notation; their semantic is not precisely specified: it is intended to permit different interpretations. UCM allow the reader to have a global idea of the system at a glance. In addition, a tool, the *Use Case Maps Navigator* [38, 17], assists the designers in the design and the syntactic correctness of the UCM model they develop.

This section gives an overview of the UCM with respect to the elements of the notation used in the design of the Mitel system. A few additional elements are presented in order to give to the reader a general understanding. More complete dictionaries of the notation can be found in [15, 17, 16].

3.1.1 Basic Notation & Interpretation

The basic notation is composed of components, scenario paths and responsibilities. Fig 3.1 shows an example of a simple system represented using UCM.

Components – They are used to specify the structure of the system. A system can contain several components and a component can contain several sub-components. Different shapes can be used to represent several kinds of objects depending on what the designer wants to express.

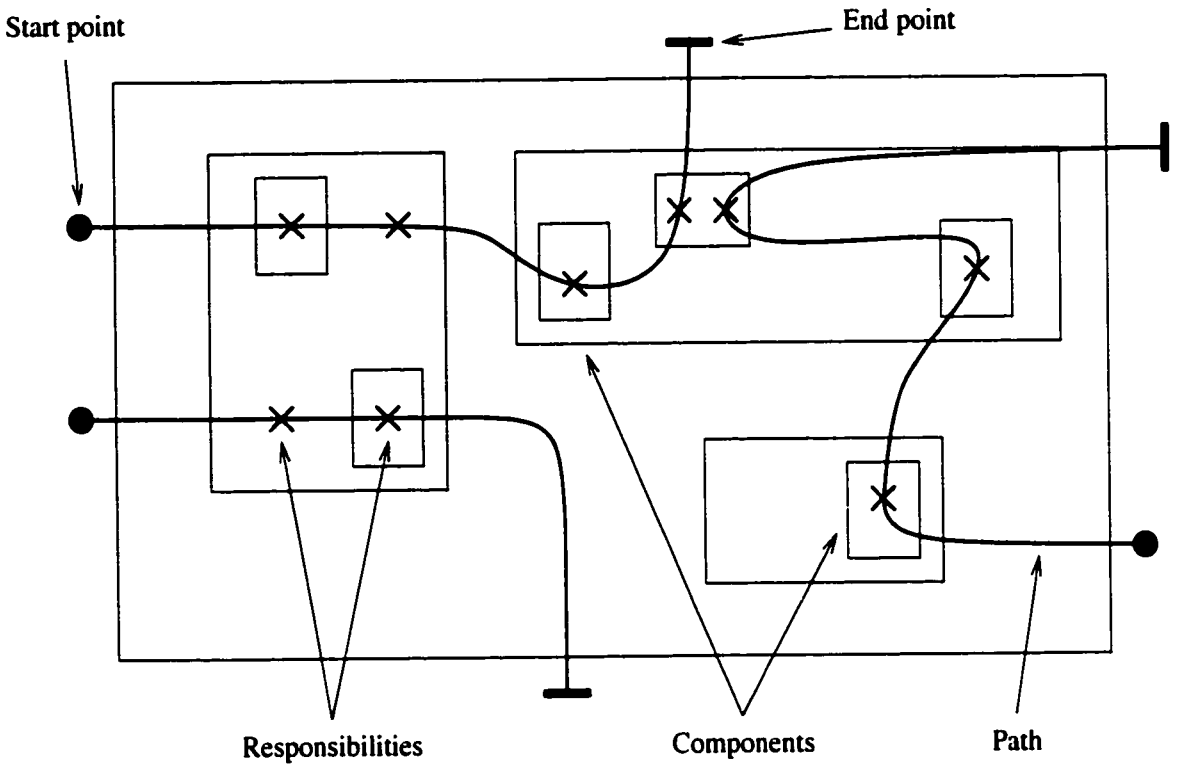


Figure 3.1: Use Case Maps Basic Notation

Scenario paths - Also called routes, they are represented by curved lines and are used to express the behavior of the scenario along time. A path starts with a start point, represented as a bullet, and terminates with an end point, represented as a bar. Paths navigate across the components to mark the places where responsibilities take place.

Responsibility points - Responsibilities are represented by crosses and are situated on paths. A responsibility is used to represent some action that takes place during the scenario. It can represent different types of actions depending on the system designed. For instance, in case of a telephony system, a responsibility can be used to represent an off-hook event.

3.1.2 Scenario Paths Notation

The scenario paths notation contains all that is needed to express scenarios using paths. This include interactions between paths, shared routes and concurrent routes. This section explains the elements of the notation. In addition, we show some variations that can be obtained by combining elements of the notation together.

Shared Routes

Shared routes represent segments of paths that can be shared by two paths. They can be used to express the fact that scenarios have common behavior parts. They are also used to express the possibility of a choice between two routes. Fig. 3.2 shows the operators used.

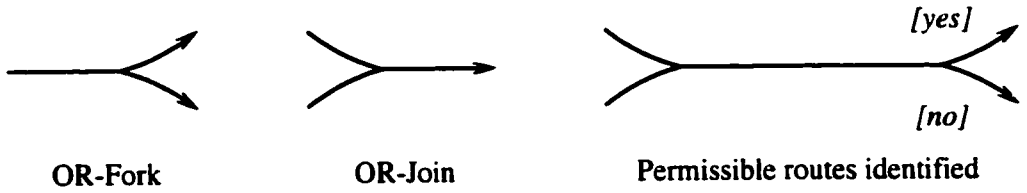


Figure 3.2: Shared Routes

OR-Fork – The OR-Fork expresses a choice. It represents the exclusive OR operation on a path divided into two possible routes. Only one of the two possible routes is followed.

OR-Join – The OR-Join expresses the fact that two different possible paths of a scenario join and share the same segment of path. Note that this does not express parallelism. In that case, only one of the two paths is active. The fact that the paths share the same segment means that, whatever path is active, the same actions take place.

Permissible routes identified – To indicate which path must be followed, alternatives may be identified by labels or by conditions associated to the corresponding sub-path.

Concurrent Routes

Concurrent routes are used to express parallelism. They allow someone to indicate a point where a scenario splits into two (or more) different branches that continue their execution independently. They also allow someone to indicate a point where branches need to merge. Fig. 3.3 presents the operators used.

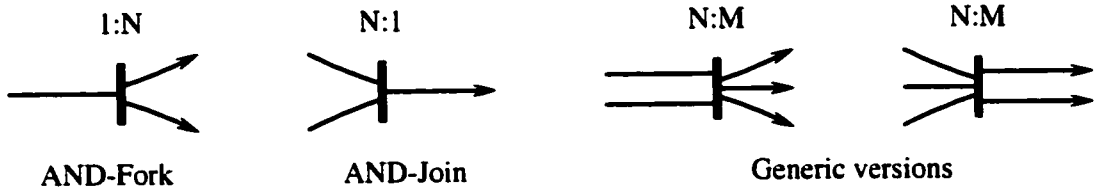


Figure 3.3: Concurrent Routes

AND-Fork – The AND-Fork splits one path into two paths that proceed in parallel. Paths are fully independent from each other. Thus, this operator expresses parallelism and can be used to model concurrent scenarios.

AND-Join – AND-Join indicates that two paths join each other to continue together (as being the same one). This operator can be used to model concurrent scenarios that need to do common actions at the same time.

Generic versions – As shown by fig. 3.3, AND-Fork relationships are usually 1 to N and AND-Joins relationships are usually N to 1. However, it is possible to use a generic version of them to express the fact that N branches are splitting into M ones, or to express the fact that N branches are joining to become M ones.

Variations on AND-Forks/Joins

Composing the Fork and Join operators is possible and can be used to express temporary concurrency, synchronization or rendezvous, as illustrated by fig. 3.4.

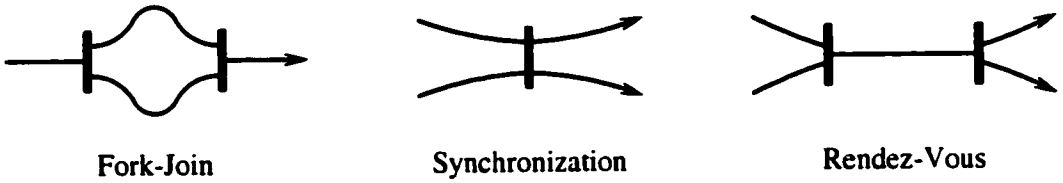


Figure 3.4: Variations on AND-Forks/Joins

Fork-Join – Fork-Join is a composed operator using an AND-Fork followed by an AND-Join. It expresses the fact that a path splits into two parallel sub-paths that join later on and can be used to express temporary parallel behaviors.

Synchronization – Synchronization is used to denote that, at a certain point, paths must synchronize. Two paths are able to pass the synchronization point only once they have both arrived. A generic version of this operator can be used to make more than two paths synchronize.

Rendezvous – Rendezvous is a composed operator using an AND-Join followed by an AND-Fork. It expresses the fact that two parallel paths are joining and synchronizing along the same one for a time, and then fork again. This can be used to express temporary common behaviors of synchronization of scenarios.

Paths Interactions

Path interactions are used to represent two (or more) routes interacting together. We distinguish two kinds of interactions: synchronous and asynchronous, as illustrated by fig. 3.5.

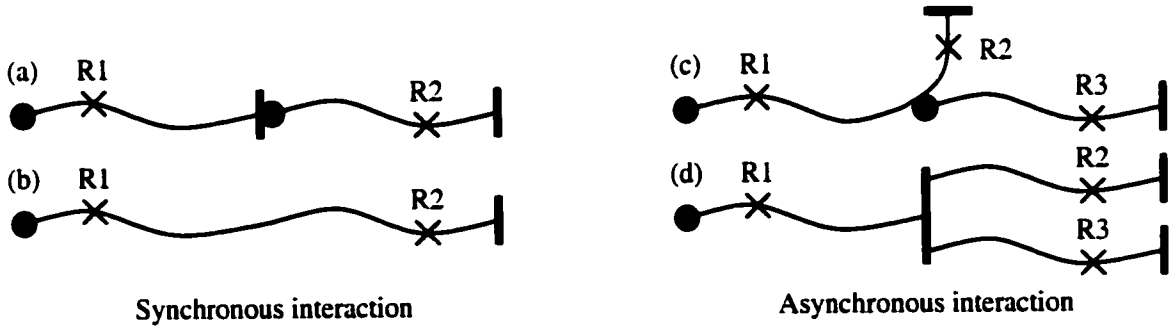


Figure 3.5: Paths Interactions

Synchronous interaction – The synchronous interaction between two paths is represented by fig. 3.5-(a) and shows that reaching the end of the first path triggers the second

one. The effect is similar of one longer path with the constituent segments joined end to end, represented by fig. 3.5-(b).

Asynchronous interaction – The asynchronous interaction, represented by fig. 3.5-(c) represents the fact that, the second path is triggered by the first one and that the first one continues its execution. The effect is similar to one path splitting in two concurrent segments, as shown by fig. 3.5-(d).

3.1.3 Components Notation

Representing components in Use Case Maps can be done using different shapes depending on the object that has to be represented. A component can be static or dynamic. Static components are parts of the system that do not change. Dynamic components are components that can be dynamically created, moved or destroyed. A component can have different attributes. The shape of a component illustrates its attributes. This section introduces the different components and their corresponding representation, as well as the representation of their attributes.

Component Types

Fig. 3.6 illustrates the different shapes that can be used to model static and dynamic components. Each shape corresponds to a particular kind of component.

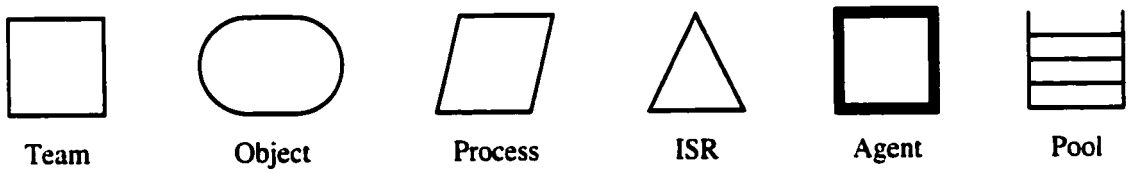


Figure 3.6: Component Types

Team – The team component is a generic container. It may contain components of any types. It is a sort of default component that is generically used in Use Case Maps models as illustrated by fig. 3.1.

Object – Objects represent passive components. They are used to represent parts of the system that are not executing. Such objects act as data and can represent specific components such as databases or more abstract concepts such as a call in a telephony system.

Process – While objects represent passive components, processes represent components that are active. Such components represent executing parts of a system, a computation for instance. Processes also permit the representation of distributed systems and their execution.

ISR – ISR, Interrupt Service Request is used to model the requests of service interruption that occur in the system.

Agent - Agents are used to represent agents of a system.

Pool - Pools are containers for Dynamic Components that are not executing. Since such components are considered as data, a pool can be used to represent data storage.

Component Attributes

Different types of components exist and can have different attributes depending on their functionality. Shapes presented in fig. 3.7 are not bounded to the team component. They are generic and illustrate the possible attributes for any component.

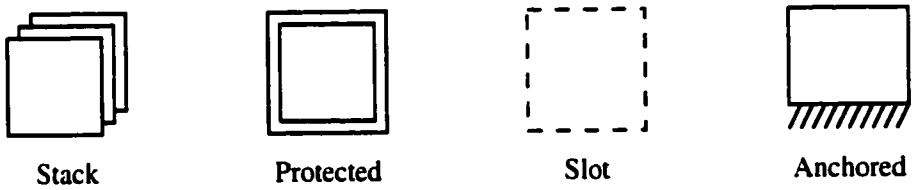


Figure 3.7: Component Attributes

Stack - Stacks are used to represent multiple instances of an object. For example, a stack of agents can represent multiple instances of the same kind of agent. Note that these are not stacks in the sense of first-in-last-out data structures.

Protected - The attribute is used to express mutual exclusion on a component. Using it permits the modeling of mutual exclusion concepts such as semaphores and monitors.

Slot - A slot is a placeholder for dynamic components as operational units. Slots may be populated with different instances of different components at different times.

Anchored - Anchored components are found in plug-ins. A component that is anchored refers to a component that is defined in another map. This can be used to model the fact of exporting components.

3.1.4 Additional Notations

Use Case Maps provide additional notation for extending the modeling possibility: stubs and plug-ins, timers, aborts, grounds and shared responsibilities. These are presented below.

Stubs & Plug-Ins

Stubs can be seen as modules of functions. They are containers for sub-maps (called plug-ins). Using stubs makes it possible to abstract from parts of the system by encapsulating them. Moreover, stubs can be used to decompose a system into different levels, permitting more modularity.

Two kinds of stubs are identified: static stubs and dynamic stubs. Fig 3.8 shows an example of a static and a dynamic stub with their plug-ins.

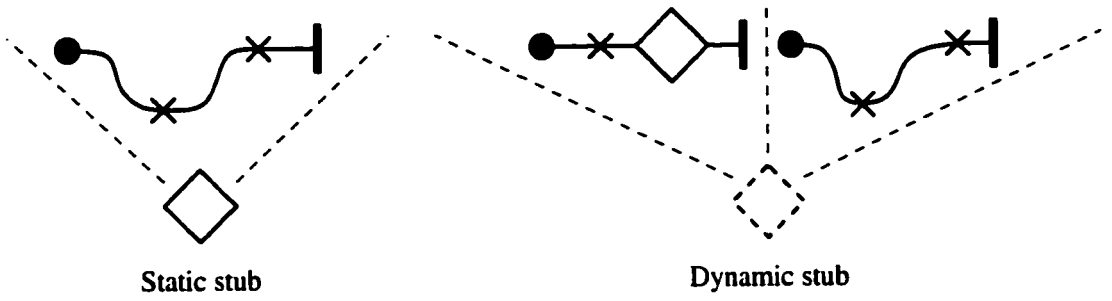


Figure 3.8: Static & Dynamic Stubs

Static Stub - Static Stubs are represented as solid diamonds. They contain one sub-map (plug-in) and are mainly used for the decomposition of system into sub parts.

Dynamic Stub - Dynamic stubs, represented as dashed diamonds, can contain several plug-ins. When entering a dynamic stub, the selection of the plug-in to use can be determined by using selection policies usually described with pre-conditions. It is possible to select several plug-ins at the same time, for sequential or parallel composition, but adequate detailed documentation needs to be provided outside of the UCM.

Timers, Aborts, Grounds & Shared Responsibilities

The elements illustrated in fig. 3.9 are used to indicate possible failures, aborts, timeouts and shared responsibilities.

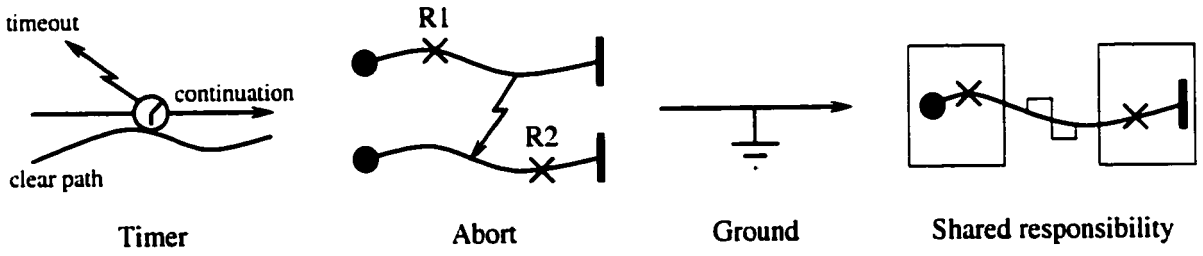


Figure 3.9: Timers, Aborts, Grounds & Shared Responsibilities

Timer - A timer is a special waiting place. It is associated with a continuation, a timeout and a clear path. If the timer is reached before it times out, the continuation path is followed. If the timer is not reached and a timeout occurs, the timeout path is the one followed. In any case, the clear path can be used to reset the timer.

Abort - The abort operator expresses the fact that one path may abort another. Fig. 3.9 shows us an example where the top paths aborts the bottom one after the execution of responsibility $R1$.

Ground - The ground element indicates a potential failure point along a path.

Shared responsibility - Shared responsibility represent a complex activity that involves negotiation between two or more components.

3.2 Mitel System

This section presents an overview of the structure of the system on which our work is based, as well as a description of the features integrated to this system.

3.2.1 Use Case Maps Model

The Mitel system architecture, including its various features, was captured using UCM, for which an example is shown in fig. 3.10. This figure shows part of the Root UCM, limited to the originating DA, CA and LA elements.

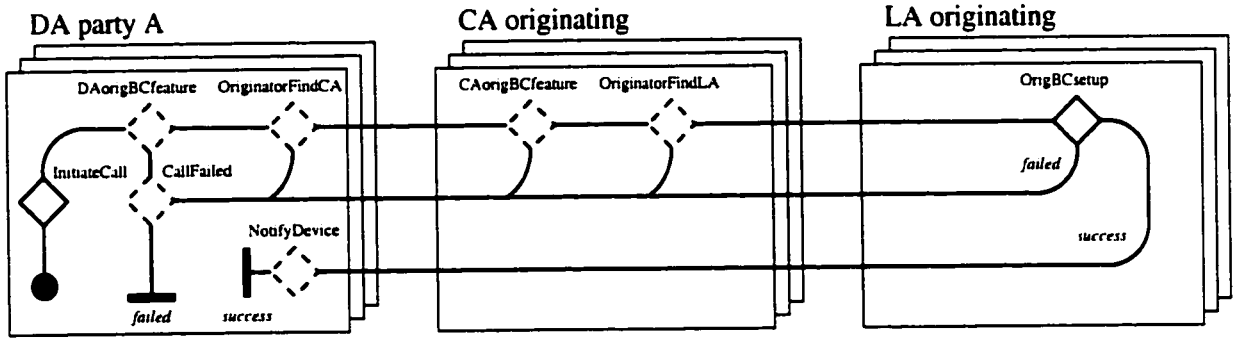


Figure 3.10: Part of the Basic Call Root UCM

Logical Agent (LA) – A LA represents a logical endpoint of the call in a switch. A phone number is associated with a LA. A LA is often associated with the location.

Communicating Agent (CA) – A CA represents a user. It knows about and deals with restrictions and privileges given to a specific user. A CA can be seen as associated to the *role* of a user. A given user can be associated with different CAs.

Device Agent (DA) – A DA represents the physical endpoint of a call. This can be an actual telephone, a computer capable of Voice over IP or some other device. A user can have several DAs depending on the number of communication devices he possesses.

Call Object (CO) – A CO is an intermediary between two LAs participating in a call.

Let us consider a company where employees wear identification badges. These badges broadcast their location to the phone system of the company. The phone system treats incoming phone calls in the following way: as soon as a call comes into the phone system for a particular employee, the system locates the employee and rings the nearest phone.

Let us imagine that a call enters a system. This call enters through a DA which represents the trunk on which the call is coming in. That DA then routes the call to a CA which represents an outside user of the system. The CA involved then chooses a LA built for an outside user. The chosen LA creates the Call Object which can find the LA representing the number being called. The call goes to that LA. The LA finds a CA which represents the user corresponding to the number that was dialed. The CA then locates the user in the building.

Using the information it has about the user's location and the locations of the phones in the building, it chooses a DA. The DA then rings the phone it represents (the nearest available phone).

As mentioned, our case study is based on the Use Case Maps model of an agent system given by Mitel Corp. The Use Case Maps model is composed of 110 different maps distributed in 7 different levels of sub-maps. The UCM were structured in a hierarchical way to describe the 9 features that were considered in this project. Nearly 60 stubs were used along the way, and many plug-in UCM were reused in more than one stub. The responsibilities are found in the plug-ins, sometimes several levels deep in the hierarchy.

Due to the already large number of stubs and complex functionality in the current UCM, and given the reusability of these, we expect to need few additional UCM to capture extra features. This section presents the structure of Use Case Maps model representing the system and its components.

3.2.2 Features Considered

The use of UCM can help directly in the detection and avoidance of many undesirable actions before any prototype is generated. By inspecting specific locations such as dynamic stubs, designers may select appropriate strategies. For instance, a selection policy where preconditions are incomplete or overlapping (non-deterministic) is likely to cause problems. The detection and resolution of many problems can hence be done locally at the stub level, which is, in our experience, far better than doing it at the system level. The correctness is further validated when testing the LOTOS specification. The features considered are described below:

Call Forward Always – Party A calls party B. Party A ends up ringing party C, as a result of party B's request.

Call Forward On Busy – Party A calls party B. Party B is busy. Party A ends up ringing party C, as a result of party B's request.

Auto-Recall – Party A calls party B who is busy. When party B cease being busy, party B is notified of the auto-recall. When party B responds to the notification, a call is made from party B to party A.

Call Waiting – Party A is talking to Party C. Party B calls Party A. Party A receives a notification that party B is on the line. Party A may then switch between talking to party C and party B.

Timed Reminder – Party A sets a timed reminder for a specific time. At the previously set time, party A's phone rings and when party A answers, a message is played.

OutgoingCall Screening – Party A is not permitted to call party B.

Call Hold – Parties A and B are connected. Party B presses a *Hold* button. Party A ends up *on hold*. Party B may retrieve party A at any time.

Call Pickup – This feature allows a party to use his phone and dial a code to answer a call made to another party.

Call Transfer (unsupervised) – Party C is connected to party A. A presses the *Transfer* button, dials B's number, and C ends up ringing B.

3.3 From Use Case Maps to LOTOS

The translation of a UCM model into a LOTOS [19, 21] specification corresponds to the formalization and mapping of an abstract and informal model into a more specific one, while remaining at the design level. In order to do this mapping, the UCM system must be analyzed such that any missing information, confusion or error is detected and corrected.

3.3.1 LOTOS Concepts

LOTOS [6] represents the behavior of a system by using *actions* and *behavior expressions*. Actions represent basic behaviors of the system, for example, *offhook* or *ring*. There is also an internal (invisible) action written as **i**. Three basic behavior expressions are **stop** (also called *deadlock* or *unsuccessful termination*), **exit** (successful termination) and process instantiation P[G](V), where P is the name of a LOTOS process, G is a set of gate parameters (see below for the concept of gates), and V is a set of value parameters. Given behavior expressions B, B1, B2, etc. and actions a, a1, a2, etc., LOTOS operators can be used to construct more complex behavior expressions as it is shown below.

a; B – The *action prefix* operator ; means that the system offers an action a followed by behavior B.

B1 [] B2 – The *choice* operator [] means that the next action offer can be obtained either from B1 or from B2. The other behavior expression is discarded.

B1 || B2 – The *full synchronization* parallel operator || means that a common next action from behavior expressions B1 and B2 has to be found in order for the system to proceed. If such actions exist, they are offered (synchronization) and then the next common action is obtained and so on.

B1 ||| B2 – The *interleaving* operator ||| means that at any point, independent actions from behavior expression B1 or B2 can be offered.

B1 |[a1, a2, ..., an]| B2 – The *selective synchronization* operator |[]| is a generalization of the full synchronization and interleave operators. It means that on actions a1, a2, ...an, B1 and B2 must synchronize; on other actions they interleave.

B1 [> B2 – The *disable* operator [> means that at any time during the execution of B1, B2 can take over, thus terminating B1.

B1 >> B2 – The *enable* operator >> means that, provided that B1 has completed successfully (**exit** behavior), B2 can start.

hide a in B – this internalizes (transforms to **i**) all offerings of actions a in B.

LOTOS includes also a basic Abstract Data Type formalism, called ACT ONE, which is used to represent data abstractions. Data can be associated with actions in two ways: **!**, meaning value offer, and **?**, meaning value query. These can be combined in actions:

```
switch !subscriber ?destination:destination_sort
```

denotes an action on gate *switch* where the current value of the value identifier subscriber is offered, and a value for *destination* is queried simultaneously. Offers and queries are called *experiments*. Note that when no data is involved, actions are simply gate names. With data, actions are gate names with data offers and queries.

A basic concept in LOTOS is the *expansion*. Any LOTOS behavior expression can be rewritten as an equivalent expression containing only choice, action prefix, **stop** and **exit** (although this expression can be infinite). An expanded LOTOS specification represents directly the *labelled transition system* (LTS) of the system in consideration (an LTS is like a Finite-State Machine, having states and transitions labelled by actions denoting synchronization events). Each alternative path in an expanded specification, or each branch in an LTS, explicitly represents one possible sequence of actions in the system. Sequences of actions are called *traces*. Traces must include only visible actions, however invisible actions often are also shown, for completeness.

LOTOS has two main assets in the area of specifying telephony systems and their features: it is capable of representing clearly system structures, and it has a good set of validation tools. Concerning the representation of telephony system structures, LOTOS operators allow us to represent clearly agents that coexist independently (interleave), communicate (selective synchronization), follow each other's actions (enable), or can interrupt each other (disable). Internal system actions can be hidden. Several *specification styles* are possible in LOTOS. Some styles are appropriate for representing abstractly system requirements, while others are appropriate for representing implementation structures. Let us consider the two following processes DA and CA:

```
process DA[ DE_to_CE,  CE_to_DE ]: exit:=
  DE_to_CE; exit
endproc

process CA[ CE_to_DE, DE_to_CE ]: exit:=
  DE_to_CE; exit
endproc
```

These can be combined to require synchronization over all their gates as follows:

```
DA[ DE_to_CE, CE_to_DE ]
  ||
CA[ CE_to_DE, DE_to_CE ]
```

The synchronizing points, or *rendezvous* points, between processes DA and CA are then all actions that involve the gates DE_to_CE and CE_to_DE. If one of the processes wants to

fire an action with gate DE_to_CE or CE_to_DE, the other process must also be willing to do accordingly for the action to become executable. The two processes then synchronize or *get together* to execute this common action. Synchronizing on an action also implies an agreement on the values of the parameters of the action, namely the experiments. It is possible to accept *(query)* a value, represented by **GATE ?valueId:sort**, or to offer *(shriek)* one, represented by **GATE !value**. In all cases involving synchronization, there must be a match in the number of experiments, their order, their sorts and their values. Let us consider two processes P1 and P2 synchronizing on gate COM with an experiment of sort Natural. The four following situations can occur:

1. P1 = **COM !1** and P2 = **COM ?y:Natural**
   P1 offers **1** to P2, which binds it to the value identifier **y** of sort Natural.

2. P1 = **COM !1** and P2 = **COM !1**
   P1 and P2 agree on the value 1.

3. P1 = **COM !1** and P2 = **COM !2**
   P1 and P2 do ***not*** agree on the value and synchronization does ***not*** occur.

4. P1 = **COM ?x:Natural** and P2 = **COM ?y:Natural**
   A value of sort Natural is determined non-deterministically (e.g. provided by the environment) and is bound to both **x** and **y**.

Combined with the interleaving operator (|||), experiments may also be used to provide a more selective kind of synchronization. This subtlety is called *gate splitting*. A typical application is a switch that may wish to synchronize with any phone, but one at a time. The phone processes are interleaved while the switch is ready to synchronize with any of them:

```
switch[ COM ]
  |[ COM ]|
(
  phone[ COM ](p_1) ||| phone[ COM ](p_2) ||| ... ||| phone[ COM ](p_n)
)
```

Then, an experiment is included in an action to uniquely identify each phone such that the switch may selectively synchronize with a specific one. To query a telephone number from the process phone p_1, the switch can specify the action:

```
COM !p_1 !dialTone ?telephone:number
```

Let us consider that a DA must communicate with two different CAs. In a simplistic approach, one can consider a set of gates (one for each direction) for the communication with each of the two CAs. This produces an undesirable static structure: notice that the set of gates of a DA has to be changed whenever a CA is added or removed. On the other hand, *gate splitting* makes it possible to have more dynamic structures. The previous definitions of DA and CA can be extended with gate splitting to a system containing three processes (e.g. two CAs and one DA). CA processes are instantiated with a parameter that predefines their identity. The process definitions become:

```
process DA[ DE_to_CE, CE_to_DE ]: exit:=
  DE_to_CE !cA_a !connect; exit
   []
  DE_to_CE !cA_b !disconnect; exit
endproc

process CA[ CE_to_DE, DE_to_CE ](cA:CAID):exit :=
  DE_to_CE !cA ?message:Data; exit
endproc
```

And the system behavior is the following :

```
DA[ DE_to_CE, CE_to_DE ]
  |[ DE_to_CE, CE_to_DE ]|
(
  CA[ CE_to_DE, DE_to_CE ](cA_a of CAID)
   |||
  CA[ CE_to_DE, DE_to_CE ](cA_b of CAID)
)
```

Each instance of CA shrieks its own identifier (cA_a and cA_b, respectively) and queries a message of sort *Data*, on the same gate DE_to_CE. The DA may either synchronize with CA cA_a through the event DE_to_CE !cA_a !connect or with CA cA_b through the event DE_to_CE !cA_b !disconnect. In all cases, the DA shrieks both the identifier of the CA and the message it wishes to send. For any action to be executed, one CA needs to agree on the first experiment (the CA identifier) and accept the second one (the message).

3.3.2 Analysis & Decisions

During the analysis phase, the LOTOS specifier acquires a global picture of the system. This knowledge is used to take decisions on the specificity of the mapping. Decisions are taken regarding the translation of UCM into LOTOS. They relate to the representation of static and dynamic stubs, plug-ins, data structures, databases, and agents (DA, CA, etc). The system contains the following types of components:

- **Agents** – are DAs, CAs, LAs and Call Objects instantiations.
- **Stubs** – are all the possible static and dynamic stubs used in the agents.
- **Plug-ins** – are all possible plug-ins that can be plugged in the dynamic stubs.
- **Database** – refers to all databases grouped in a single one.

Components are modeled as processes. The database and agents are instantiations at the behavior level of the specification. Stub and plug-in instantiations are at lower levels, inside agent processes. In essence, the architecture of the LOTOS specification reflects that of the UCM model. We have defined guidelines, presented below, to construct LOTOS specifications from UCM.

General Translation Guidelines

- Start points and end points are usually represented by LOTOS gates in the prototype. They can then be controlled and observed during the validation.
- UCM responsibilities are also represented as gates, sometimes with additional message exchanges (to insure causality across components).
- LOTOS gates representing UCM responsibilities and channels that are not observable are hidden through the hide operator.
- UCM components are represented as processes synchronized on their shared channels/gates. The structure is specified mostly in a resource-oriented style [39], with multi-way synchronization (|[...]|) and interleaving (|||) operators.
- Containment of components is maintained. If a component, represented by a LOTOS process, contains sub-components, then the processes representing these are instantiated (and possibly defined) within the former process.
- If multiple path segments (possibly from different UCM scenarios) cross one component, they are integrated together in the same LOTOS process, often as alternatives.
- Elementary processes are specified mostly in a state-oriented style, with choice ([]) and action prefix (;) operators, and with guarded behaviors ([...] $\rightarrow$).
- Abstract data types are used to represent operations, and conditions (LOTOS guard expressions).
- Symmetry is enforced in synchronized actions: actions in one component/process must be mirrored in the other synchronized processes, unless locally hidden.
- Components with stubs have sub-processes, one for each stub. For dynamic stubs, these processes specify the selection policy, i.e. the type of composition between the possible plug-ins.
- Dynamic stubs may have multiple sub-processes, one for each plug-in, whereas static stubs are refined directly by the process representing their only plug-in.
- Stub processes receive a list of entry/exit points as input and then output another such list upon termination.
- Time is not intrinsic to LOTOS. Timers are represented as processes and timeouts are represented using actions. A timeout is simulated by the execution of its corresponding action.
- OR-Forks are represented using the *choice* operator.
- OR-Join, AND-Fork/Join and rendezvous are represented using the LOTOS synchronization mechanisms.

Stubs & Plug-ins

Stubs and plug-ins are elements of agents. DA, CA, LA and CallObject are agents composed of several static and dynamic stubs. These stubs can in turn contain other stubs and so on. Stubs and plug-ins are translated into LOTOS processes and instantiations. A static stub is represented as a process. A dynamic stub is represented by the instantiation of a process of the same name. This process handles the instantiation of one or many of the sub-processes based on selection policies; each sub-process corresponding to a plug-in of the dynamic stub. Those choices result in a LOTOS specification which is close to the UCM model. It also makes the translation of features a straightforward technique, consisting in following the UCM model and choosing adequate plug-ins.

Agents & Database

Agents and Database are at the same level, namely the root of the system. The database is represented as a process that is initialized at the time of its instantiation. All agents are uniquely identified.

DAs are static: for any particular DA, there exists one instance. Each DA process is instantiated with values corresponding to its identifier (ID) and the user with whom it is associated. CAs, LAs and Call Objects are managed in a more dynamic fashion. CAs and LAs can be involved in several communication scenarios at a time. To permit this, CA and LA processes are designed such that there may be many instances of them at a give time. In addition to their unique *process* identifier, each *instance* of a given CA or LA process is also uniquely identified. These agents are therefore identified by a tuple $\{id, instance\}$.

The first parameter, *id*, is the main identifier, much like DA's ID; the second parameter, *instance*, represents a specific instance of the CA or the LA process. Immediately after the initialization of the system, there exists only one instance of each agent. When a CA is contacted, it handles the communication and concurrently replicates itself into a new instance. This new instance satisfies future requests. The synchronization between distinct instances is done using *gate splitting*. Hence, two CAs can use the same LA without conflict; that is, each CA communicates with a different LOTOS instance of the same LA.

Call Objects are managed similarly. A Call Object process is identified solely by its instance identifier. Immediately after the initialization of the system, there is no Call Object process in existence. The creation of Call Object instances is handled dynamically by a special process (Call Object Creator) which initially waits for requests. When a LA asks for the creation of a call, the Call Object Creator creates a new Call Object instance and gives it a specific instance identifier to avoid the conflicts between existing instances.

3.3.3 Features Integration

Once the basic LOTOS specification is completed and conforms to the UCM model, features can be integrated. Ten features have been implemented: *Call Forward Always (CFA)*, *Call Forward on Busy (CFB)*, *Outgoing Call Screening (OCS)*, *Call Hold (CH)*, *Call Pickup (CP)*, *Call Transfer (CT)*, *Auto Recall (AR)*, *Time Reminder (TMR)* and *Call Waiting (CW)*. The LOTOS specification being close to the Use Case Maps model, implementing a feature consists in using its plug-ins and modifying the dynamic stubs composing the feature.

The system contains about one hundred and ten stubs and locating them is a tedious task. The difficulty increases with the depth of recursion. The use of specialized techniques reduces this difficulty. One such technique consists in representing the components as a graph, using tools like *GraphViz (Graph Visualization*, a tool from Bell Labs). For any feature, the UCM are browsed while taking note of the components used and their relationships. This textual description is fed by *GraphViz*, which generates a graph from it. The graphical representation of the components helps to identify the parts that need be modified.

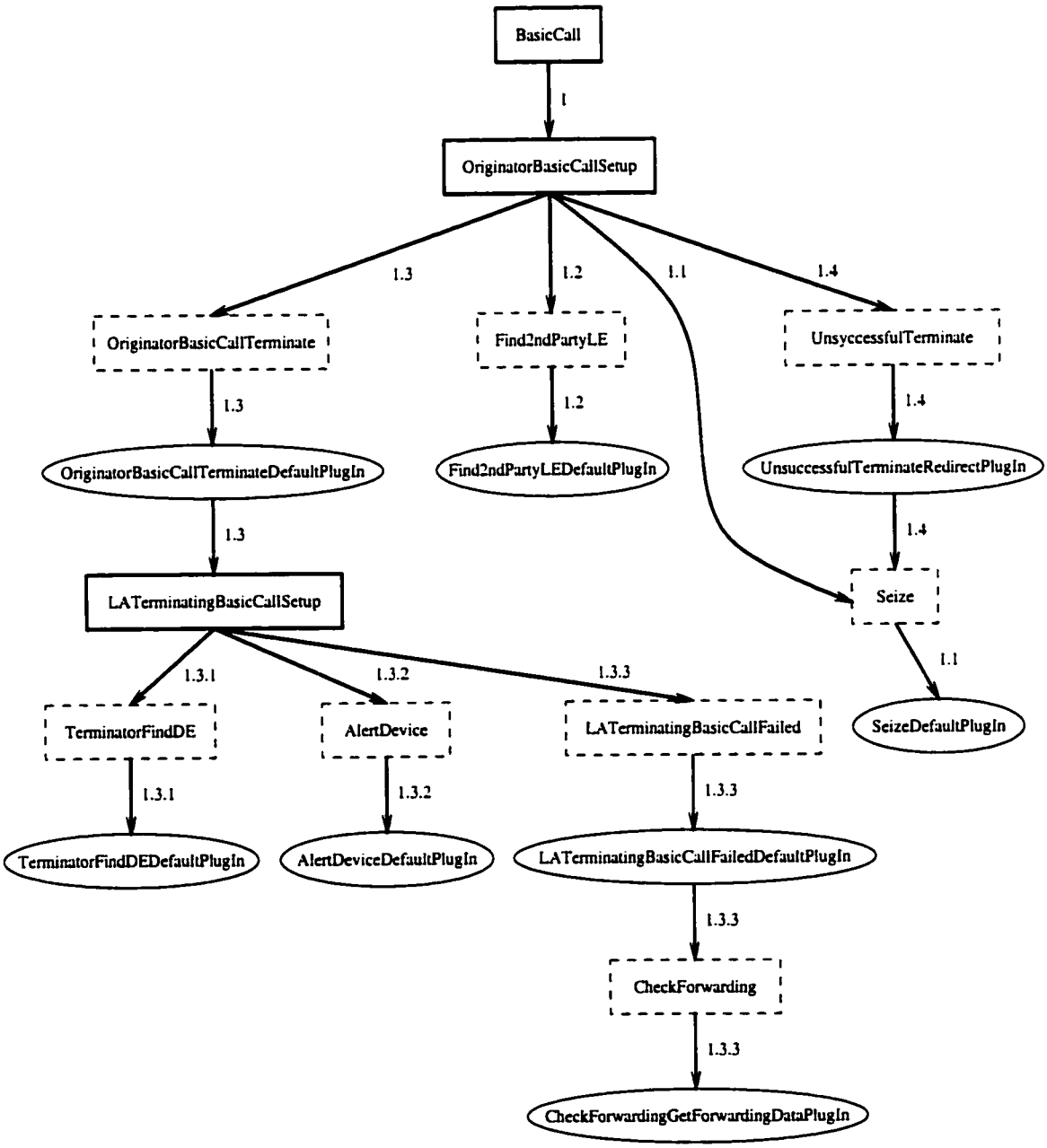


Figure 3.11: Graph of Stubs & Plug-Ins of CFB

Fig. 3.11 illustrates the graph generated by *GraphViz* for *Call Forward on Busy*, where static stubs are represented by solid boxes, dynamic stubs by dashed boxes and plug-ins by ellipses.

The LOTOS specification is 2410 lines long. This breaks down into 450 lines of Abstract Data Types (ADT) and 1960 lines of behavior.

3.4 In Summary

In this chapter, we presented the specification part of our project. We presented the Use Case Maps notation as well as the basic LOTOS concepts. In addition, we presented the system and the features provided by Mitel Corp. as well as guidelines for the translation of the UCM model into a LOTOS specification. This permits expressing requirements in a formal and executable model and, thus, to perform tool-aided validation of the model.

Chapter 4

Identification of Incoherences – Theory

As mentioned in chapter 2, identifying feature interactions is not a trivial task and it is believed that formal and automatic methods are helpful. Moreover, identifying interactions early in the development process reduces the cost of testing. We propose a filtering method that is based on the principle of logic and resolution proof [40]. Our method allows to identify incoherences corresponding to potential pair-wise feature interactions at the requirements stage.

This method deals with features without considering the system as a whole specification. It only identifies incoherences on which the tester should focus. We define a simple formal notation for the description of features using information obtained from requirements. We define and verify formulas relative to specific types of incoherences that correspond to possible pair-wise feature interactions.

4.1 Features Representation

A feature implements a functionality that applies to specific states or situations of a call process. When a call process is in a state for which a given feature is defined, the functionality of the feature is used instead of, or in addition to, the functionality provided by the base system. Since a call serves the communication purposes of users, features are associated, or bound, to users. We see the evolution of a call process as a sequence of events bringing the system from a state to another. Thus, the activation of a feature, and the use of its functionality, depends on the state in which the system is and the event that occurs. A feature in a certain state is triggered by some specific event and the use of such a feature leads to some specific results according to the functionality provided.

4.1.1 Basic Principles

A feature implementing a functionality that presents behaviors at different states is broken into several description parts, each one representing a different behavior. Hence, the name of a description part is composed of the name of the feature itself and the part concerned. Each description part is decomposed into 4 specific groups of properties: *pre-conditions*, *triggering*

events, *results* and *constraints*. Properties are composed of a name and a set of elements that can be variables or sub-properties.

Pre-conditions – represent the mandatory conditions for the activation of the feature. These conditions describe the state in which the system must be before the activation of the feature can be considered.

Triggering events – represent the action(s) triggering the feature

Results – are the actions produced by the execution of the feature and the state in which the system is after such execution.

Constraints – are the restrictions relative to the variables used in pre-conditions, triggering events and results describing the feature.

4.1.2 Formalism

As mentioned, properties are composed of a name and a set of elements. The set of elements can be empty (i.e. the property is only a name) or can contain a list of variables representing the users concerned with the property. In order to distinguish names from variables, property names start with a lowercase while variable names start with an uppercase. The grammar defining the syntax of a property is the following:

```
PROPERTY    ::=    NAME "(" VARLIST ")" | NAME
VARLIST     ::=    VAR | VARLIST "," VAR
NAME        ::=    [a-z][A-Za-z0-9]*
VAR         ::=    [A-Z][A-Za-z0-9]*
```

A property is formally denoted $\gamma(v_1, ..., v_i)$, $i > 0$, where γ stands for the name of the property and the list $v_1, ..., v_i$ are variables representing entities concerned with the property. Property names are not limited by any specific rule. They only need to be homogeneous with respect to different features. Two names are however reserved for specific mandatory properties; `subs` and `concerns`, defined below:

- `subs(U, F)` – represents a user, denoted by the variable U, that subscribes to a feature, denoted by variable F. For instance, the fact that the user Bob subscribes to the feature Call Waiting, represented by the acronym `cw`, is denoted: `subs(bob, cw)`, where U = bob and F = cw.

- `concerns(U, F)` – stipulates that the usual behavior of user U can be influenced by the feature F. For instance, Call Waiting implements the possibility, for its subscriber, to hold an incoming call when already involved in another call. Call Waiting also specifies that if someone is held and receives a call, this latter user must be considered as being busy. In that case, the user concerned is not the subscriber but the one on hold. Alice being concerned with Call Waiting is denoted `concerns(alice, cw)`.

A feature X is denoted $\mathcal{F}_X$ and can be broken into several description parts, each one relative to one of the behaviors of the functionality. A description part is denoted $\mathcal{D}_{X,m}$ where X stands for the name of the feature and m is an integer identifying the description part. For example, a feature X implementing with three behavioral parts is broken into three description parts $\mathcal{D}_{X,1}$, $\mathcal{D}_{X,2}$ and $\mathcal{D}_{X,3}$. The feature $\mathcal{F}_X$ can be stated as $\mathcal{F}_X = \{\mathcal{D}_{X,1}, \mathcal{D}_{X,2}, \mathcal{D}_{X,3}\}$.

As for a description part, each of the four groups of properties composing it is denoted by a letter that corresponds to its type and is indexed with the name of the feature and the number of the description part. Given a description part $\mathcal{D}_{X,m}$, the groups are denoted as follows:

- The set of pre-conditions is denoted $\mathcal{P}_{X,m}$. It is an ordered set of properties formally defined as: $\mathcal{P}_{X,m} = \{\gamma_1(v_{1,1}, ..., v_{1,i}), ..., \gamma_j(v_{j,1}, ..., v_{j,i})\}$, $i, j > 0$, where $\gamma_1, ..., \gamma_j$ are properties representing pre-conditions. Pre-conditions must always contain the two properties **subs** and **concerns**, denoting the subscriber and the user influenced by the functionality (in most cases, the subscriber itself).

- Following the same pattern, the set of triggering events is an ordered set denoted $\mathcal{T}_{X,m}$, and we express $\mathcal{T}_{X,m} = \{\gamma_1(v_{1,1}, ..., v_{1,i}), ..., \gamma_j(v_{j,1}, ..., v_{j,i})\}$, $i, j > 0$, where $\gamma_1, ..., \gamma_j$ are properties representing triggering events.

- Based again on the same pattern, the set of results is an ordered set denoted $\mathcal{R}_{X,m}$, and we express $\mathcal{R}_{X,m} = \{\gamma_1(v_{1,1}, ..., v_{1,i}), ..., \gamma_j(v_{j,1}, ..., v_{j,i})\}$, $i, j > 0$, where $\gamma_1, ..., \gamma_j$ are properties representing results.

- The set of constraints does not contain properties but restrictions. It identifies restrictions on the variables involved, which represent users. This set is denoted $\mathcal{C}_{X,m}$, and, $\mathcal{C}_{X,m} = \{(v_1 \Re_1 v_2), ..., (v_i \Re_j v_{i+1})\}$, $i, j \geq 0$, where $v_1, ..., v_{i+1}$ are the variables used in $\mathcal{P}_{X,m}$, $\mathcal{T}_{X,m}$ and $\mathcal{R}_{X,m}$, and, $\Re_1, ..., \Re_j$ represent the relations between variables regarding equality ($=$) or inequality ($\neq$). Such constraints can represent the fact that a caller must be different from the callee, or that they must be the same user.

4.1.3 Example

Let us consider an example of feature definition: `Call Waiting`. `Call Waiting` implements the following functionality: when the subscriber is busy and receives a call, the caller is put on hold and the subscriber receives a notification signal. Then, pressing a special button, the subscriber can switch from the talking party to the held party and vice-versa. If a call request is received by the held party, a busy notification is sent to the caller. We distinguish three behaviors and thus, three description parts that are:

1. The subscriber is talking, receives a call and gets a notification.

2. The subscriber is holding a caller and switches from the talking party to the held one.

3. The held party receives a call request and sends a busy indication.

In order to formalize the description of this feature, we use the acronym `cw` to represent Call Waiting and we establish the list of properties corresponding to the requirements above:

- `subs(A, cw)` - User A subscribes to Call Waiting.
- `concerns(B, cw)` - Call Waiting concerns user B.
- `busy(B)` - User B is busy.
- `talk(A, B)` - User A is talking with user B.
- `call(A, C)` - Call attempt from a user A to another user C.
- `hold(A, C)` - User A holds user C.
- `cw_notify(A)` - User A receives a *hold* notification.
- `flash(A)` - User A presses flash button to switch between parties.
- `busy_ind(C, D)` - User C sends a busy indication to user D.

We formally express the feature description parts 1, 2 and 3 as:

- $\mathcal{D}_{cw,1}$, Call Waiting, description part 1: subscriber A receives a call from C.
  $\mathcal{P}_{cw,1}$ = { subs(A, cw), concerns(A, cw), busy(A), talk(A, B) }
  $\mathcal{T}_{cw,1}$ = { call(C, A) }
  $\mathcal{R}_{cw,1}$ = { hold(A, C), cw_notify(A) }
  $\mathcal{C}_{cw,1}$ = { (A $\neq$ B), (A $\neq$ C), (B $\neq$ C) }
- $\mathcal{D}_{cw,2}$, Call Waiting, description part 2: subscriber A switches between parties.
  $\mathcal{P}_{cw,2}$ = { subs(A, cw), concerns(A, cw), busy(A), talk(A, B), hold(A, C) }
  $\mathcal{T}_{cw,2}$ = { flash(A) }
  $\mathcal{R}_{cw,2}$ = { hold(A, B), talk(A, C) }
  $\mathcal{C}_{cw,2}$ = { (A $\neq$ B), (A $\neq$ C), (B $\neq$ C) }
- $\mathcal{D}_{cw,3}$, Call Waiting, description part 3: held party C receives a call from D.
  $\mathcal{P}_{cw,3}$ = { subs(A, cw), concerns(C, cw), busy(A), talk(A, B), hold(A, C) }
  $\mathcal{T}_{cw,3}$ = { call(D, C) }
  $\mathcal{R}_{cw,3}$ = { busy_ind(C, D) }
  $\mathcal{C}_{cw,3}$ = { (A $\neq$ B), (A $\neq$ C), (A $\neq$ D), (B $\neq$ C), (B $\neq$ D), (C $\neq$ D) }

Parts 1 & 2 of the feature concern the subscriber since the functionality is added to him. In part 3, the user concerned is not the subscriber but the held party. We consider that the held party is the concerned user because his behavior is modified by the feature (which is external to him) in case of an incoming call. All users are assumed to be different because we are addressing only the general case. Particular cases such as the one where a user calls itself were not considered. Such cases can be explicitly considered by using appropriate constraints.

4.2 Incoherences

The identification of incoherences between two features is based on the identification of specific incoherences between their set of properties. We represent these incoherences in the form of rules and check whether one or more of the rules can be satisfied or not by binding variables of feature descriptions to users. If a rule is satisfied, an incoherence is identified.

Two properties can present a contradiction. When formally defining a feature, contradictory pairs of properties must be identified and listed for future analysis purposes. A contradiction taking place between two properties is stated using a binary relation denoted $\mathcal{K}$ $(\gamma_1(\upsilon_{1,1}, ..., \upsilon_{1,i}), \gamma_2(\upsilon_{2,1}, ..., \upsilon_{2,j}))$, i, j > 1, where γ_1 and γ_2 are the two properties considered and υ_1, υ_2 are the variables used to represent users.

Variables allow to express particular restrictions such as the users concerned with two properties must be the same. For instance, let us say that `busy(X)` denotes that a user, represented by the variable X, is busy. And let us say that `idle(Y)` denotes that a user, represented by the variable Y, is idle. A user being idle and busy at the same time not being permitted can be expressed as contradiction: $\mathcal{K}$(`busy(A)`, `idle(A)`). The use of the same variable `A` indicates that the two properties are in contradiction when they refer to the same user. This can be similarly expressed by $\mathcal{K}$(`busy(A)`, `idle(B)`) if `A` = `B`.

Unbound properties are properties for which variables have no fixed values. As mentioned, they are represented with their name and variables such as: $\gamma_1(\upsilon_1, ..., \upsilon_i)$. To simplify the notation, properties for which variables are bound (have been given a value) are represented without their variables. An unbound property, denoted $\gamma_1(\upsilon_1, ..., \upsilon_i)$ is denoted π_1 when bound.

Note that, in order to simplify the notation, we use the same names $\mathcal{F}_X$, $\mathcal{D}_{X,m}$, etc. to denote sets of properties over variables or over constants.

4.2.1 Definitions

In order to formalize incoherences identification, a few definitions need to be stated:

- We define the combination of two feature descriptions as the ordered pair of the two descriptions where the variables of a description are kept distinct from the ones of the other description, even if they have the same names. Given two feature descriptions $\mathcal{D}_{X,m}$ and $\mathcal{D}_{Y,n}$, we formally denote their combination $|\mathcal{D}_{X,m} \bullet \mathcal{D}_{Y,n}|$.
- The same user cannot subscribe to the same feature more than once. We express this fact as a general contradiction denoted $\mathcal{K}$(subs(A, X), subs(A, X)) where the variable A represents the user and variable X represents the name of the feature.
- We denote $\mathcal{S}(\mathcal{C}_{X,m})$ the fact that all constraints of the constraint set $\mathcal{C}_{X,m}$ are satisfied.
- We define Υ the finite set of all possible users that can be used to bind variables in properties as $\Upsilon = \{\upsilon_1, ..., \upsilon_n\}$, for some n representing the number of users.
- Two distinct ordered sets $\mathcal{E}_1$ and $\mathcal{E}_2$ present a contradiction if there exists $\pi_1 \in \mathcal{E}_1$ and a $\pi_2 \in \mathcal{E}_2$ such that $\mathcal{K}(\pi_1, \pi_2)$ is satisfied. Note that this is decidable because only a finite number of variables and values are in discussion.

- Two distinct ordered sets $\mathcal{E}_1$ and $\mathcal{E}_2$ are equal to each other ($\mathcal{E}_1 = \mathcal{E}_2$) if they contain the same elements in the same order. Otherwise, they are different ($\mathcal{E}_1 \neq \mathcal{E}_2$).

Incoherences refer to incoherences between behaviors. Hence, an incoherence between two features $\mathcal{F}_X$ and $\mathcal{F}_Y$ is characterized by an incoherence present between a description part $\mathcal{D}_{X,m}$ of $\mathcal{F}_X$ and another description part $\mathcal{D}_{Y,n}$ of $\mathcal{F}_Y$. Different types of incoherences can be identified. An incoherence can be characterized by non determinism, contradiction, or loops. Considering pair-wise feature combinations, we identify six global types of incoherences. The types are gathered in two main categories:

- Direct incoherences - for which identification rules are presented in sub-section 4.2.2
- Transitive incoherences - for which identification rules are presented in sub-section 4.2.3

Variables are bound to users. The number of users to use depends on the number of variables used in the properties of feature descriptions. It must be possible to bind all variables contained in the properties of a feature description to different values. In order to fulfill this requirement, the number of required users may reach up to twice the number of variables of the description that contains the greater number of variables. However, in most cases, only half of these users may be sufficient.

4.2.2 Direct Incoherence Rules

Direct incoherences are present when two description parts $\mathcal{D}_{X,m}$ and $\mathcal{D}_{Y,n}$ are linked together, present the same triggers but lead to different or contradictory results. In both cases, non determinism arises since two possible results, valid from each feature's point of view, are possible. Two feature descriptions $\mathcal{D}_{X,m}$ and $\mathcal{D}_{Y,n}$ can be linked together and present an incoherence for the following reasons:

1. The same user subscribes to both features $\mathcal{F}_X$ and $\mathcal{F}_Y$. An incoherence can be present if:
   - 1a. Features can be triggered by the same event and present results that are different but not contradictory.
   - 1b. Features can be triggered by the same event and present contradictory results.
2. Different users U and V respectively subscribe to $\mathcal{F}_X$ and $\mathcal{F}_Y$ and U is concerned with both features $\mathcal{F}_X$ and $\mathcal{F}_Y$. Again, an incoherence can be present for two reasons:
   - 2a. Features can be triggered by the same event and present results that are different but not contradictory.
   - 2b. Features can be triggered by the same event and present contradictory results.

All cases (1a, 1b, 2a, 2b) result in non-determinism except that, in cases 1a and 2a, results are different and do not present contradictions while in cases 1b and 2b, results are different and present a contradiction. We distinguish four rules for direct incoherences, each one associated to a case: direct incoherence rule #d1, associated to case 1a; direct incoherence rule #d2, associated to case 1b; direct incoherence rule #d3, associated to case 2a; direct incoherence rule #d4, associated to case 2b.

Direct Incoherence Rule #d1

Rule #d1 identifies incoherences between features subscribed by the same user, triggered by the same event and yielding different but not contradictory results. Let us consider two feature descriptions $\mathcal{D}_{X,m}$ and $\mathcal{D}_{Y,n}$ combined in a pair $|\mathcal{D}_{X,m} \bullet \mathcal{D}_{Y,n}|$. An incoherence is present if, from Υ, the set of all possible users, it is possible to find a binding of property variables such that the following holds:

1. The same user subscribes to both $\mathcal{F}_X$ and $\mathcal{F}_Y$,
2. $\mathcal{P}_{X,m}$ and $\mathcal{P}_{Y,n}$ do not present a contradiction,
3. $\mathcal{T}_{X,m}$ and $\mathcal{T}_{Y,n}$ are the same,
4. $\mathcal{R}_{X,m}$ and $\mathcal{R}_{Y,n}$ are different,
5. $\mathcal{R}_{X,m}$ and $\mathcal{R}_{Y,n}$ do not present a contradiction,
6. $\mathcal{S}(\mathcal{C}_{X,m})$ and $\mathcal{S}(\mathcal{C}_{Y,n})$, constraints of both descriptions are satisfied.

Due to its configuration (features are subscribed by the same user, have the same triggering events and different results), the incoherences identified by this rule are *symmetric*. If the rule identifies an incoherence in $|\mathcal{D}_{X,m} \bullet \mathcal{D}_{Y,n}|$, it also identifies the same incoherence in $|\mathcal{D}_{Y,n} \bullet \mathcal{D}_{X,m}|$. Similarly, if no incoherence is found in one pair, none will be found in the other one. Hence, the rule only needs to be applied to one of the two pairs. The rule is formally represented as follows:

Variables of properties are bound using values provided by Υ such that:

1. $\exists\, \upsilon \in \Upsilon \mid (\text{subs}(\upsilon, X) \in \mathcal{P}_{X,m} \wedge \text{subs}(\upsilon, Y) \in \mathcal{P}_{Y,n})$
2. $\nexists(\pi_1 \in \mathcal{P}_{X,m}, \pi_2 \in \mathcal{P}_{Y,n}) \mid \mathcal{K}(\pi_1, \pi_2)$
3. $\mathcal{T}_{X,m} = \mathcal{T}_{Y,n}$
4. $\mathcal{R}_{X,m} \neq \mathcal{R}_{Y,n}$
5. $\nexists(\pi_3 \in \mathcal{R}_{X,m}, \pi_4 \in \mathcal{R}_{Y,n}) \mid \mathcal{K}(\pi_3, \pi_4)$
6. $\mathcal{S}(\mathcal{C}_{X,m}) \wedge \mathcal{S}(\mathcal{C}_{Y,n})$

As an example, let us consider two features: **Call Forward on Busy** and **Incoming Call Screening**. **Call Forward on Busy** implements the possibility, for the subscriber, to forward incoming calls to another party when busy. **Incoming Call Screening** allows the subscriber to establish a screening list and to deny calls incoming from users contained in this list. The list of properties corresponding to the requirements follows:

- **busy(B)** – User B is busy.

- `call(A, B)` - Call attempt from a user A to another user B.
- `cfb(C)` - C is the party to which calls are forwarded.
- `redirected(A, C)` - User A is forwarded to user C.
- `ics_list(A)` - User A is in the screening list of the subscriber.
- `deny_call(B, A)` - Calls incoming from A are denied by user B.
- `call_denied(B, A)` - User B notifies user A that the call has been denied.

Each feature only implements one functionality. Hence, one description part is sufficient to formalize each feature. In order to consider the general cases of the use of **Call Forward on Busy** and **Incoming Call Screening**, all variables involved in a feature description must have different values. This eliminates particular cases such as those where a user forwards all its calls to himself when he is busy. We use the acronyms **cfb** to denote **Call Forward on Busy** and **ics** to denote **Incoming Call Screening**; the number 1 denotes the (single) description part of each feature:

- $\mathcal{D}_{cfb,1}$ Call Forward on Busy, description part 1, subscriber B is busy and receives a call from A:
  $\mathcal{P}_{cfb,1}$ = { subs(B, cfb), concerns(B, cfb), cfb(C), busy(B) }
  $\mathcal{T}_{cfb,1}$ = { call(A, B) }
  $\mathcal{R}_{cfb,1}$ = { redirected(A, C), call(A, C) }
  $\mathcal{C}_{cfb,1}$ = { (A $\neq$ B), (A $\neq$ C), (B $\neq$ C) }
- $\mathcal{D}_{ics,1}$ Incoming Call Screening, description part 1, screened party C calls B:
  $\mathcal{P}_{ics,1}$ = { subs(B, ics), concerns(B, ics), ics_list(A) }
  $\mathcal{T}_{ics,1}$ = { call(A, B) }
  $\mathcal{R}_{ics,1}$ = { deny_call(B, A), call_denied(B, A) }
  $\mathcal{C}_{ics,1}$ = { (A $\neq$ B) }

Performing pair-wise combination over $\mathcal{D}_{cfb,1}$ and $\mathcal{D}_{ics,1}$, we observe that $|\mathcal{D}_{cfb,1} \bullet \mathcal{D}_{ics,1}|$ presents an incoherence since it is possible to find a binding of variables that satisfies direct incoherence rule #d1. Let us use the notation $X \leftarrow val$ to denote the binding of variable X to value *val* and let us consider three users: *Alice*, *Bob*, *Carol*. We bind the variables of the properties of $\mathcal{D}_{cfb,1}$ and $\mathcal{D}_{ics,1}$ to the following values:

- $\mathcal{D}_{cfb,1}$ - $A \leftarrow$ alice, $B \leftarrow$ bob, $C \leftarrow$ carol.
  $\mathcal{P}_{cfb,1}$ = { subs(bob, cfb), concerns(bob, cfb), cfb(carol), busy(bob) }
  $\mathcal{T}_{cfb,1}$ = { call(alice, bob) }
  $\mathcal{R}_{cfb,1}$ = { redirected(alice, carol), call(alice, carol) }
  $\mathcal{C}_{cfb,1}$ = { (alice $\neq$ bob), (alice $\neq$ carol), (bob $\neq$ carol) }
- $\mathcal{D}_{ics,1}$ - $A \leftarrow$ alice, $B \leftarrow$ bob.
  $\mathcal{P}_{ics,1}$ = { subs(bob, ics), concerns(bob, ics), ics_list(alice) }
  $\mathcal{T}_{ics,1}$ = { call(alice, bob) }
  $\mathcal{R}_{ics,1}$ = { deny_call(bob, alice), call_denied(bob, alice) }
  $\mathcal{C}_{ics,1}$ = { (alice $\neq$ bob) }

Rule #d1 is satisfied for this binding since:

1. subs(bob, cfb) $\in \mathcal{P}_{cfb,1} \wedge$ subs(bob, ics) $\in \mathcal{P}_{ics,1}$
2. $\nexists\, \pi_1 \in \mathcal{P}_{cfb,1}, \pi_2 \in \mathcal{P}_{ics,1} \mid \mathcal{K}(\pi_1, \pi_2)$
3. $\mathcal{T}_{cfb,1} = \mathcal{T}_{ics,1}$
4. $\mathcal{R}_{cfb,1} \neq \mathcal{R}_{ics,1}$
5. $\nexists\, \pi_3 \in \mathcal{R}_{cfb,1}, \pi_4 \in \mathcal{R}_{ics,1} \mid \mathcal{K}(\pi_3, \pi_4)$
6. $\mathcal{S}(\mathcal{C}_{cfb,1}) \wedge \mathcal{S}(\mathcal{C}_{ics,1})$

If Alice calls Bob (call(alice, bob)), **Call Forward on Busy** should be activated and Alice should be redirected to Carol (redirected(alice, carol)) and call Carol (call(alice, carol)). At the same time, **Incoming Call Screening** should also be activated and Alice's call should be denied (deny_call(bob, alice)). We conclude that $\mathcal{D}_{cfb,1}$ and $\mathcal{D}_{ics,1}$ present an incoherence characterized by non-determinism.

Direct Incoherence Rule #d2

Direct incoherence rule #d2 identifies incoherences characterized by contradictory results. Similarly to rule #d1, it identifies *symmetric* incoherences. An incoherence is present if, from Υ, it is possible to find a binding of property variables such that the following holds:

1. The same user subscribes to both $\mathcal{F}_X$ and $\mathcal{F}_Y$,
2. $\mathcal{P}_{X,m}$ and $\mathcal{P}_{Y,n}$ do not present a contradiction,
3. $\mathcal{T}_{X,m}$ and $\mathcal{T}_{Y,n}$ are the same,
4. $\mathcal{R}_{X,m}$ and $\mathcal{R}_{Y,n}$ present a contradiction,
5. $\mathcal{S}(\mathcal{C}_{X,m})$ and $\mathcal{S}(\mathcal{C}_{Y,n})$, constraints of both descriptions are satisfied.

The rule is formally defined as follows:

Variables of properties are bound using values provided by Υ such that:

1. $\exists\, \upsilon \in \Upsilon \mid (\text{subs}(\upsilon, X) \in \mathcal{P}_{X,m} \wedge \text{subs}(\upsilon, Y) \in \mathcal{P}_{Y,n})$
2. $\nexists(\pi_1 \in \mathcal{P}_{X,m}, \pi_2 \in \mathcal{P}_{Y,n}) \mid \mathcal{K}(\pi_1, \pi_2)$
3. $\mathcal{T}_{X,m} = \mathcal{T}_{Y,n}$
4. $\exists(\pi_3 \in \mathcal{R}_{X,m}, \pi_4 \in \mathcal{R}_{Y,n}) \mid \mathcal{K}(\pi_3, \pi_4)$
5. $\mathcal{S}(\mathcal{C}_{X,m}) \wedge \mathcal{S}(\mathcal{C}_{Y,n})$

As an example, let us consider the feature definition of **Call Waiting** given in section 4.1.3 and combine it with the feature definition of **Incoming Call Screening** given in section 4.2.2. In addition to the list of properties previously defined for these two features, we consider that a user cannot hold a call and deny it at the same time. This contradiction is expressed by the following relation:

- $\mathcal{K}$(hold(B, A), deny_call(B, A)) - A user B cannot, at the same time, hold and deny a call incoming from user A.

Let us consider that the system contains 3 possible users: *Alice*, *Bob*, *Carol* and let us bind variables of $\mathcal{D}_{cw,1}$ and $\mathcal{D}_{ics,1}$ to users such that their constraints are satisfied. A binding of properties variables of both descriptions satisfying the rule #d2 is:

- $\mathcal{D}_{cw,1}$ - $A \leftarrow$ alice, $B \leftarrow$ bob, $C \leftarrow$ carol.
  $\mathcal{P}_{cw,1}$ = { subs(bob, cw), concerns(bob, cw), busy(bob) talk(bob, carol) }
  $\mathcal{T}_{cw,1}$ = { call(alice, bob) }
  $\mathcal{R}_{cw,1}$ = { hold(bob, alice), cw_notify(bob) }
  $\mathcal{C}_{cw,1}$ = { (alice $\neq$ bob), (alice $\neq$ carol), (bob $\neq$ carol) }
- $\mathcal{D}_{ics,1}$ - $A \leftarrow$ alice, $B \leftarrow$ bob.
  $\mathcal{P}_{ics,1}$ = { subs(bob, ics), concerns(bob, ics), ics_list(alice) }
  $\mathcal{T}_{ics,1}$ = { call(alice, bob) }
  $\mathcal{R}_{ics,1}$ = { deny_call(bob, alice), call_denied(bob, alice) }
  $\mathcal{C}_{ics,1}$ = { (alice $\neq$ bob) }

The rule is satisfied since:

1. subs(bob, cw) $\in \mathcal{P}_{cw,1} \wedge$ subs(bob, ics) $\in \mathcal{P}_{ics,1}$
2. $\nexists\, \pi_1 \in \mathcal{P}_{cw,1}, \pi_2 \in \mathcal{P}_{ics,1} \mid \mathcal{K}(\pi_1, \pi_2)$
3. $\mathcal{T}_{cw,1} = \mathcal{T}_{ics,1}$
4. $\exists\, \pi_3 \in \mathcal{R}_{cw,1}, \pi_4 \in \mathcal{R}_{ics,1} \mid \mathcal{K}(\pi_3, \pi_4)$
5. $\mathcal{S}(\mathcal{C}_{cw,1}) \wedge \mathcal{S}(\mathcal{C}_{ics,1})$

If Alice calls Bob (call(alice, bob)), **Call Waiting** should be activated, and hold Alice (hold(bob, alice)). At the same time, **Incoming Call Screening** should also be activated and deny the call incoming from Alice (deny_call(bob, alice)) since Alice is in the screening list. If Alice is held by **Call Waiting**, she will be able to talk to Bob later, which should not happen. Hence, $\mathcal{D}_{cw,1}$ and $\mathcal{D}_{ics,1}$ present an incoherence characterized by a contradiction.

Direct Incoherence Rule #d3

Rule #d3 identifies incoherences taking place between features subscribed by different users and both concerning the subscriber of one of the features. An incoherence is present if the following characteristics hold:

1. A user v_1 subscribes to $\mathcal{F}_X$,
   a different user v_2 subscribes to $\mathcal{F}_Y$,
   the user v_2 is concerned with both descriptions $\mathcal{D}_{X,m}$ and $\mathcal{D}_{Y,n}$,
2. $\mathcal{P}_{X,m}$ and $\mathcal{P}_{Y,n}$ do not present a contradiction,
3. $\mathcal{T}_{X,m}$ and $\mathcal{T}_{Y,n}$ are the same,
4. $\mathcal{R}_{X,m}$ and $\mathcal{R}_{Y,n}$ are different,
5. $\mathcal{R}_{X,m}$ and $\mathcal{R}_{Y,n}$ do not present a contradiction,
6. $\mathcal{S}(\mathcal{C}_{X,m})$ and $\mathcal{S}(\mathcal{C}_{Y,n})$, constraints of both features are satisfied.

Rule #d3 identifies similar incoherences as rule #d1 where $\mathcal{T}_{X,m} = \mathcal{T}_{Y,n}$ and $\mathcal{R}_{X,m} \neq \mathcal{R}_{Y,n}$. Given two feature descriptions $\mathcal{D}_{X,m}$ and $\mathcal{D}_{Y,n}$, the rule stipulates that $\mathcal{D}_{X,m}$ must have a user v_1 as a subscriber and must concern a different user v_2 and that $\mathcal{D}_{Y,n}$ must have v_2 as a subscriber and concerned user.

Identifying an incoherence in $|\mathcal{D}_{X,m} \bullet \mathcal{D}_{Y,n}|$ does not insure that the same incoherence is also present in $|\mathcal{D}_{Y,n} \bullet \mathcal{D}_{X,m}|$, which means that, in order to consider all cases, incoherence rule #d3 must be applied to both combinations. In other words, contrary to rule #d1, rule #d3 is not symmetric.

The rule is formally defined as follows:

Variables of properties are bound using values provided by Υ such that:

1. $\exists\, v_1, v_2 \in \Upsilon, v_1 \neq v_2 \mid$
   $(\text{subs}(v_1, X) \in \mathcal{P}_{X,m} \wedge \text{concerns}(v_2, X) \in \mathcal{P}_{X,m}) \wedge$
   $(\text{subs}(v_2, X) \in \mathcal{P}_{Y,n} \wedge \text{concerns}(v_2, Y) \in \mathcal{P}_{Y,n})$
2. $\nexists(\pi_1 \in \mathcal{P}_{X,m}, \pi_2 \in \mathcal{P}_{Y,n}) \mid \mathcal{K}(\pi_1, \pi_2)$
3. $\mathcal{T}_{X,m} = \mathcal{T}_{Y,n}$
4. $\mathcal{R}_{X,m} \neq \mathcal{R}_{Y,n}$
5. $\nexists(\pi_3 \in \mathcal{R}_{X,m}, \pi_4 \in \mathcal{R}_{Y,n}) \mid \mathcal{K}(\pi_3, \pi_4)$
6. $\mathcal{S}(\mathcal{C}_{X,m}) \wedge \mathcal{S}(\mathcal{C}_{Y,n})$

As an example, let us consider **Call Waiting** (defined in section 4.1.3) and **Call Forward on Busy** (defined in section 4.2.2). We observe that $|\mathcal{D}_{cw,3} \bullet \mathcal{D}_{cfb,1}|$ presents an incoherence since it is possible to find a binding of variables that satisfies incoherence rule #d3. Let us consider 4 possible users: *Alice*, *Bob*, *Carol*, *Dave*, and use them to bind variables of $\mathcal{D}_{cw,3}$ and $\mathcal{D}_{cfb,1}$ such that their constraints are satisfied and such that rule #d3 is also satisfied:

- $\mathcal{D}_{cw,3}$ - $A \leftarrow$ alice, $B \leftarrow$ bob, $C \leftarrow$ carol, $D \leftarrow$ dave.

  $\mathcal{P}_{cw,3}$ = { subs(alice, cw), concerns(carol, cw), busy(alice), talk(alice, bob), hold(alice, carol) }
  $\mathcal{T}_{cw,3}$ = { call(dave, carol) }
  $\mathcal{R}_{cw,3}$ = { busy_ind(carol, dave) }
  $\mathcal{C}_{cw,3}$ = { (alice $\neq$ bob), (alice $\neq$ carol), (alice $\neq$ dave), (bob $\neq$ carol), (bob $\neq$ dave), (carol $\neq$ dave) }

- $\mathcal{D}_{cfb,1}$ - $A \leftarrow$ dave, $B \leftarrow$ carol, $C \leftarrow$ alice.

  $\mathcal{P}_{cfb,1}$ = { subs(carol, cfb), concerns(carol, cfb), cfb(alice), busy(carol) }
  $\mathcal{T}_{cfb,1}$ = { call(dave, carol) }
  $\mathcal{R}_{cfb,1}$ = { redirected(dave, alice), call(dave, alice) }
  $\mathcal{C}_{cfb,1}$ = { (dave $\neq$ carol), (dave $\neq$ alice), (carol $\neq$ alice) }

The rule is satisfied since:

1. (subs(alice, cw) $\in \mathcal{P}_{cw,3} \wedge$ concerns(carol, cw) $\in \mathcal{P}_{cfb,1}) \wedge$
   (subs(carol, cfb) $\in \mathcal{P}_{cw,3} \wedge$ concerns(carol, cfb) $\in \mathcal{P}_{cfb,1}$)
2. $\nexists\, \pi_1 \in \mathcal{P}_{cw,3}, \pi_2 \in \mathcal{P}_{cfb,1} \mid \mathcal{K}(\pi_1, \pi_2)$
3. $\mathcal{T}_{cw,3} = \mathcal{T}_{cfb,1}$
4. $\mathcal{R}_{cw,3} \neq \mathcal{T}_{cfb,1}$
5. $\nexists\, \pi_3 \in \mathcal{R}_{cw,3}, \pi_4 \in \mathcal{R}_{cfb,1} \mid \mathcal{K}(\pi_3, \pi_4)$
6. $\mathcal{S}(\mathcal{C}_{cw,3}) \wedge \mathcal{S}(\mathcal{C}_{cfb,1})$

Features have different subscribers but concern the same user. Carol is busy and held by Alice. When Carol receives a call (call(dave, carol)), `Call Waiting`, on Alice side, informs that the switch should send a busy indication to the caller (busy_ind(carol, dave)), while `Call Forward on Busy`, on her side, informs that the call should be forwarded (redirected(dave, alice)). We face a non determinism since both feature parts are triggered by the same event and lead to different results. Thus, $\mathcal{D}_{cw,3}$ and $\mathcal{D}_{cfb,1}$ present an incoherence characterized by a non-determinism.

Direct Incoherence Rule #d4

Similarly to direct incoherence rule #d3, rule #d4 identifies incoherences present between features subscribed by different users and concerning the same user. Again, for partitioning issues, this rule is similar to #d2 since it concerns the set of incoherences characterized by contradictory results. Also, as for #d3, due to its configuration, this rule does not identify *symmetric* incoherences.

As for all other rules defined before, we consider two feature descriptions $\mathcal{D}_{X,m}$ and $\mathcal{D}_{Y,n}$. An incoherence is present if, from Υ, the set of all possible users, it is possible to find a binding of property variables such that the combination of the two features considered are the following:

1. A user v_1 subscribes to $\mathcal{F}_X$,
   a different user v_2 subscribes to $\mathcal{F}_Y$,
   the user v_2 is concerned with both descriptions $\mathcal{D}_{X,m}$ and $\mathcal{D}_{Y,n}$,
2. $\mathcal{P}_{X,m}$ and $\mathcal{P}_{Y,n}$ do not present a contradiction,
3. $\mathcal{T}_{X,m}$ and $\mathcal{T}_{Y,n}$ are the same,
4. $\mathcal{R}_{X,m}$ and $\mathcal{R}_{Y,n}$ present a contradiction,
5. $\mathcal{S}(\mathcal{C}_{X,m})$ and $\mathcal{S}(\mathcal{C}_{Y,n})$, constraints of both features are satisfied.

The rule is formally defined as follows:

Variables of properties are bound using values provided by Υ such that:

1. $\exists\, v_1, v_2 \in \Upsilon, v_1 \neq v_2 \mid$
   $(\text{subs}(v_1, X) \in \mathcal{P}_{X,m} \wedge \text{concerns}(v_2, X) \in \mathcal{P}_{X,m}) \wedge$
   $(\text{subs}(v_2, X) \in \mathcal{P}_{Y,n} \wedge \text{concerns}(v_2, Y) \in \mathcal{P}_{Y,n})$
2. $\mathcal{B}(\pi_1 \in \mathcal{P}_{X,m}, \pi_2 \in \mathcal{P}_{Y,n}) \mid \mathcal{K}(\pi_1, \pi_2)$
3. $\mathcal{T}_{X,m} = \mathcal{T}_{Y,n}$
4. $\exists(\pi_3 \in \mathcal{R}_{X,m}, \pi_4 \in \mathcal{R}_{Y,n}) \mid \mathcal{K}(\pi_3, \pi_4)$
5. $\mathcal{S}(\mathcal{C}_{X,m}) \wedge \mathcal{S}(\mathcal{C}_{Y,n})$

As an example, let us consider again the feature definitions of `Call Waiting` (defined in section 4.2.2) and `Incoming Call Screening` (defined in section 4.1.3). Let us also consider that sending a busy indication and denying a call presents a contradiction if applied to the same user and let us denote this using the contradiction relation:

- $\mathcal{K}$(busy_ind(B, A), deny_call(B, A)) - User B cannot send a busy indication to A and deny the call from A at the same time.

Performing pair-wise combination over the two feature descriptions $\mathcal{D}_{cw,3}$ and $\mathcal{D}_{ics,1}$, we observe that $|\mathcal{D}_{cw,3} \bullet \mathcal{D}_{ics,1}|$ presents an incoherence since it is possible to find a binding of variables that satisfies the incoherence rule. Let us consider that the system contains 4 possible users: *Alice*, *Bob*, *Carol*, *Dave*, and let us bind variables of $\mathcal{D}_{cw,3}$ and $\mathcal{D}_{ics,1}$ to users, such that their constraints are satisfied:

- $\mathcal{D}_{cw,3}$ - $A \leftarrow$ alice, $B \leftarrow$ bob, $C \leftarrow$ carol, $D \leftarrow$ dave.
  $\mathcal{P}_{cw,3}$ = { subs(alice, cw), concerns(carol, cw), busy(alice),
  talk(alice, bob), hold(alice, carol) }
  $\mathcal{T}_{cw,3}$ = { call(dave, carol) }
  $\mathcal{R}_{cw,3}$ = { busy_ind(carol, dave) }
  $\mathcal{C}_{cw,3}$ = { (alice $\neq$ bob), (alice $\neq$ carol), (alice $\neq$ dave), (bob $\neq$ carol),
  (bob $\neq$ dave), (carol $\neq$ dave)}

- $\mathcal{D}_{ics,1}$ - $A \leftarrow$ dave, $B \leftarrow$ carol, $C \leftarrow$ alice.
  $\mathcal{P}_{ics,1}$ = { subs(carol, ics), concerns(carol, ics), ics_list(dave) }
  $\mathcal{T}_{ics,1}$ = { call(dave, carol) }
  $\mathcal{R}_{ics,1}$ = { deny_call(carol, dave) }
  $\mathcal{C}_{ics,1}$ = { (dave $\neq$ carol), (dave $\neq$ alice), (carol $\neq$ alice) }

The rule is satisfied since:

1. (subs(alice, cw) $\in \mathcal{P}_{cw,3} \wedge$ concerns(carol, cw) $\in \mathcal{P}_{ics,1}) \wedge$
   (subs(carol, ics) $\in \mathcal{P}_{cw,3} \wedge$ concerns(carol, ics) $\in \mathcal{P}_{ics,1}$)
2. $\nexists\ \pi_1 \in \mathcal{P}_{cw,3}, \pi_2 \in \mathcal{P}_{ics,1} \mid \mathcal{K}(\pi_1, \pi_2)$
3. $\mathcal{T}_{cw,3} = \mathcal{T}_{ics,1}$
4. $\exists\ \pi_3 \in \mathcal{R}_{cw,3}, \pi_4 \in \mathcal{R}_{ics,1} \mid \mathcal{K}(\pi_3, \pi_4)$
5. $\mathcal{S}(\mathcal{C}_{cw,3}) \wedge \mathcal{S}(\mathcal{C}_{ics,1})$

Features have different subscribers but concern the same user. Carol is busy and held by Alice. When Carol receives a call (call(dave, carol)), `Call Waiting`, on Alice's side, informs that she should send a busy indication to the caller (busy_ind(carol, dave)), while `Incoming Call Screening`, on her side, informs that the call should be denied (deny_call(carol, dave)). We face a contradiction. Thus, we can conclude that $\mathcal{D}_{cw,3}$ and $\mathcal{D}_{ics,1}$ present an incoherence characterized by a contradiction.

4.2.3 Transitive Incoherence Rules

The word *transitive* refers to transitivity with respect to features. Namely, the two features could be $\mathcal{D}_{X,m}$ and $\mathcal{D}_{Y,n}$, subscribed by a different or the same user, where the results of $\mathcal{D}_{X,m}$ trigger $\mathcal{D}_{Y,n}$. We identify two possible incoherences for this situation: contradictions, when the results of the two descriptions present a contradiction, and loops, when the results of a description trigger the other one and vice-versa. Contradictions are identified by transitive incoherence rule #t1 and loops by transitive incoherence rule #t2.

The transitivity relation is an implication relation where $\mathcal{D}_{X,m}$ *implies* $\mathcal{D}_{Y,n}$, which can be denoted $\mathcal{D}_{X,m} \Rightarrow \mathcal{D}_{Y,n}$ or also $\mathcal{R}_{X,m} \Rightarrow \mathcal{T}_{Y,n}$.

Transitive Incoherence Rule #t1

Transitive rule #t1 identifies incoherences caused by the results of a feature triggering another feature that has results presenting a contradiction with results of the first feature. As usual, we consider two feature descriptions $\mathcal{D}_{X,m}$ and $\mathcal{D}_{Y,n}$. An incoherence is present if, from Υ, the set of all possible users, it is possible to find a binding of property variables such that the combination of the two features present the following characteristics:

1. $\mathcal{P}_{X,m}$ and $\mathcal{P}_{Y,n}$ do not present a contradiction,
2. $\mathcal{R}_{X,m} \Rightarrow \mathcal{T}_{Y,n}$ ($\mathcal{R}_{X,m}$ is a superset of $\mathcal{T}_{Y,n}$, or, $\mathcal{R}_{X,m} = \mathcal{T}_{Y,n}$)

3. $\mathcal{R}_{X,m}$ and $\mathcal{R}_{Y,n}$ present a contradiction,

4. $\mathcal{S}(\mathcal{C}_{X,m})$ and $\mathcal{S}(\mathcal{C}_{Y,n})$, constraints of both features are satisfied.

The rule is formally defined as follows:

Variables of properties are bound using values provided by Υ such that:

1. $\mathcal{A}(\pi_1 \in \mathcal{P}_{X,m}, \pi_2 \in \mathcal{P}_{Y,n}) \mid \mathcal{K}(\pi_1, \pi_2)$

2. $\mathcal{R}_{X,m} \supseteq \mathcal{T}_{Y,n}$

3. $\exists(\pi_3 \in \mathcal{R}_{X,m}, \pi_4 \in \mathcal{R}_{Y,n}) \mid \mathcal{K}(\pi_3, \pi_4)$

4. $\mathcal{S}(\mathcal{C}_{X,m}) \wedge \mathcal{S}(\mathcal{C}_{Y,n})$

As an example, let us consider a new feature **Outgoing Call Screening** and combine it with the feature **Call Forward on Busy** defined in section 4.2.2. **Outgoing Call Screening** allows the subscriber to establish a screening list and block the calls made to the users that are in this list. The list of properties corresponding to the requirements of **Outgoing Call Screening** is the following:

- `call(A, B)` – Call attempt from user A to user B.
- `ocs_list(A)` – User A is in the screening list of subscriber.
- `block_call(A, B)` – User A blocks outgoing call to user B.
- `call_blocked(A, B)` – User B is notified that his call has been blocked by user A.

Also, we must state that a call cannot be performed and blocked at the same time:

- $\mathcal{K}$(call(A, B), block_call(A, B)) – User A cannot perform a call to user B and block it at the same time.

Using the acronym **ocs** to denote **Outgoing Call Screening**, and number 1 for the single description part, the formal description of feature $\mathcal{D}_{ocs,1}$ is the following:

- $\mathcal{D}_{ocs,1}$ *Outgoing Call Screening, part 1*:
  $\mathcal{P}_{ocs,1}$ = { subs(B, ocs), concerns(B, ocs), ocs_list(A) }
  $\mathcal{T}_{ocs,1}$ = { call(B, A) }
  $\mathcal{R}_{ocs,1}$ = { block_call(B, A), call_blocked(B, A) }
  $\mathcal{C}_{ocs,1}$ = { (A $\neq$ B) }

Performing pair-wise combination over the two feature descriptions $\mathcal{D}_{cfb,1}$ and $\mathcal{D}_{ocs,1}$, we observe that $|\mathcal{D}_{cfb,1} \bullet \mathcal{D}_{ocs,1}|$ presents an incoherence since it is possible to find a binding of variables that satisfies the rule. Let us consider a system containing 3 possible users: *Alice*, *Bob*, *Carol*, and let us find a combination of those users to bind property variables of $\mathcal{D}_{cfb,1}$ and $\mathcal{D}_{ocs,1}$ such that their respective constraints are satisfied, and rule #t1 is also satisfied:

- $\mathcal{D}_{cfb,1}$ - $A \leftarrow$ alice, $B \leftarrow$ bob, $C \leftarrow$ carol.

  $\mathcal{P}_{cfb,1}$ = { subs(bob, cfb), concerns(bob, cfb), cfb(carol), busy(bob) }
  $\mathcal{T}_{cfb,1}$ = { call(alice, bob) }
  $\mathcal{R}_{cfb,1}$ = { redirected(alice, carol), call(alice, carol) }
  $\mathcal{C}_{cfb,1}$ = { (alice $\neq$ bob), (alice $\neq$ carol), (bob $\neq$ carol) }

- $\mathcal{D}_{ocs,1}$ - $A \leftarrow$ carol, $B \leftarrow$ alice.

  $\mathcal{P}_{ocs,1}$ = { subs(alice, ocs), concerns(alice, ocs), ocs_list(carol) }
  $\mathcal{T}_{ocs,1}$ = { call(alice, carol) }
  $\mathcal{R}_{ocs,1}$ = { block_call(alice, carol), call_blocked(alice, carol) }
  $\mathcal{C}_{ocs,1}$ = { (carol $\neq$ alice) }

The rule is satisfied since:

1. $\nexists\, \pi_1 \in \mathcal{P}_{cfb,1}, \pi_2 \in \mathcal{P}_{ocs,1} \mid \mathcal{K}(\pi_1, \pi_2)$
2. $\mathcal{R}_{cfb,1} \supseteq \mathcal{T}_{ocs,1}$
3. $\exists\, \pi_3 \in \mathcal{R}_{cfb,1}, \pi_4 \in \mathcal{R}_{ocs,1} \mid \mathcal{K}(\pi_3, \pi_4)$
4. $\mathcal{S}(\mathcal{C}_{cfb,1}) \wedge \mathcal{S}(\mathcal{C}_{ocs,1})$

If Alice calls Bob (call(alice, bob)), **Call Forward on Busy** is activated and Bob forwards Alice to carol (redirected(alice, carol)). Alice calls Carol (call(alice, carol)) and **Outgoing Call Screening** blocks the call (block_call(alice, carol)). Bob's intention is that Alice calls Carol and Alice's intention is to block calls going to Carol.

Transitive Incoherence Rule #t2

Transitive incoherence rule #t2 identifies incoherences characterized by loops. Two features $\mathcal{D}_{X,m}$ and $\mathcal{D}_{Y,n}$ enter a loop when $\mathcal{D}_{X,m}$ implies $\mathcal{D}_{Y,n}$ and vice-versa. The fact that their results present a contradiction or not is not checked since the goal is to identify loops caused by transitivity of features, not contradictions, already identified by the rule #t1. Given two feature descriptions $\mathcal{D}_{X,m}$ and $\mathcal{D}_{Y,n}$, an incoherence is present if, from Υ, the set of all possible users, it is possible to find a binding of property variables such that the combination of the two features presents the following characteristics:

1. $\mathcal{P}_{X,m}$ and $\mathcal{P}_{Y,n}$ do not present a contradiction,
2. $\mathcal{R}_{X,m} \Rightarrow \mathcal{T}_{Y,n}$ ($\mathcal{R}_{X,m}$ is a superset of $\mathcal{T}_{Y,n}$, or, $\mathcal{R}_{X,m} = \mathcal{T}_{Y,n}$)
3. $\mathcal{R}_{Y,n} \Rightarrow \mathcal{T}_{X,m}$ ($\mathcal{R}_{Y,n}$ is a superset of $\mathcal{T}_{X,m}$, or, $\mathcal{R}_{Y,n} = \mathcal{T}_{X,m}$),
4. $\mathcal{S}(\mathcal{C}_{X,m})$ and $\mathcal{S}(\mathcal{C}_{Y,n})$, constraints of both features are satisfied.

The rule is formally defined as follows:

Variables of properties are bound using values provided by Υ such that:

1. $\not\exists(\pi_1 \in \mathcal{P}_{X,m}, \pi_2 \in \mathcal{P}_{Y,n}) \mid \mathcal{K}(\pi_1, \pi_2)$
2. $\mathcal{R}_{X,m} \supseteq \mathcal{T}_{Y,n}$
3. $\mathcal{R}_{Y,n} \supseteq \mathcal{T}_{X,m}$
4. $\mathcal{S}(\mathcal{C}_{X,m}) \wedge \mathcal{S}(\mathcal{C}_{Y,n})$

As an example, let us consider the feature definition of **Call Forward on Busy** given in section 4.2.2. Let us consider two different instances of the same feature description $\mathcal{D}_{cfb,1}$. Let us differentiate them using $\mathcal{D}_{\alpha-cfb,1}$ for the first instance and $\mathcal{D}_{\beta-cfb,1}$ for the second one, such that their composition is denoted $|\mathcal{D}_{\alpha-cfb,1} \bullet \mathcal{D}_{\beta-cfb,1}|$.

Performing pair-wise combination on them, $|\mathcal{D}_{\alpha-cfb,1} \bullet \mathcal{D}_{\beta-cfb,1}|$ presents an incoherence since it is possible to find a binding of variables that satisfies transitive incoherence rule #t2. Let us consider that the system contains 3 possible users: *Alice*, *Bob*, *Carol*, and let us bind variables of the two instances of $\mathcal{D}_{cfb,1}$ to users, such that their constraints are satisfied:

- $\mathcal{D}_{\alpha-cfb,1}$ - $A \leftarrow$ alice, $B \leftarrow$ bob, $C \leftarrow$ carol.
  $\mathcal{P}_{\alpha-cfb,1}$ = { subs(bob, cw), concerns(bob, cw), cfb(carol), busy(bob)}
  $\mathcal{T}_{\alpha-cfb,1}$ = { call(alice, bob) }
  $\mathcal{R}_{\alpha-cfb,1}$ = { redirected(alice, carol), call(alice, carol) }
  $\mathcal{C}_{\alpha-cfb,1}$ = { (alice $\neq$ bob), (alice $\neq$ carol), (bob $\neq$ carol) }
- $\mathcal{D}_{\beta-cfb,1}$ - $A \leftarrow$ alice, $B \leftarrow$ carol, $C \leftarrow$ bob.
  $\mathcal{P}_{\beta-cfb,1}$ = { subs(carol, cfb), concerns(carol, cfb), cfb(bob), busy(carol) }
  $\mathcal{T}_{\beta-cfb,1}$ = { call(alice, carol) }
  $\mathcal{R}_{\beta-cfb,1}$ = { redirected(alice, bob), call(alice, bob) }
  $\mathcal{C}_{\beta-cfb,1}$ = { (alice $\neq$ carol), (alice $\neq$ bob), (carol $\neq$ bob) }

The rule is satisfied since:

1. $\not\exists\, \pi_1 \in \mathcal{P}_{\alpha-cfb,1}, \pi_2 \in \mathcal{P}_{\beta-cfb,1} \mid \mathcal{K}(\pi_1, \pi_2)$
2. $\mathcal{R}_{\alpha-cfb,1} \supseteq \mathcal{T}_{\beta-cfb,1}$
3. $\mathcal{R}_{\beta-cfb,1} \supseteq \mathcal{T}_{\alpha-cfb,1}$
4. $\mathcal{S}(\mathcal{C}_{\alpha-cfb,1}) \wedge \mathcal{S}(\mathcal{C}_{\beta-cfb,1})$

The subscriber of each instance forwards its incoming calls to the subscriber of the other instance. The trigger of $\mathcal{D}_{\alpha-cfb,1}$ is call(alice, bob). Among the results of $\mathcal{D}_{\alpha-cfb,1}$ we have call(alice, carol). This is the trigger of $\mathcal{D}_{\beta-cfb,1}$ that has among its results call(alice, bob), trigger of $\mathcal{D}_{\alpha-cfb,1}$. As soon as $\mathcal{D}_{\alpha-cfb,1}$ is triggered (call(alice, bob)), its results trigger $\mathcal{D}_{\beta-cfb,1}$ (call(alice, carol)). The results of $\mathcal{D}_{\beta-cfb,1}$ then trigger $\mathcal{D}_{\alpha-cfb,1}$ (call(alice, bob)), and the system enters a loop.

We can then conclude that two different instances α and β of the feature description $\mathcal{D}_{cfb,1}$ can present a loop. This particular incoherence can be avoided by the use of a counter that is bounded to a certain number of iterations. Hence, a user does not forward the same call more than once.

4.3 Priorities

Most of the incoherences are due to the fact that two features try to influence the system in two different ways at the same time. Most of these incoherences can be avoided by using a priority mechanism. Priorities can be represented by the use of contradiction pairs. However, this technique has a limitation, presented in section 4.3.2.

4.3.1 Representing Priorities Using Contradiction Pairs

Let us define a property `priority(P)` where P denotes the priority of a feature. Let us state that two different priorities present a contradiction using the contradiction $\mathcal{K}$(priority(P), priority(Q)) if and only if P $\neq$ Q. By expressing that two features descriptions having different priorities present a contradiction, we state that only features having the same priorities can possibly present an incoherence.

For instance, to express that Call Forward Always has priority over Call Forward on Busy, we give them different priorities. The pre-conditions of the two feature descriptions respectively become:

$\mathcal{P}_{cfa,1}$ = { subs(B, cfb), concerns(B, cfb), cfb(C), priority(0) }

$\mathcal{P}_{cfb,1}$ = { subs(B, cfb), concerns(B, cfb), cfb(C), busy(B), priority(1) }

Due to the contradiction pair $\mathcal{K}$(priority(P), priority(Q)) if and only if P $\neq$ Q, $\mathcal{P}_{cfa,1}$ and $\mathcal{P}_{cfb,1}$ will present a contradiction since their priorities are different and thus, no incoherence will be identified. Moreover, the priorities stated in pre-conditions of each feature provides information to the designer.

4.3.2 Limitation & Possible Solution

The representation used in section 4.3.1 is not sufficient to model complex priority relations such as a feature having priority only over certain features and not over some others.

Let us consider three feature descriptions $\mathcal{D}_{X,1}$, $\mathcal{D}_{Y,1}$ and $\mathcal{D}_{Z,1}$, respectively denoted X, Y and Z to simplify the reading. Let us reason about assigning them priorities:

1. X has priority over Y. This can be modeled by the fact that priority(0) is assigned to X and priority(1) is assigned to Y.

2. Y has priority over Z. This can be modeled by keeping the priority previously assigned to Y (priority(1)) and assigning the priority(2) to Z.

3. X and Z have the same level of priority. This implies to modify the priority assigned to one of the two descriptions. Unfortunately, changing such priority invalidates one on the priorities relations defined in steps 1 and 2.

Hence, the three priority relations we reasoned about cannot be modeled at the same time. This is a limitation in the sense that a designer could not express such relations.

This problem can be solved by adding a property to the pre-conditions of feature descriptions: `priority(Features)`, where the variable *Features* is a set of feature description names containing the names of the feature descriptions, over which the feature description considered has priority.

Such a solution implies to modify incoherences rules such that $|\mathcal{D}_{X,m} \bullet \mathcal{D}_{Y,n}|$ may present an incoherence if, and only if the priorities of their pre-conditions present an incoherence:

- Both descriptions have priority over each other; $\mathcal{D}_{X,m}$ has priority over $\mathcal{D}_{Y,n}$ and $\mathcal{D}_{Y,n}$ has priority over $\mathcal{D}_{X,m}$.
- None of the description has priority over the other one; $\mathcal{D}_{X,m}$ does not have priority over $\mathcal{D}_{Y,n}$ and $\mathcal{D}_{Y,n}$ does not have priority over $\mathcal{D}_{X,m}$.

4.4 In Summary

We have presented six rules that allow to identify incoherences between features at the requirements level and thus to identify feature interactions that can be present in a specification corresponding to the requirements. The rules are formally defined. As mentioned in section 4.2.1, variables are bound to users. The use of a finite number of values (users) insures a finite number of possible binding combinations for variables. Hence, the identification of incoherences is a decidable problem. Thus, such identification can be automated by the use of a logic predicate language. Chapter 5 presents our implementation of these rules as a tool for automatic filtering of incoherences and feature interaction detection.

Chapter 5

A Tool for Feature Interaction Detection

Test scenarios can be used for partial validation of systems, in the sense that whether a system can or cannot execute a scenario can reveal or exclude the existence of a particular error (although not necessarily all). The method proposed for filtering incoherences considers pairs of features. Given two features presenting an incoherence, the corresponding interaction can be detected in the specification using validation scenarios. The rules defined in chapter 4 allow manual identification of incoherences as well as manual test derivation to test a specification. However, automatic incoherences identification and automatic test suite generation for corresponding feature interactions detection are preferred.

Based on the rules presented in chapter 4, a tool for the automatic filtering of incoherences and detection of feature interactions is proposed. Given feature representations, the tool analyzes incoherences and generates reports. Based on these reports, validation test suites can be automatically obtained, thus, the existence of the corresponding interactions in the specification can be verified. The tool we propose is an application implementing two behaviors: the filtering of incoherences and the derivation of test suites, corresponding to the incoherences found for feature interaction detection.

In this chapter, we first present the scenario based validation of the LOTOS specification as it was manually done at the beginning of the project (section 5.1). Then, we present the tool . We first present the implementation of the filtering process (section 5.2). Then, we present the automatic derivation of test suites and explain what needs to be provided for such generation (section 5.3).

5.1 Scenario Based Validation

LOTOS allows many types of sophisticated validation procedures. However, some of these can be impractical or time consuming on large specifications. We propose a pragmatic approach, that can be quickly adopted in an industrial environment. For this purpose, validation of the LOTOS specification is based on scenarios. Test cases relative to the behavior of the system are produced and applied against the specification. The results of the scenarios are interpreted and allow the user to determine whether the model is correct or not, with respect to the scenarios applied against the specification. This section gives an

overview of the manual validation performed on the specification.

5.1.1 Testing Strategy

In this method, scenarios are manually extracted from the UCM model. Note that the latter is a scenario-based model to start with. This allows the tester to derive scenarios by simply following possible paths in the model. This works well for the derivation of basic system and individual feature scenarios because these are based on the structure and syntax of the specification.

The sequence of activities that composes a given path is extracted from the UCM. That sequence is translated into a LOTOS process. The resulting scenario can be used for testing the LOTOS specification of the system. The purpose of scenarios derived from the UCM model is twofold: to validate the LOTOS specification against the UCM requirements and to test the system. Furthermore, scenarios are derived with three types of tests in mind:

Basic system properties – consist in testing the basic properties that the system must possess (the basic call). The scenarios involve simple test cases such as calling another party which is idle and calling another party which is busy.

Individual features properties – focus on testing features taken individually. Each feature needs to be working properly by itself before being integrated with other features for detecting interactions.

Feature interactions – focus on the detection of *interfering* feature interactions. These test scenarios aim at revealing problems that a given combination of features might inflict upon users.

The derivation of scenarios for feature interaction testing is less direct than for basic system and individual feature properties. Feature interaction involves the properties of the features and their behavior relating to each other and to the system. Some features can interact with many while others can interact only with a few. The strategy used consists in analyzing the features and targeting those pairs of features that have the highest probability of interaction. This analysis requires a good knowledge of the system and of its features.

5.1.2 Testing Using LOLA

The testing strategy consists in deriving scenarios from the UCM model, then applying them to the specification. A test suite composed of twenty scenarios testing the basic and individual properties of the system and its features were manually produced. The pairs of features having the highest likelihood of showing interactions were identified.

Test scenarios are specified in terms of LOTOS behaviors and applied against the LOTOS specification of the system by means of tools such as LOLA *(LOtos LAboratory, Universidad Politécnica de Madrid)*. Under LOLA, test scenarios are expected to exist as individual LOTOS processes within the LOTOS specification. This involves embedding the LOTOS behavior of each test scenario into a process structure. Provided with a targeted *success* event, LOLA synchronizes the given test process with the specification and attempts to follow all possible execution paths. The resulting verdict is then one of the following:

- MUST PASS – when all execution paths lead to the targeted *success* event
- REJECT – when the targeted *success* event is never reached
- MAY PASS – otherwise.

LOLA may also be instructed on the kind of exploration to perform:

Full Exploration – consists in exploring all the possible paths that can be reached when synchronizing with a test scenario. All cases conforming to those of the test scenario are analyzed. The execution time depends on the number of possible cases.

Partial Exploration with Random Walk – consists of a partial exploration of paths. The paths executed are selected using heuristics [35]. This is useful with tests that would require long execution times because heuristics provide much faster results (few seconds to few minutes). Even if the coverage is not 100% complete, experience shows that heuristic testing detects most of the specification faults quickly.

Unique Exploration – consists in visiting a single successful execution path. This is a convenient way to import of export scenarios to other tools. Such a trace, corresponding to *one* possible path, can be converted from or into a Message Sequence Charts (MSC) that can be used by other tools like SDT.

5.1.3 Feature Interaction Testing

Feature interaction detection is an important goal of the testing process. But, as mentioned, before testing for interactions between features, it is essential to test the basic properties of the system as well as the individual properties of the features. Once the system has been validated, specific FI tests derived from the UCM model are applied against the LOTOS specification.

All test scenarios are applied against the LOTOS specification using LOLA. Most of the scenarios are analyzed within a few seconds. Some scenarios can take days due to the number of states generated.

Test scenarios can be executed in such a way that LOLA displays intermediate results every ***n*** states (where ***n*** is an arbitrary number specified by the tester). This enables the tester to monitor the running test process.

The execution of scenarios can be driven by a script to automatically run them, possibly concurrently on different machines.

As an example, let us consider the following scenario: the user **B** is registered to *Call Forward on Busy* and forwards the calls to user **C**. User **A** is registered to *Originating Call Screening* and screens calls from user **C**. Given that **B** is busy and that **A** calls **B**, then **A** should be forwarded to **C** (because of **B**'s **CFB**) but it should also get a dial-Tone (because of **A**'s **OCS**). The corresponding LOTOS process is then coded as follows:

```
process ScenarioFI_OCS_CF_2r2 [ USER_to_DE, DE_to_USER,
                                 Init, scenarioFI_OCS_CF_2r2 ]: noexit:=
  let specificDB:Database =
```

```
        UF(userB, FD(cfb, fArg(2003, noArg), endFeatureSet),
        UF(2002, FD(cfb, fArg(2003, noArg), endFeatureSet),
        UF(userA, FD(ocs, fArg(2003, noArg), endFeatureSet),
        UF(2001, FD(ocs, fArg(2003, noArg), endFeatureSet), emptyDB)))) in (
    Init !specificDB;

  USER_to_DE !f1 !userB !debB0 !offHook;          (* B goes offHook *)
  DE_to_USER !f2 !debB0 !userB !dialTone;         (* B gets dialTone *)
  USER_to_DE !f1 !userA !debA0 !offHook;          (* A goes offHook *)
  DE_to_USER !f2 !debA0 !userA !dialTone;         (* A gets dialTone *)
  USER_to_DE !f0 !userA !debA0 !dial !2002;       (* A dials B *)
  DE_to_USER !f2 !debA0 !userA !callInProgress;   (* A gets callInProgress *)
  DE_to_USER !f2 !debA0 !userA !dialTone;         (* A gets dialTone *)
  USER_to_DE !f1 !userA !debA0 !onHook;           (* A goes onHook *)
  DE_to_USER !f2 !debA0 !userA !toneOff;          (* A gets toneOff *)

  scenarioFI_OCS_CF_2r2; stop                     (* success *)
)
endproc
```

The output of the execution, with LOLA, of this test scenario on the specification is:

```
Analyzed states = 863
Generated transitions = 1121
Duplicated states = 0
Deadlocks = 0

Process Test = scenariofi_ocs_cf_2r2
Test result = MUST PASS.

successes = 259
stops = 0
exits = 0
cuts by depth = 0
```

This output provides information about the number of states of the graph generated, the number of generated transitions, and the number of successes. The number of successes equals the number of paths executed when the verdict is MUST PASS.

The first validation of the LOTOS specification has been done with scenarios that were derived by hand from UCM. This can lead to different problems. Not enough scenarios are produced or some important cases are missing. The validation is not complete. On the other hand, too many scenarios can be produced, which can lead to redundancies. Moreover, deriving all possible test cases (brute force) is not always a good choice, considering the processing time and the state explosion problem.

A solution is provided by our tool. Incoherences can be identified at the requirements level by using the filtering rules proposed in chapter 4. This provides useful information for

testing the system. Later, this information is used to derive test scenarios in order to validate the specification. Thus, the specification is derived with knowledge of feature combinations presenting incoherences and, later, test scenarios allow to insure that these combinations do not lead to problems and that the specification is free of interactions.

5.2 Automatic Incoherences Filtering & FI Detection

This section presents the implementation of our filtering method for the automatic identification of incoherences. Given a set of feature descriptions, it is possible to identify incoherences manually by applying each rule to each possible pair. Thus, for n feature descriptions, n^2 pairs must be considered. Feature descriptions and rules are formally defined and based on sets of properties. Given a specific rule, the strategy followed for the identification of an incoherence is to bind the variables of properties to a proper combination of users such that the two features satisfy the rule. Such a task can require a significant amount of time, even for a small number of features.

Such a task can be automated by the use of a logic programming language. Representing rules and feature descriptions as logic predicates, the identification of an incoherence can be conducted using a logic programming interpreter. Such an interpreter tries to satisfy the predicates, and thus, tries to bind variables to all combinations until a combination that satisfies the predicates is found or all combinations of users have been tried. Logic programming languages such as Prolog [36, 37] or others can be used. Prolog is the language chosen. It has been developed at the University of Marseilles (France) [36] as a practical tool for programming in logic. The Prolog interpreter chosen is SWI-Prolog [41].

5.2.1 Implementation of Feature Descriptions & Rules into Prolog

Rules are translated into Prolog using the following mapping:

- Variables remain variables.
- Properties are represented as facts. Their syntax is the same as the one used for facts in Prolog.
- Sets of properties are represented as ordered lists of facts.
- A feature description is represented as a predicate, *feature*. The head of the predicate contains four lists: the name, pre-conditions, triggering events and results. The body contains the constraints, stated as relations between variables.

Let us consider the definition of Call Forward on Busy:

- $\mathcal{D}_{cfb,1}$ *Call Forward on Busy, part 1*:
  $\mathcal{P}_{cfb,1}$ = { subs(B, cfb), concerns(B, cfb), cfb(C), busy(B) }
  $\mathcal{T}_{cfb,1}$ = { call(A, B) }
  $\mathcal{R}_{cfb,1}$ = { redirected(A, C), call(A, C) }
  $\mathcal{C}_{cfb,1}$ = { $(A \neq B)$, $(A \neq C)$, $(B \neq C)$ }

The translation into Prolog is straightforward. The signature of the predicate *feature* is `feature(V, W, X, Y) :- Z.` Where variables V, W, X, Y respectively represent the lists containing name, pre-conditions, triggering events and results, and Z represents the body containing the restrictions.

```
feature([cfb, 1],
        [subs(B, cfb), concerns(B, cfb), cfb(C), busy(B)],
        [call(A, B)],
        [redirected(A, C), call(A, C)]
    ):-
        A \= B, A \= C, B \= C.
```

In order to be able to bind variables to proper values (i.e. user names), users must be defined and the fact that variables must be bound to users must be specified explicitly. If not, the tool will still work but the Prolog interpreter will bind variables to internal default values. These are usually composed of an underscore, followed by a letter and numbers such as `_A340`. Even if still valid, such values can be difficult to interpret for the designer. It is easier to read `call(alice, bob)` than `call(_B747, _A340)`.

Users can be defined using a fact *user* containing the name of a user. Thus, defining four users for the system is done by defining four *user* facts, each one representing a different user: `user(alice).`, `user(bob).`, `user(carol).`, `user(dave).` In Prolog, the body of a feature definition contains a call to the *user* facts in order to bind variables to users. The complete definition of a feature description is then:

```
feature([cfb, 1],
        [subs(B, cfb), concerns(B, cfb), cfb(C), busy(B)],
        [call(A, B)],
        [redirected(A, C), call(A, C)]
    ):-
        user(A), user(B), user(C),
        A \= B, A \= C, B \= C.
```

When the predicate is called, Prolog binds variables A, B and C to different values taken from the *user* fact in such a way that relations between variables (constraints) are satisfied.

Contradictions between properties must be also clearly identified. If they are not, the tool will still work but some incoherences will not be identified. We use the predicate *contradiction_pair* to state a contradiction between two properties (facts). An example of a contradiction pair is that a user X cannot be busy and idle at the same time. This is expressed by the following Prolog fact:

```
contradiction_pair(busy(X), idle(X)).
```

A contradiction exists between two sets if it is possible to find two elements, one from each set, presenting a contradiction. The predicate *contradiction* is used to define possible contradictions between two facts. The code of the *contradiction* predicate follows:

```
contradiction(Set1, Set2):-
```

```
        member(X, Set1),
        member(Y, Set2),
        (contradiction_pair(X, Y) ; contradiction_pair(Y, X)), !.
```

As for feature descriptions, rules are translated into Prolog. Rules are implemented in the form of a predicate called *fi_check*. The signature of the predicate is the following: `fi_check(R, [F1, Fx], [F2, Fy], PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2).`

The variables of this predicate respectively denote: the name of the rule, the names (split in two parts) of the two features to be considered, the pre-conditions, triggering events and results of the first feature, and the pre-conditions, triggering events and results of the second one. The Prolog definition of rule #d1, given in chapter 4, section 4.2.2, is the following:

```
fi_check(d1, [F1, Fx], [F2, Fy], PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2):-
        feature([F1, Fx], PcnF1, TrgF1, ResF1), !,
        feature([F2, Fy], PcnF2, TrgF2, ResF2),

        member(subs(U, F1), PcnF1),
        member(subs(U, F2), PcnF2),
        not(contradiction(PcnF1, PcnF2)),
        TrgF1 = TrgF2,
        ResF1 \= ResF2,
        not(contradiction(ResF1, ResF2)), !.
```

Prolog always tries to satisfy a predicate. It contains a mechanism that allows the self-binding of variables in order to try to satisfy predicates and facts. For instance, if a query composed the fact *user* is called with a fixed value: `user(alice)`, the Prolog interpreter simply checks if a definition exists for such a fact and returns `true` or `false` whether such a fact exists or not. If the fact is called with a variable instead of a given value, the Prolog interpreter looks into the list of users and returns the next value. If this value does not satisfy the predicate, Prolog fetches the next one. And so on until the predicate is satisfied or no more values are available. Three kinds of queries can be made with the predicate *fi_check*:

- Variables are bound to values: the interpreter simply checks if the predicates listed in *fi_check* are satisfied on those values and returns `true` or `false` depending on whether the predicates are satisfied or not.

- Variables are unbound: the interpreter tries to bind variables to values that satisfy the predicate.

- Variables are partially bound: the interpreter tries to bind variables not bound to values that satisfy the predicate with respect to the values already given.

Thus, given two feature descriptions, a single query to one of the rules allows to find all incoherences that can be identified by this rule. Section 5.2.2 describes the application of the tool with respect to the identification of incoherences from the given descriptions.

5.2.2 Incoherences Filtering

Feature descriptions are contained into a text file. When the tool is invoked, the name of this file is given as a command line argument. The tool then elaborates all possible pairwise combinations and analyzes all pairs with each of the six rules to check whether an incoherence is present or not. In order not to report the same incoherence twice, the tool builds an internal database of incoherences found. In addition, each time an incoherence is found, a report is generated. The report contains a detailed description of the rule used and the incoherence found, and contains the incoherence itself.

An incoherence is represented by a fact called `fi` containing the description of the features involved. This fact is composed of nine parameters (1 to 9): the rule used (1), names of the features (2,3) and the pre-conditions (4,7), triggering events (5,8), and results (6,9) of both features. As an example, incoherence between $\mathcal{D}_{cfb,1}$ and $\mathcal{D}_{ics,1}$ (presented in chapter 4, section 4.2.2) follows. The predicate has been reformatted to improve its readability. Text parts preceded by a % are comments.

```
fi(
    d1,                                                          % (1)
    [cfb, 1],                                                    % (2)
    [ics, 1],                                                    % (3)
    [subs(bob, cfb), concerns(bob, cfb), cfb(carol), busy(bob)], % (4)
    [call(alice, bob)],                                          % (5)
    [redirected(alice, carol), call(alice, carol)],              % (6)
    [subs(bob, ics), concerns(bob, ics), ics_list(alice)],       % (7)
    [call(alice, bob)],                                          % (8)
    [deny_call(bob, alice), call_denied(bob, alice)]             % (9)
).
```

Each report is composed of two parts: detailed information about the incoherence and the incoherence itself (`fi` fact). Detailed information is presented in plain English in a format tailored to the rule concerned. It contains the name of the rule, a brief description of the type of incoherence, as well as the complete information about the incoherence.

The first part, presenting the information about the incoherence, is written as Prolog comments. The second part is written as a Prolog `fi` fact, similar to the formatted one previously presented. Hence, a file containing reports is a Prolog database containing already known incoherences recorded as `fi` facts as well as user's readable information written as comments, thus ignored by the interpreter.

As an example, the report obtained for the incoherence between $\mathcal{D}_{cfb,1}$ and $\mathcal{D}_{ics,1}$ identified by rule #d1 follows. The report is shown in its initial form, with comments describing the incoherence. The `fi` fact following the description, already presented in reformatted form, is not shown here.

```
%-------------------------------------------------------------------------
% * Rule #d1 -> [cfb, 1] & [ics, 1]
%   Same user subscribed to two features having the
%   same triggering events and different results
```

```
%      + Features pre-conditions

%         - Pre-conditions of [cfb, 1]
%             subs(bob, cfb)
%             concerns(bob, cfb)
%             cfb(carol)
%             busy(bob)

%         - Pre-conditions of [ics, 1]
%             subs(bob, ics)
%             concerns(bob, ics)
%             ics_list(alice)

%      + Same triggering events

%          call(alice, bob)

%      + Different results

%         - Resulting events of [cfb, 1]
%             redirected(alice, carol)
%             call(alice, carol)
%             ring(carol)

%         - Resulting events of [ics, 1]
%             deny_call(bob, alice)
%             call_denied(bob, alice)
%-------------------------------------------------------------------------
```

5.2.3 Filtering Algorithm

Considering $\mathcal{S}$, the set of all features of a specification, the list of all possible pair-wise combinations is obtained by a binary relation (*one to one* pairs) on $\mathcal{S} \times \mathcal{S}$. Let us consider a set of new features $\mathcal{T}$, and the new set of all features $\mathcal{U} = \mathcal{S} \cup \mathcal{T}$. The list of all combinations are obtained by $\mathcal{U}$, basically, $\mathcal{S} \cup \mathcal{T} \times \mathcal{S} \cup \mathcal{T}$, which means that pair-wise combinations already analyzed have to be considered again, while they need not be since their results are already known. In order to avoid that, our tool implements two solutions:

The first solution consists in giving the set of already known incoherences such that for a given rule, only pairs not known as presenting an incoherence are considered. As we previously mentioned, the data in the report is a database of known incoherences. Thus, the set of already known incoherences can be obtained by loading a previous report into the tool. However, this only gives information about known incoherences. Other pairs, already analyzed but not presenting any incoherence are processed again.

The second solution is to consider two different sets $\mathcal{S}$ and $\mathcal{T}$ and to determine the pairs to be considered using a relation on $\mathcal{S} \times \mathcal{T}$. To distinguish them, we respectively call them

left hand set and *right hand set.* For instance, if a new feature description $\mathcal{D}_{X,m}$ is added, one can specify two sets such that $\mathcal{S}$ contains all features of the old specification and $\mathcal{T}$ contains the newly added feature. Then, $\mathcal{S} \times \mathcal{T}$ determines all pairs containing the new feature. Moreover, all pairs previously analyzed may be ignored. *Left hand set* and *right hand set* features are respectively represented using lists of Prolog facts called `lfeature` and `rfeature`.

Considering 3 feature descriptions $\mathcal{D}_{cfb,1}$, $\mathcal{D}_{ics,1}$, $\mathcal{D}_{ocs,1}$ already analyzed, and a new feature description $\mathcal{D}_{cw,1}$, the sets $\mathcal{S} = \{\mathcal{D}_{cfb,1}, \mathcal{D}_{ics,1}, \mathcal{D}_{ocs,1}, \mathcal{D}_{cw,1}\}$ and $\mathcal{T} = \{\mathcal{D}_{cw,1}\}$ are represented using the following facts:

```
lfeature([cfb, 1]).
lfeature([ics, 1]).
lfeature([ocs, 1]).
lfeature([cw, 1]).

rfeature([cw, 1])
```

And, $\mathcal{S} \times \mathcal{T} = \{|\mathcal{D}_{cw,1} \bullet \mathcal{D}_{cfb,1}|, |\mathcal{D}_{cw,1} \bullet \mathcal{D}_{ocs,1}|, |\mathcal{D}_{cw,1} \bullet \mathcal{D}_{ics,1}|, |\mathcal{D}_{cw,1} \bullet \mathcal{D}_{cw,1}|\}$ is the set of new combinations that need to be analyzed by the tool.

The first solution implies that only already known interacting pairs are not considered; all other pairs, even those already analyzed, are considered again. The second solution avoids this by directly targeting new pairs.

However, depending on the size of the specification, reusing already known incoherences, automatically obtained from reports, can take less time than building and managing the two sets of features to be considered. The availability of the two solutions allows the user to choose the best one depending on the number of descriptions to analyze.

The invocation of the tool must be done with at least one command line argument: the name of the file containing the specification. Three options can be specified:

`-g <file>` - Name of the file containing *right hand* and *left hand* sets.

`-k <file>` - Name of the file containing the list of already known incoherences.

`-i <file>` - Name of the file to which the report about incoherences are written.

When invoked, the interpreter loads the needed files. If the options `-k` is not used, no file containing the right and left hand sets is given. With respect to the solution previously described, the tool then builds the two sets such that both of them contains all the descriptions, thus, all combinations are analyzed. Then, the tool looks for incoherences by applying each rule to each pair and generates a report for each incoherence found. The complete description of the procedure follows:

1. Load the features description file.
2. If a file containing left and right hand sets is given in the command line, load it. Else, build sets such that each one contains all features.
3. If a file containing already known incoherences is given in the command line, load it.
4. For each existing rule
   (a) For each pair in $\mathcal{S} \times \mathcal{T}$

i. Try all possible values for variables to satisfy the rule

ii. If an incoherence is found, update the internal database and generate a report. Else do nothing.

Prolog tries to satisfy predicates thus insuring that all possible values are tried. Finding a proper combination of users is part of the task automatically managed by the interpreter. Pair-wise combinations can also be automatically built by a single call to a predicate. The algorithm is split in two symmetric parts: one concerning pairs (s, t) and the other concerning pairs (t, s) where $s \in \mathcal{S}$ and $t \in \mathcal{T}$. The Prolog implementation of part 4 of the algorithm is given below:

```
% default predicate
fi_search_left_to_right:-
        rule(R),          % rule
        lfeature(F1),     % feature description from left hand set
        rfeature(F2),     % feature description from right hand set

        % no FI has been found yet for this combination with rule R
        not(fi(R, F1, F2, _, _, _, _, _, _)),

        % try to find a possible FI
        fi_check(R, F1, F2, PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2),

        % add FI to the internal database
        assert(fi(R, F1, F2, PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2)),

        % generate report corresponding to the FI
        fi_report(R, F1, F2, PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2),

        % loop (recursive call)
        fi_search_left_to_right, !.

% predicate used if default one can not be satisfied (FI already found)
fi_search_left_to_right:-
        rule(R),
        lfeature(F1), rfeature(F2),
        (fi(R, F1, F2, _, _, _, _, _, _)), !.
```

5.3 Automatic Test Suite Generation

Incoherences identification is used as a filtering method. Since the information about features is obtained from the requirements, identifying an incoherence does not imply that the corresponding interaction exists in the specification. Nevertheless, a designer being aware of such an incoherence can take extra care in designing the system. Hence, this can lead to the avoidance of several interactions at the design stage.

The information provided by the reports can be used to derive validation test suites in order to verify that interactions corresponding to incoherences are not present in the specification. However, depending on the number of incoherences, manual test suite derivation can be time consuming. Considering that the specification follows the requirements, establishing mapping rules from properties (used to formalize feature descriptions) to scenario parts enables the automatic generation of scenarios, and thus, test suites.

5.3.1 From Properties to Scenario Parts

Properties used to describe a feature illustrate the state in which the feature must be, the event(s) triggering the feature and the results produced by the activation of this feature. Properties are obtained from the requirements. The LOTOS specification is also derived from the requirements. Depending on the language used, features are represented differently and thus, validation scenarios use different representations such as Message Sequence Charts in SDL or actions in LOTOS.

Given a LOTOS specification involving a pair of features for which an incoherence was identified according to the method described in section 5.2, a validation scenario is a sequence of actions, representing the behavior of a feature interaction. Each action represents a part of the behavior, such as the establishment of a call or a dial-tone. As properties describe the features, actions of a test scenario correspond to actions of the behavior described in requirements. It is possible to establish a direct mapping between properties and scenario parts. Then, given an incoherence, a scenario can be automatically derived by following the model presented in section 5.2.3 and using the mappings between properties and scenario parts.

Establishing mapping rules to go from a feature description to a LOTOS test scenario consists in identifying the LOTOS action(s) corresponding to each property of the set of feature descriptions. The mapping is done by associating each property to an action of the specification. To generate a test scenario, the algorithm uses five predicates for which signatures (names and parameters) are pre-defined. The body of the predicates needs to be programmed by the tester to implement scenario generation. The predicates are respectively:

test_header(*[F1, PndF1], [F2, PndF2], Result, R***)** – This predicate builds the header of the test scenario. It expects the names of two features *F1* and *F2* and the lists of their pre-conditions: *PndF1* and *PndF2*. In addition, the predicate also expects the name of the feature in which the scenario results, *Result*, and the rule used, *R*.

test_pre_conditions(*List***)** – This predicate is translates a list of pre-conditions into a list of scenario parts (e.g. actions, in LOTOS). It expects a list of pre-conditions, decomposes it into single elements (properties) and then calls a predicate corresponding to each of them.

test_triggering_events(*List***)** – This predicate translates a list of triggering events into a list of scenario parts (e.g. actions, in LOTOS). It expects a list of triggering events, decomposes it into single elements (properties) and then calls a predicate corresponding to each one of them.

test_results(*List*) – This predicate translates a list of results into a list of scenario parts (e.g. actions, in LOTOS). It expects a list of results, decomposes it into single elements (properties) and then calls a predicate corresponding to each one of them.

test_footer(*List*) – This predicate generates the end of the test scenario. The end of the scenario does not contain any specific information. The parameter list is for future use.

The test generation algorithm is based on the five predicates presented above. The tester must provide the internal code of each one of them such that they can be used by the interpreter. The body of these predicates is defined depending on what needs to be provided for test generation. Additional predicates can be defined, such as those needed to map users to phone numbers.

A property may have different meanings depending on the set (pre-conditions, triggering events or results) they belong to. As an example, let us consider Call Forward, defined in chapter 4, section 4.2.2. The property `call` composed of two variables is present in both triggering events and results. However, the meaning of a call is not the same in the triggering events where it represents an incoming call than in the results where it represents an outgoing forwarded call. It is more convenient to use different sub-predicates to map properties from pre-conditions, triggering events and results. Pre-conditions are mapped using a predicate called *pre-condition*, triggering events are mapped using a predicate called *triggering_event* and results are mapped using a predicate *results*. Such a partitioning allows specify different mappings for the same property, depending on the set it belongs to in the feature description.

Pre-conditions, triggering events and results describe part of the behavior and are considered as ordered sets. Elements composing a results must appear in a certain order, defined in requirements. The use of Prolog lists for feature description insures that the order between properties is respected in the scenario generation algorithm. This gives the possibility to the testers and designers to insure that events occur in a specific order.

5.3.2 Test Suite Principles

Feature interaction validation of a specification assumes that the basic system properties have been validated and the features work properly in isolation. Nevertheless, the tool offers the automatic generation of single scenarios for single features validation. The behavior of a feature is contained in the feature description, thus, using the predicates presented section 5.3.1, a given feature description can be mapped into a single scenario corresponding to its behavior.

Applied against the specification, the scenario can be successful, meaning that the basic behavior of the feature is working, or, rejected (or not always successful), meaning that the feature is not working properly in isolation. Even if single feature testing is not the main goal of the tool, this offers the possibility of automatically insuring that the basic behavior of the features has been correctly specified.

Feature interaction testing scenarios are more complex. An interaction, corresponding to an incoherence identified in the filtering process, can be characterized by non determinism, a contradiction (which can be also seen as a non determinism) or a loop. We define a general model corresponding to each of the three possible kinds of interactions. The models defined

are then used for the derivation of test scenarios and analysis of the results following their application on the specification using LOLA.

Non-determinism – Non-determinism implies that the result is ambiguous. The tool is not able to know which result is intended by the designer. Thus, interaction detection is done using a test suite composed of two scenarios. Both scenarios model a case where both features are present. The first scenario illustrates the case where the results of the first feature occur and the second one illustrates the case where the results of the second feature occur. The result of the application of both scenarios with LOLA is interpreted in the following manner:

- Both scenarios are rejected (REJECT or MAY PASS). This definitely indicates a problem. It tells the tester that in any case, neither feature behaves properly.
- Both scenarios are successful (MUST PASS) means that both results can occur. This indicates a problem in the specification because only one feature should be activated. However, depending on the features, this result can be acceptable. It is up to the tester to insure that the behavior is the one intended.
- One scenario is successful (MUST PASS) and the other is rejected (REJECT or MAY PASS). This indicates that, at least, no non-determinism exists. However, the behavior may not be the one intended. If the tester expects the rejected scenario to be successful and vice-versa, then a problem is present. As for the previous case, it is up to the tester to insure that the behavior is the one intended.

Contradictions – A Contradiction also presents non-determinism. The results of the two involved features are in contradiction, and, as for non-determinism, the result is not known. Thus, scenarios for this kind of interaction follow the same model as the one followed by the non-deterministic test suites.

Loops – The model followed by loop scenarios is different. A loop can either occur or not occur. Then, only one scenario is needed to detect such interaction. The scenario illustrates the case where the loop is occurring. Thus, rejection of the scenario indicates that no loop occurs. If the scenario is successful, then it can be concluded that the loop effectively occurs in the specification.

A tester having knowledge of the models used can interpret the results of the application of scenarios and give a verdict. Explanations concerning the derivation of scenarios follows.

5.3.3 Scenarios Generation

This section presents the implementation of scenario generation. It includes single feature validation scenarios and feature interaction validation scenarios. Examples, based on the Mitel specification, are given in chapter 6.

Single Feature Test Scenarios Generation

The goal of single feature test scenario generation is to translate the behavior contained in a feature description into a scenario. This is implemented as a predicate calling the five mandatory predicates previously defined in section 5.3.1. The Prolog predicate used is:

```
isolation_test(F1, PcnF1, TrgF1, ResF1):-
        test_header([F1, PcnF1]),
        test_pre_conditions(PcnF1, []),
        test_triggers(TrgF1),
        test_results(ResF1),
        test_footer.
```

Since only one feature is considered, a different predicate `test_header` is used. In addition, the predicate `test_pre_conditions` are called with an empty list as second argument instead of the pre-conditions of the second feature.

Feature Interaction Test Suites Generation

As explained in section 5.3.2, we identified three different kinds of interactions. Non-deterministic and contradictory interactions can be tested based on the non-determinism they present. They both need two scenarios per interaction. However, the kind of rule (direct incoherence rules or transitive incoherence rules) corresponding to them must be taken in account. Interactions presenting a loop only need one scenario.

All direct incoherence rules (#d1, #d2, #d3, #d4) identify non-determinism or contradictions. The corresponding validation test suites can be derived using the same model. Rule #t1 identifies contradictions but is a transitive rule, thus, a different model must be used. Rule #t2 identifies loops and, thus, also needs to use a different model. Hence, we distinguish three kind of test suites:

Direct test suites – Concerning rules #d1, #d2, #d3 and #d4, these test suites are composed of two test scenarios, respectively presenting the results of the first and the second feature. The predicate used is the following:

```
fi_test(R, F1, F2, PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2):-
        ((R = d1); (R = d2);(R = d3); (R = d4)),
        test_header([F1, PcnF1], [F2, PcnF2], F1, R),   % F1
        test_pre_conditions(PcnF1, PcnF2),
        test_triggers(TrgF1),
        test_results(ResF1),
        test_footer,

        test_header([F1, PcnF1], [F2, PcnF2], F2, R),   % F2
        test_pre_conditions(PcnF1, PcnF2),
        test_triggers(TrgF2),
        test_results(ResF2),
        test_footer.
```

Transitive test suites – Concerning rule #t1, these test suites are also composed of two test scenarios. One illustrates the case where the scenario stops after the first feature, and thus, the second one is not activated. The other illustrates the case where the second feature is activated. The Prolog predicate is:

```
fi_test(t1, F1, F2, PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2):-
        test_header([F1, PcnF1], [F2, PcnF2], F1, t1),   % F1
        test_pre_conditions(PcnF1, PcnF2),
        test_triggers(TrgF1),
        test_results(ResF1),
        test_footer,

        test_header([F1, PcnF1], [F2, PcnF2], F2, t1),   % F2
        test_pre_conditions(PcnF1, PcnF2),
        test_triggers(TrgF1),
        test_results(TrgF2),
        test_results(ResF2),
        test_footer.
```

Loop test suites – Concerning rule #t2, these test suites are composed of only one test scenario. The test scenario illustrates the activation of the first feature, followed by the activation of the second one, followed, again, by the activation of the first feature. The corresponding predicate is:

```
fi_test(t2, F1, F2, PcnF1, TrgF1, ResF1, PcnF2, _, ResF2):-
        test_header([F1, PcnF1], [F2, PcnF2], ['a', 'loop'], t2),
        test_pre_conditions(PcnF1, PcnF2),
        test_triggers(TrgF1),
        test_results(ResF1),
        test_results(ResF2),
        test_results(ResF1),
        test_footer.
```

The production of validation test suites needs the file containing the feature descriptions and the file containing the mapping rules. A test suite is produced by calling the `fi_test` predicate relative to the rule identifying the incoherence corresponding to the interaction. Known incoherences can be optionally specified. In such a case, the corresponding test suites are generated; else, the filtering algorithm is applied and the corresponding test suites are generated. Four possible invocations are possible:

- Incoherences filtering,
- Test suite derivation for features in isolation,
- Test suite derivation for feature interactions,
- Both filtering and test suite generation combined together.

The complete Prolog implementation can be found in appendix B. The tasks are chosen at invocation, by the use of arguments. If invoked with no arguments (or wrong arguments) the tool prints its usage, as follows:

```
Usage: fi-lookup <command>
    * where command must be one of the following:

    -filtering <features> [options]
    --> find incoherences
        * where options include:
        -g <file> --- feature sets to combine
        -k <file> --- list of already known incoherences
        -i <file> --- file to which reports are written

    -testiso <features> <mapping-rules> [options]
    --> produces test suites for features in isolation
        * where options include:
        -t <file> --- file to which test suites are written

    -testfi <mapping-rules> <known incoherences> [options]
    --> produces test suites for given incoherences
        * where options include:
        -t <file> --- file to which test suites are written

    -filttest <features> <mapping-rules> [options]
    --> finds incoherences & produces corresponding test suites
        * where options include:
        -g <file> --- feature sets to combine
        -k <file> --- list of already known incoherences
        -i <file> --- file to which reports are written
        -t <file> --- file to which test suites are written
```

5.4 In Summary

We showed that the automation of incoherences filtering and test suite generation for feature interaction detection is possible using a logic programming language. The complete Prolog code of the tool is 850 lines long (commented). An evaluation, based on the Mitel specification, and on the Feature Interaction Workshop 2000 [4] is presented in chapter 6 with detailed examples of incoherences filtering. Test suite generation is also presented.

Chapter 6

Application & Evaluation

Our incoherences identification approach is based on a filtering process followed by a validation process that uses automatic generation of validation scenarios. This chapter presents the results obtained from the application of our method on two different specifications. The algorithmic complexity of the analysis as well as the usefulness of the method are presented. In addition, a comparison with another approach is discussed.

6.1 Application

We applied our approach to two different case studies: Mitel specification and the Feature Interaction Workshop 2000 [4] specification [42]. The application on the Mitel specification consisted in filtering incoherences, deriving test suites for the corresponding interactions and using these suites to test the specification. The application on the Feature Interaction Workshop consisted in applying the filtering process to identify incoherences.

6.1.1 Case Study 1 – Mitel

As mentioned in chapter 2 and chapter 3, the approach applied to the Mitel project consists in expressing requirements with a Use Case Maps model, deriving a LOTOS specification corresponding to the UCM model and validating the LOTOS specification. Mitel Corp. provided us with the UCM model from which a LOTOS specification was derived. We first manually derived feature interaction test suites to validate the specification. Then, we built our tool and applied the filtering process and the derivation of test suites to automatically validate the LOTOS specification.

Part 1 – Filtering Process

Using Use Case Maps and additional documentation provided with the set of UCM by Mitel Corp., the features were modeled and our filtering method was applied to it. The features considered were: Outgoing Call Screening (OCS), Incoming Call Screening (ICS), Call Forward Always (CFA), Call Forward on Busy (CFB), Call Transfer (CT), Call Pickup (CP), Call Waiting (CW), Automatic ReCall (ARC) and Time Reminder (TMR). These features are presented in chapter 3, section 3.2.2.

Some of the features are represented using a single description while some others need up to three or four. Call Waiting is split in four parts. Call Transfer, and Automatic Recall are split in three parts. Time Reminder is split in two parts. The other features are represented using only one description. They add up to twenty feature descriptions. As an example of features definitions, let us consider Call Forward Always and Call Forward on Busy. Both features are defined using the following common properties:

- `call(A, B)` - Call attempt from user A to user B.
- `redirected(A, C)` - User A is forwarded to user C.
- `ring(A, C)` - User A rings user C.

Call Forward Always uses the additional properties:

- `subs(A, cfa)` - User A subscribes to Call Forward Always.
- `concerns(A, cfa)` - Call Forward Always concerns user A.
- `cfa(C)` - Calls are forwarded to user C.

And Call Forward on Busy uses the additional following ones:

- `subs(A, cfb)` - User A subscribes to Call Forward on Busy.
- `concerns(A, cfb)` - Call Forward on Busy concerns user A.
- `busy(A)` - User A is busy.
- `cfb(C)` - Calls are forwarded to C.

Each feature only implements one functionality, hence only one *feature* predicate is needed for each one of them. The Prolog implementation of features follows:

- Call Forward Always - A call from user A to user B is forwarded to the specified destination (user C) stated by the property $cfa(C)$.

```
feature([cfa, 1],
        [subs(B, cfa), concerns(B, cfa), cfa(C)],
        [call(A, B)],
        [redirected(A, C), call(A, C), ring(A, C)]
    ):-
        user(A),
        user(B), A \= B,
        user(C), C \= A, C \= B.
```

- Call Forward on Busy – When user B is busy, a call from user A to user B is forwarded to the specified destination (user C) stated by the property $cfb(C)$.

```
feature([cfb, 1],
        [subs(B, cfb), concerns(B, cfb), cfb(C), busy(B)],
        [call(A, B)],
        [redirected(A, C), call(A, C), ring(A, C)]
    ):-
        user(A),
        user(B), A \= B,
        user(C), C \= A, C \= B.
```

In addition, we consider that a user A making a call to a user B is in contradiction with the same user A making a call to a different user C. This contradiction is stated as:

- $\mathcal{K}$(call(A, B), call(A, C)) if and only if $B \neq C$.

This is implemented in Prolog using the predicate *contradiction_pair* in the following form:

- contradiction_pair(call(A, B), call(A, C)):- B \= C

Our filtering method identified a total of 43 incoherences. Eight of them concern incoherences involving Call Forward Always and Call Forward on Busy. Brief explanations, based on the reports given by the tool, are given here. The 43 reports can be found in appendix A.

Rule #d2 → [cfb, 1] & [cfa, 1] - Contradiction between Call Forward Always and Call Forward on Busy when some user subscribes to both features and the features forwards his calls to different users.

Rule #t1 → [cfb, 1] & [cfb, 1] - Contradiction between Call Forward on Busy and Call Forward on Busy when some user subscribes to Call Forward on Busy and forwards his calls to a user that also subscribes to Call Forward on Busy but forwards his calls to someone else.

Rule #t1 → [cfb, 1] & [cfa, 1] - Contradiction between Call Forward on Busy and Call Forward Always when some user subscribes to Call Forward on Busy and forwards his calls to a user that subscribes to Call Forward Always but forwards his calls to someone else.

Rule #t1 → [cfa, 1] & [cfb, 1] - Contradiction between Call Forward Always and Call Forward on Busy when some user subscribes to Call Forward Always and forwards his calls to a user that subscribes to Call Forward on Busy but forwards his calls to someone else.

Rule #t1 → [cfa, 1] & [cfa, 1] - Contradiction between Call Forward Always and Call Forward Always when some user subscribes to Call Forward Always and forwards his calls to a user that also subscribes to Call Forward Always but forwards his calls to someone else.

Rule #t2 → [cfb, 1] & [cfb, 1] - Loop between Call Forward on Busy and Call Forward on Busy when two users subscribes to Call Forward on Busy and each one forwards his calls to the other.

Rule #t2 → [cfb, 1] & [cfa, 1] - Loop between Call Forward on Busy and Call Forward Always when two users respectively subscribe to cfb and Call Forward Always, and each one forwards his calls to the other.

Rule #t2 → [cfa, 1] & [cfa, 1] - Loop between Call Forward Always and Call Forward Always when two users subscribe to Call Forward on Busy and each one forwards his calls to the other.

A detailed summary of the incoherences identified between all pairs of feature descriptions is presented in Table 6.1:

Incoherences									
	OCS	ICS	CFA	CFB	CT	CP	CW	ARC	TMR
OCS			1	1	2	1		1	
ICS	—		2	2	1	1	3	2	
CFA	—	—	2	4	1	1	3	2	
CFB	—	—	—	2	1		3	2	
CT	—	—	—	—			1		
CP	—	—	—	—	—				
CW	—	—	—	—	—	—	1	3	
ARC	—	—	—	—	—	—	—		
TMR	—	—	—	—	—	—	—	—	

Number of incoherences per rule and type						
Rule	#d1	#d2	#d3	#d4	#t1	#t2
Number of incoherences	13	3	4	1	19	3
Ratio per rule	30%	7%	9.5%	2.5%	44%	7%
Ratio per type	49%				51%	

Table 6.1: Mitel Case Study: Incoherence Filtering Results

Spaces indicate that the tool did not identify any incoherence. A distinction is made between direct and transitive incoherences. In addition, we make a distinction between direct incoherences concerning features subscribed by the same user and those concerning features subscribed by different users. Such distinctions allow to classify incoherences into three groups, presented below.

Rules d1 & d2 - identify direct incoherences present between two features having the same subscriber and presenting incoherent results.

Rules d3 & d4 - identify direct incoherences present between two features having different subscribers and presenting incoherent results.

Rules t1 & t2 - identify transitive incoherences present between two features where one triggers the other and the results are incoherent or lead to a loop.

The results show that most of the incoherences identified are either direct incoherences present between two features subscribed by the same user (rules #d1 and #d2) or transitive incoherences (rules #t1 and #t2). Only a few direct incoherences are identified for pairs of features subscribed by different users (rules #d3 and #d4).

We interpret the results as follows: incoherences identified by rules #d1, #d2, #t1 and #t2 are easier to foresee. Incoherences identified by rules #d3 and #d4 are much more complex and thus, can be more difficult to avoid. Hence, having only a few incoherences identified by rules #d3 and #d4, as it is the case here, indicates that the designers have to deal with simple incoherences and thus, refining requirements or specification to avoid incoherences takes less time and effort.

Part 2 – Test Suites Generation

As explained in chapter 5, the tool produces test suites composed of one or two scenarios. Scenarios are produced using five predicates: *test_header*, *test_pre_conditions*, *test_triggers*, *test_results* and *test_footer*.

These predicates must be coded by the tester and must contain the mapping rules needed for the scenarios derivation. We modeled the mapping rules using the five conventional predicates as well as some additional ones when needed:

- **test_header** – The LOTOS specification implements a database that allows dynamic specification of feature attributes (e.g. which user to screen, where to call forward). These attributes refer to the subscriber as well as the users concerned with the feature. This database must be initialized with proper values at the beginning of each test scenario. The `test_header` predicate is used to produce the LOTOS test header and to initialize the database.

  The initialization of the database is constructed using a sub-predicate `lotos_database` that builds the initialization instructions. This predicate calls another sub-predicate `database_arg` to build the arguments of the database corresponding to the attributes of the features.

- **test_pre_conditions** – This predicate is used to decompose the list of pre-conditions properties and to produce the LOTOS event(s) corresponding to each property. The decomposition is done sequentially in the same order as the properties. The corresponding LOTOS action(s) are produced by calling a sub-predicate `test_pre_condition` containing the action(s) to produce.

- **test_triggers** – This predicate is built on the same principles as the previous one and uses the sub-predicate `test_trigger` to produce LOTOS line(s) corresponding to each and every property.

- **test_results** – This predicate is built on the same principles as the previous one and uses the sub-predicate `test_result` instead of `test_trigger`.

- **test_footer** – This predicate is simply used to produce the end of the LOTOS test process, which does not contain anything special.

In the LOTOS specification, a user is associated with a phone number and a DA (see chapter 3, section 3.2 for more details).

Moreover, users in the LOTOS specification have different names than the users of the Prolog features specification. Mapping between users and their phone numbers and DAs is done by using a predicate `lotos_user`. Examples of the predicates we defined are:

- Mapping rule used to build the test header:

```
test_header([[F1, Fx], CndF1], [[F2, Fy], CndF2], [F, X], R):-
        write('process ScenarioFI_R'), write(R), write('_'),
        write(F1), write('_'), write(Fx), write('_and_'),
        write(F2), write('_'), write(Fy), write('_resulting_in_'),
        write(F), write('_'), write(X), write('_['), nl,
        write('\tUSER_to_DE, DE_to_USER, Init, success ]: noexit:=\n'),
        write('\n  let specificDB:Database ='),
        lotos_database([[F1, CndF1], [F2, CndF2]]),
        write(' in ( \n\n    Init !specificDB;\n\n').
```

- Mapping rule used to decompose sets of pre-conditions:

```
test_pre_conditions(CndF1, CndF2):-
        test_pre_conditions(CndF1),
        test_pre_conditions(CndF2), !.

test_pre_conditions([]).

test_pre_conditions([Elem|List]):-
        test_pre_condition([Elem]),
        test_pre_conditions(List).
```

- Mapping rule used map between a user and his phone number and DA.

```
lotos_user(alice, 'userA', '2001', 'debA0').
lotos_user(bob, 'userB', '2002', 'debB0').
lotos_user(carol, 'userC', '2003', 'debC0').
lotos_user(dave, 'userD', '2004', 'debD0').
```

- Examples of mapping rules for the pre-condition property `busy(X)`:

```
test_pre_condition([busy(X)]) :-
        lotos_user(X, User, _, Deb),
        write('   USER_to_DE !'), write(User), write(' !'), write(Deb),
        write(' !offHook'), write(';\n'),
        write('   DE_to_USER !'), write(Deb), write(' !'), write(User),
        write(' !dialTone'), write(';\n').
```

Based on these mapping rules, our tool produces LOTOS test suites. Directly applied to the specification, these test suites allow us to perform feature interaction validation and to verify that no interaction can be found in the specification.

As an example, let us consider the incoherence identified by rule #d2 for Call Forward Always and Call Forward on Busy. This incoherence is characterized by the fact that if a user subscribes to both features and forwards his calls to some user X with the first feature and to some user Y with the other feature, an incoherence is present when someone calls this user when he is busy. This contradiction is due to both features being triggered and leading to contradictory results since calls are forwarded to different users. The report for this incoherence follows.

```
% * Rule #d2 -> [cfb, 1] & [cfa, 1]

%   Same user subscribes to two features having the
%   same triggering events and contradictory results

%      + Features pre-conditions
%         - Pre-conditions of [cfb, 1]
%             subs(bob, cfb)
%             concerns(bob, cfb)
%             cfb(carol)
%             busy(bob)

%         - Pre-conditions of [cfa, 1]
%             subs(bob, cfa)
%             concerns(bob, cfa)
%             cfa(dave)

%      + Same triggering events

%          call(alice, bob)

%      + Contradictory results

%         - Resulting events of [cfb, 1]
%             redirected(alice, carol)
%             call(alice, carol)
%             ring(alice, carol)

%         - Resulting events of [cfa, 1]
%             redirected(alice, dave)
%             call(alice, dave)
%             ring(alice, dave)
```

The corresponding derived test suite is composed of two scenarios: the first one where Call Forward on Busy is triggered and the second one where Call Forward Always is triggered.

As an example, one of the two LOTOS test processes produced (resulting in Call Forward on Busy) is shown below.

```
process ScenarioFI_Rd2_cfb_1_and_cfa_1_resulting_in_cfb_1[
        USER_to_DE, DE_to_USER, Init, success ]: noexit:=

  (* database initialization *)
  let specificDB:Database =
        UF(userB, FD(cfb, fArg(2003), noArg), endFeatureSet),
        UF(2002, FD(cfb, fArg(2003), noArg), endFeatureSet),
        UF(userB, FD(cfa, fArg(2004), noArg), endFeatureSet),
        UF(2002, FD(cfa, fArg(2004), noArg), endFeatureSet),
        emptyDB)))) in (
    Init !specificDB;

    (* busy(bob) *)
    USER_to_DE !userB !debB0 !offHook;
    DE_to_USER !debB0 !userB !dialTone;

    (* call(alice, bob) *)
    USER_to_DE !userA !debA0 !offHook;
    DE_to_USER !debA0 !userA !dialTone;
    USER_to_DE !userA !debA0 !dial !2002;
    DE_to_USER !debA0 !userA !callInProgress;

    (* redirected(alice, carol), call(alice, carol) *)
    DE_to_USER !debA0 !userA !redirectTone;

    (* ring(alice, carol) *)
    DE_to_USER !debC0 !userC !ringingOn;

    success; stop
  )
endproc
```

The application of the two test scenarios using LOLA allows the tester to verify that:

- First, only one of the two features is always triggered: applied to the specification using LOLA (see chapter 5, section 5.1.2), one of the test scenarios results in a MUST PASS while the other results in a REJECT.
- Second, the feature that is triggered is the intended one. This must be verified by the designer, in accordance with the results (scenario resulting in a MUST PASS) and the requirements.

The Use Case Maps model and the LOTOS specification already existed for 6 of the features in consideration when we formalized our rules and built the tool. Hence, our method couldn't

influence the Use Case Maps or the LOTOS models for these 6 features (OCS, ICS, CFA, CFB, CT, CP). However, we were able to identify incoherences for all the 9 features including some that were not identified by Mitel's designers for the three last features. All possible incoherences identified were incoherences that could possibly lead to interactions in the specification.

In addition, we derived test suites and tested the LOTOS specification for the six features already specified No feature interactions were found, due to the fact that feature studied were well known to the specifiers; interactions were avoided by Mitel designers at the Use Case Maps level.

6.1.2 Case 2 – The Feature Interaction Workshop 2000

Following the first feature interaction detection contest held in 1998 [3], a second feature interaction contest was held in May 2000 [4, 43], in conjunction with the sixth international workshop on feature interactions in telecommunications and software systems. The goal of the workshop was to *"provide a simple comparison of different automated tools for feature interaction detection"*. In this second contest, the basic call model and the features are specified using Finite State Machines [42].

Most of the features are composed of two parts: *subscriber*, describing the behavior relative to the subscriber, and *everyone*, describing the behavior relative to non-subscribers. Fig. 6.1 presents the specification of the Terminating Call Screening feature.

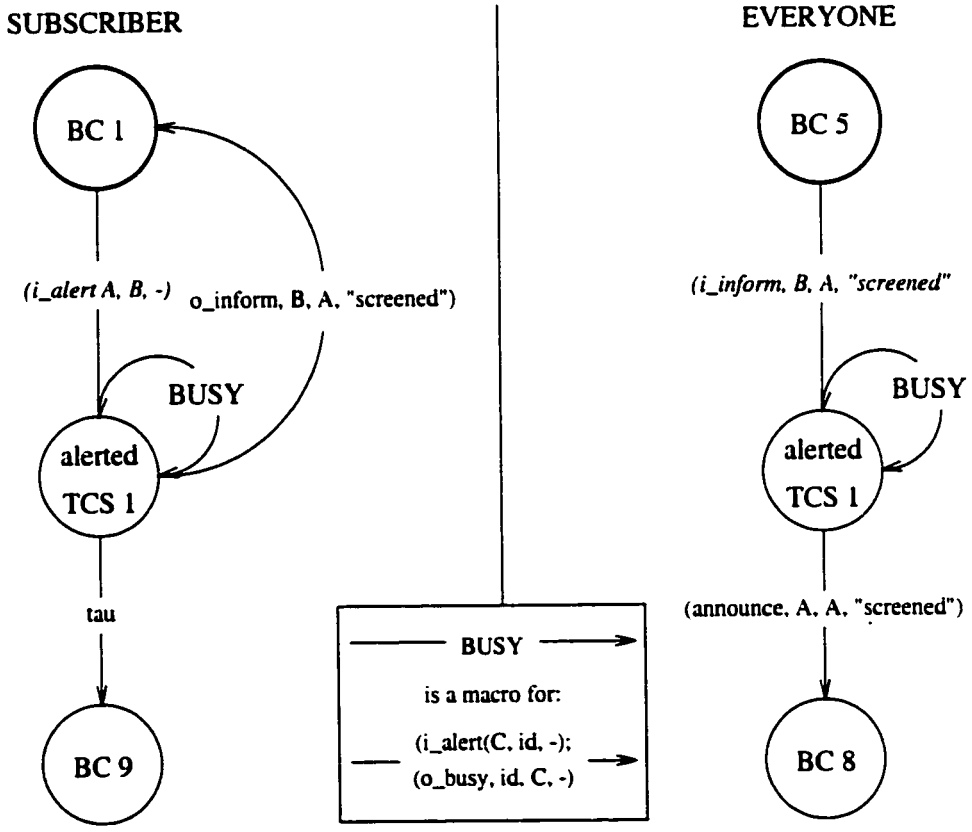


Figure 6.1: Terminating Call Screening

We only wanted to test our filtering method on a different set of features, so no test suites were derived. Moreover, we only applied our filtering method to the ten features of phase one of the contest. A formal specification of the features was already given (using Finite state machines); it was translated into our Prolog representation. The mapping of the FSM representation into Prolog feature descriptions is based on the following rules:

- The name of a description is composed of the name of the feature for the first part and the state of the feature description for the second part.
- States are mapped into pre-conditions. A list of pre-conditions includes the properties characterizing the state of the feature along with the name of the state.
- Transitions representing an incoming event are mapped into triggering events.
- Transitions representing an output are mapped into results.
- Only external observable actions, corresponding to inputs and outputs, are considered. Thus, states are only shown in pre-conditions; they do not appear in triggering events or results.
- An input followed by an output going from a state to another via an intermediary state (as is the case for *(i_alert A, B, -)* and *(o_inform B, A, "screened")* on fig. 6.1) are represented as a triggering event and a result. Intermediate states are states that do not correspond to any input or output. These states are not considered.

Our Prolog representation of Terminating Call Screening (fig. 6.1) is the following:

```
% *** Subscriber ***

% *** bc(1) ***
% B is subscriber, is idle and receives a call from A, A is screened
feature([tcs, bc(1)],
        [subs(C, tcs), concerns(C, tcs),                    % pre-conditions
         tcs_list(B), idle(C), bc(1)],
        [alert(B, C, -)],                                   % triggering events
        [inform(C, B, screened)]                            % results
    ):-
        user(B), user(C), C \= B.                           % constraints

% *** tcs(1) ***
% while screening a call, if a new call incomes, busy indication is sent
feature([tcs, tcs(1)],
        [subs(B, tcs), concerns(B, tcs), tcs_list(A),       % pre-conditions
         busy(B), alerted(B), tcs(1)],
        [alert(A, B, -)],                                   % triggering events
        [busy(B, A, -)]                                     % results
    ):-
        user(A), user(B), B \= A.                           % constraints
```

```
% *** Everyone ***

% *** bc(5) ***
% A calls B and is in screening list of B, A is screened by B
feature([tcs, bc(5)],
        [subs(B, tcs), concerns(A, tcs),                % pre-conditions
         tcs_list(A), bc(5)],
        [inform(B, A, screened)],                       % triggering events
        [announce(A, A, screened)]                      % results
    ):-
        user(A), user(B), B \= A.                       % constraints

% *** tcs(2) ***
% while being screened, if a call is received, busy indication is sent
feature([tcs, tcs(2)],
        [subs(B, tcs), concerns(A, tcs), busy(A),       % pre-conditions
         tcs_list(A), annoucement(A), tcs(2)],
        [alert(C, A, -)],                               % triggering events
        [busy(A, C, -)]                                 % results
    ):-
        user(A), user(B), B \= A,                       % constraints
        user(C), C \= A, C \= B.
```

The ten features were represented using 97 description parts. The application of our filtering

	CFB	TL	TCS	CW	TWC	RC	CNDB	RBWF	VM	SB
CFB	2	1	2	4	1	1		5	3	1
TL	—			1	7			6	3	
TCS	—	—		2				3		
CW	—	—	—		2	1		11	9	1
TWC	—	—	—	—		1	1	33	24	1
RC	—	—	—	—	—		1	1		1
CNDB	—	—	—	—	—	—				1
RWBF	—	—	—	—	—	—	—		16	1
VM	—	—	—	—	—	—	—	—		
SB	—	—	—	—	—	—	—	—	—	

—

Number of incoherences per rule and type						
Rule	#d1	#d2	#d3	#d4	#t1	#t2
Number of incoherences	94	1	49	0	2	1
Ratio per rule	64%	0.75%	33%	0%	1.5%	0.75%
Ratio per type	97.75%				2.25%	

Table 6.2: Feature Interaction Workshop Case Study: Results

method on the whole specification reported 149 incoherences. Table 6.2 summarizes the incoherences identified. Most of them were identified by direct incoherence rules.

The number of incoherences is significant but some of them are incoherences that cannot lead to feature interactions in the specification. These *false* incoherences are characterized by the fact that the tool identifies a possible incoherence between two features X and Y that are respectively in states S_m and S_n while, in fact, this situation cannot occur in the specification. This fact shows a limitation of our method.

An example of such an incoherence is the following: Alice subscribes to Voice Mail. Bob subscribes to Three Way Calling. Alice is involved in a three way conversation and at the same time, plays back the messages she received with Voice Mail. If Alice goes onhook, results are contradictory since both features react differently. However, Alice can never listen to her messages and be involved in a three way conversation at the same time.

These *false* incoherences are due to the algorithm identifying an incoherence when two features have the same triggers and different or contradictory results, while their pre-conditions show incompatible states, meaning that both features should not be triggered at the same time.

This problem is not related to the rules we developed but to the way we modeled the features: incompatible states are not identified in our model. This can be done by refining the properties used to represent states and by building contradiction pairs corresponding to incompatible states. If this is done, two features presenting incompatible states would present a contradiction in their pre-conditions, and would not be considered by the rules. This emphasizes that features must be carefully modeled and that information about contradictions between properties must be clearly stated. It also demonstrates that our method forces the designer to do a precise modeling.

6.2 More Complicated Examples

So far we have presented simple examples to facilitate the understanding of our concepts. To illustrate the power of our method, more complicated examples of incoherence identifications are presented. Examples are taken from the Mitel specification. As already mentioned, the complete report of all incoherences identified can be found in appendix A.

Rule #d3 → [cw, 3] & [cw, 1] – Alice and Bob both subscribes to Call Waiting. Alice is busy talking to Bob and Carol called Alice, which implies that Carol is held by the Call Waiting feature of Alice. If Dave calls Carol, Alice's Call Waiting specifies that Carol should send a busy signal while Carol's Call Waiting specifies that Dave should be put on hold.

Rule #t1 → [ct, 3] & [ocs, 1] – Alice subscribes to Call Transfer, holds Bob and is talking to Carol. Bob subscribes to Outgoing Call Screening. Alice wants to transfer Bob to Carol. Alice's Call Transfer intention is that Bob talks to Carol while Bob's Outgoing Call Screening is that any call to Carol is blocked.

Rule #t1 → [ct, 3] & [ics, 1] – Alice subscribes to Call Transfer, holds Bob and is talking to Carol. Carol subscribes to Incoming Call Screening. Alice wants to transfer Bob to

Carol. Alice's Call Transfer intention is that Bob talks to Carol while Carol's Incoming Call Screening is that any call from Bob is denied.

6.3 Performance Evaluation

This section presents some metrics concerning the time taken to built the Prolog representation for the Mitel features as well as for the Feature Interaction Workshop features. In addition, an overview of the algorithmic complexity of the analysis is given.

6.3.1 Human Aspects Considerations

Mitel provided us with 9 features. They were split in 20 descriptions and the modeling of these descriptions took a total of 3 days. The application of the filtering process identified 43 possible incoherences. 30 mapping rules were derived to produce test suites and no interaction was found in the LOTOS specification. Table 6.3 summarizes the results. We

Features	Descriptions	Time	Incoherences	Mapping rules	Time
9	20 (2.2 / feature)	3 days	43	30	3 days

Table 6.3: Mitel Case Study: Design Time

considered the features given in the first phase of the second Feature Interaction Workshop contest. The contest instructions [42] contain the FSM definition of 10 features. We split them into 97 descriptions (9.7/feature). The specification took a total of 15 days. 149 possible incoherences were identified by our tool. Table 6.4 summarizes the results: The

Features	Descriptions	Time	Incoherences
10	97 (9.7 / feature)	15 days	149

Table 6.4: Feature Interaction Workshop Case Study: Design Time

number of descriptions per feature for the Feature Interaction Workshop is greater than for Mitel. This is due to the fact that the Mitel features were represented at a very high level, which implies a low level of details, and thus, less description parts. The Feature Interaction Workshop features were already specified in a detailed manner (using FSM), which led to more feature descriptions, to be in accordance with the level of detail of the FSM representation.

6.3.2 Algorithmic Performance

The algorithm developed consists in taking a list of feature descriptions and applying each filtering rule on all pair-wise combinations, attempting to prove that an incoherence is present. Since we consider pair-wise combinations, the analysis of n feature descriptions results in the analysis of n^2 pairs. Each pair must be analyzed with all the rules. This results in $6n^2$ combinations to treat.

Therefore, the order of the algorithmic complexity of the analysis is n^2. The curve presented in fig. 6.2 shows the function $6n^2$. The number of combinations increases considerably along with the number of features considered. Working with pair-wise combinations always leads to this order of algorithmic complexity.

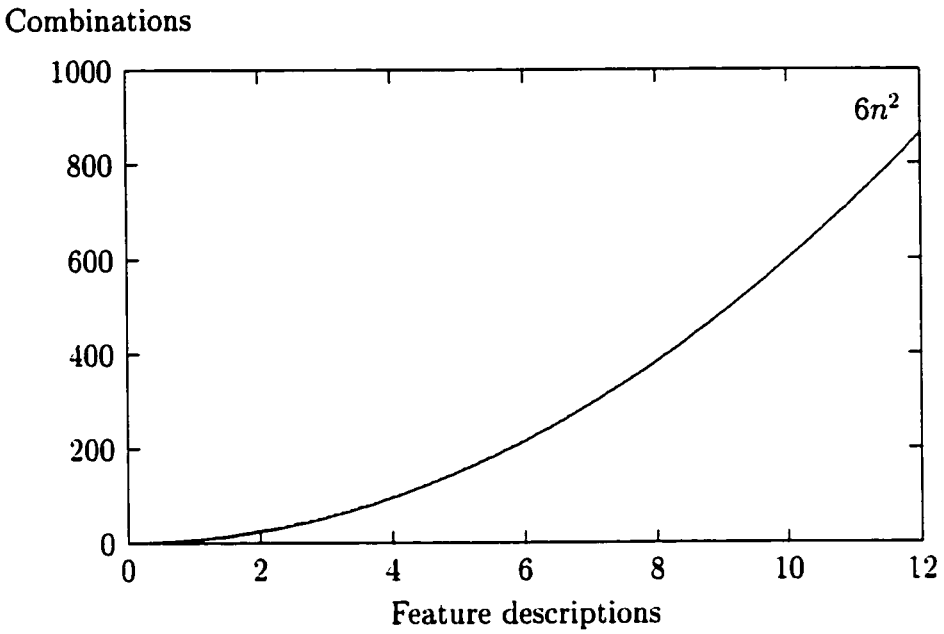


Figure 6.2: Number of Combinations over Feature Descriptions

It is not possible to reduce the order of this algorithmic complexity but Prolog interpreters are fast and the analysis of all combinations takes little time.

For a given rule r and a given combination of features $|\mathcal{D}_{X,m} \bullet \mathcal{D}_{Y,n}|$, the analysis of this combination deals with the properties and variables of both feature descriptions $\mathcal{D}_{X,m}$ and $\mathcal{D}_{Y,n}$. As mentioned, the goal is to satisfy the filtering rule. To obtain rule satisfaction, Prolog tries all possible bindings of variables until it finds a proper combination or until all combinations have been tried. The number of tries to bind variables with proper values depends on the number of users considered and on whether an incoherence is present or not.

The algorithm is built in such a way that variables of the first feature description are bound with default users, basically the first users that satisfy the constraints. Then, Prolog tries to bind variables of the second feature description with all possible combinations until an incoherence is found or everything has been tried.

Let us consider two features $\mathcal{D}_{X,m}$ and $\mathcal{D}_{Y,n}$ analyzed by a rule The analysis is done as follows:

1. Bind variables of properties of $\mathcal{D}_{X,m}$ with the first not yet tried combination of users that satisfies $\mathcal{C}_{Y,n}$ and go to step 2.

2. Value combinations remain untried ?
   - *YES* - Bind variables of properties of $\mathcal{D}_{Y,n}$ with first possible combination that satisfies $\mathcal{C}_{X,m}$ and go to step 3.
   - *NO* - Go to step 7.

3. $\mathcal{P}_{X,m}$ and $\mathcal{P}_{Y,n}$ present a contradiction ?
   - *YES* - Backtrack to step 2.
   - *NO* - Go to step 4.

4. $\mathcal{R}_{X,m} \supseteq \mathcal{T}_{Y,n}$, or, $\mathcal{R}_{X,m} = \mathcal{T}_{Y,n}$?
   - *YES* - Go to step 5.
   - *NO* - Backtrack to step 2.

5. $\mathcal{R}_{X,m}$ and $\mathcal{R}_{Y,n}$ present a contradiction ?
   - *YES* - Go to step 6.
   - *NO* - Backtrack to step 2.

6. The rule is satisfied and an incoherence is identified.

7. Rule fails, no incoherence are identifiable.

The number of times that the algorithm backtracks depends on how an incoherence is identified, if ever. Given a feature description that contains n users, if an incoherence is found within the first try, all steps are visited only once. But if no incoherence exists, all combinations of users are tried and the steps are visited n^2 times until Prolog can conclude that no incoherence has been found.

Rules #d1 and #d2 can also influence the number of iterations: since they are symmetric, only one pair needs to be analyzed. Hence, the number of combinations considered by each rule is half the number of combinations considered by any of the other rules.

It is impossible to determine the exact number of iterations for a given list of feature descriptions. However, the algorithmic complexity must be computed with respect to the number of feature descriptions and the number of variables. Let us denote f the number of features and u the number of users. The algorithmic complexity of the analysis, denoted $\mathcal{C}p$, lies between a minimum and a maximum expressed by: $5f^2 \leq \mathcal{C}p \leq u^2(5f^2)$.

Considering a specification containing 4 users, the algorithmic complexity lays between $5f^2$ and $16(5f^2)$, as illustrated by the curves presented in fig. 6.3.

Thus, the algorithmic complexity of our tool is polynomial. Such a complexity is unfortunately not reducible. Prolog must try all possible combinations to be able to give a verdict. Even if the complexity is high, the execution time remains very reasonable. Executed on a Pentium II 450 Mhz, our tool gave results in a very short time considering the number of pairs analyzed:

- Mitel specification, 20 feature descriptions (400 pairs) and 4 users: 70,122,509 inferences in 84.14 seconds. 43 possible incoherences found.

- FIW00 specification, 97 feature descriptions (9409 pairs) and 4 users: 789,818,572 inferences in 1299.93 seconds. 149 possible incoherences found.

Considering that our tool needs 21 minutes and 40 seconds to analyze 9409 description pairs and find 149 possible incoherences, we can conclude that the execution time remains very reasonable in spite of the algorithmic complexity presented.

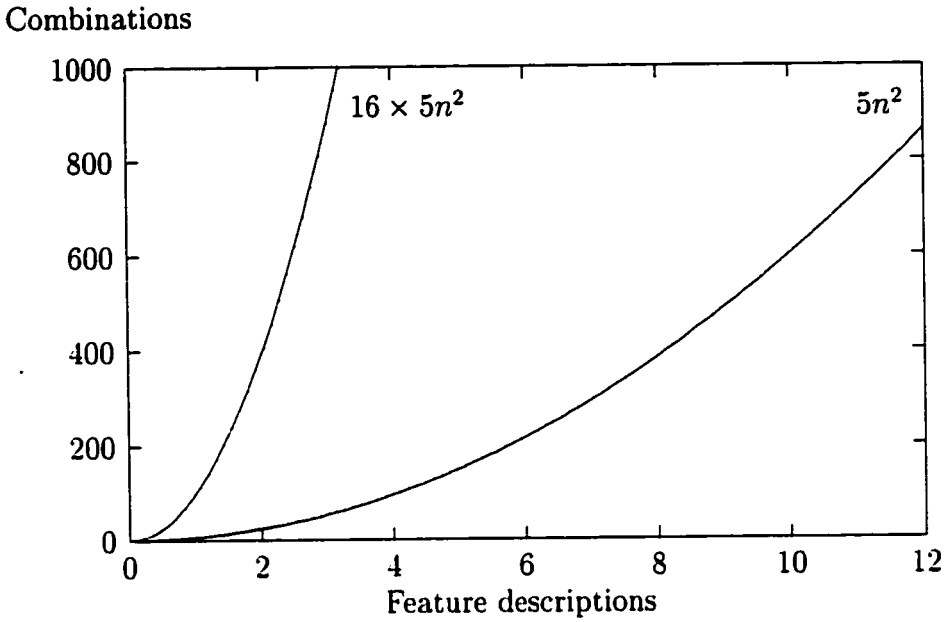


Figure 6.3: Range of Possible Complexity

6.4 Comparison

Tadashi Ohta and Tae Yoneda developed a similar method [30, 31], already mentioned in chapter 2, section 2.2.6. The most recent version of their approach [31] is illustrated in fig. 6.4. This approach starts with the specification of features using STR [44]. The syntax of STR.

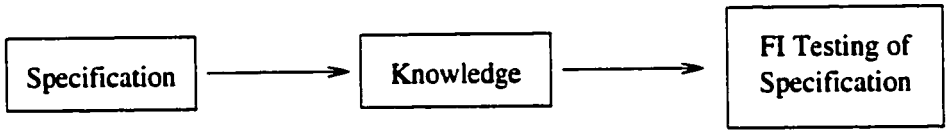


Figure 6.4: Approach Overview

composed of pre-conditions, event, and post-conditions is very similar to the one we use. which is also composed of pre-conditions, triggering-events and results. The specification is analyzed by an automatic tool that *elicits* knowledge relative to the combination of features. Such knowledge is then used to manually identify feature interactions and thus validate the specification.

The knowledge is modeled using states and transitions. A Service is represented as a set of states and rules. A system state is represented as a set of state elements called state descriptions primitives (simply called primitives).

Fig. 6.5 illustrates a case where primitives `calling(A, C)`, `idle(B)` and `m-cfv(B, C)` represent actions involved in the call forwarding and were the system goes in a state where the call from A to C is established and C rings.

A new service is added by adding new state transitions. The new transitions take precedence over the previously existing ones. Previous transitions and states to which they lead are called inhibited primitives while the new ones are called intention primitives.

Let us consider a state S_i and a rule R_e going from this state to another state S_{i+1}. Let us consider the addition of a feature implementing a new functionality. This addition,

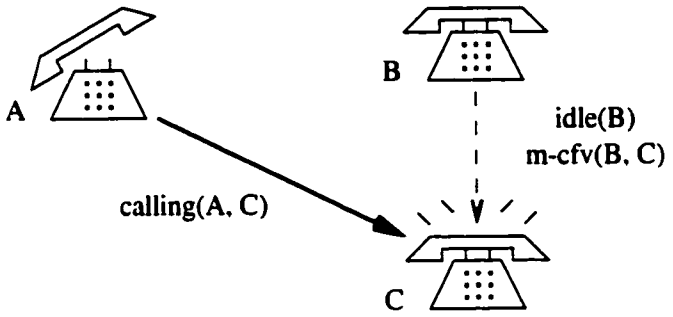


Figure 6.5: Knowledge Formal Definition: System State

illustrated by fig. 6.6, is represented as a new rule R_a going from state S_i to a new state S_j: rule R_a replaces rule R_e and thus, state P_j replaces state P_i. P_j becomes an inhibited primitive and P_j is an intention primitive. As previously mentioned, feature specifications

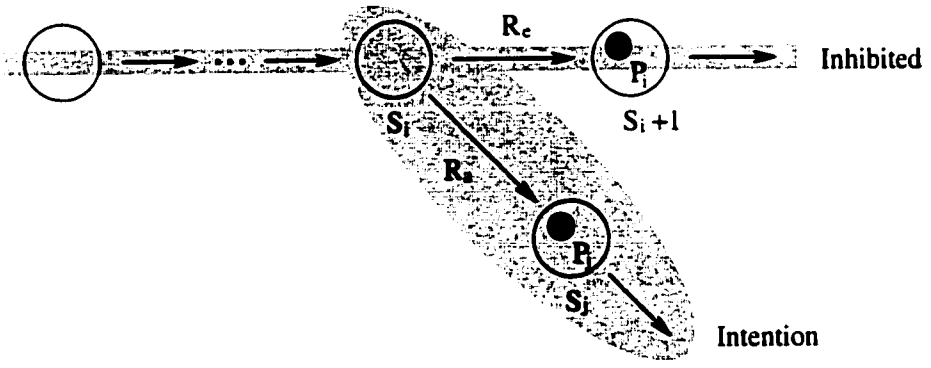


Figure 6.6: Knowledge Formal Definition: Primitives

are represented using a description language called STR (State Transition Rule) [44]. A feature is modeled using pre-conditions, events and post-conditions. This representation is very close to the one we use, based on pre-conditions, triggering events and results. However, since a feature is triggered by an event when the system is in a certain state (pre-conditions) and results in some new state (post-conditions), using STR seems to be the most appropriate to represent features.

Based on STR, T. Ohta and T. Yoneda have implemented an algorithm that *elicits* knowledge from the specification. According to our understanding, *elicitation* denotes the extraction of knowledge for the identification of inhibited and intention primitives with respect to the combinations of features. The extraction of this knowledge is automatic. Such knowledge is later manually analyzed to identify feature interactions. The interactions are characterized by specific rules. This rules are given in [31] and follow:

- *An intention primitive for one service is defined as an inhibited primitive in the other service.*

- *An inhibited primitive for one service (suppose service A) is an element of the post-pondition of a rule belonged to the other service and the rule takes precedence over the rule belonged to service A.*

- *When a rule of service A, r_a, does not take precedence over a rule of service B, r_b, r_b has not an intention primitive for service A in post-condition while r_b has the same event which is a condition for intention primitive of service A and has the same primitive in pre-condition which is a condition for intention primitive of service A.*

These rules are similar to those used in our approach. Such similarity is not surprising since features described with similar notation should lead to similar procedures. However, even if similar in the notation and rules, the approach of T. Ohta and T. Yoneda and the one we developed present substantial differences located at different stages of the process and even in the process itself. The comparison can be summarized as follows:

Design stages covered - T. Ohta and T. Yoneda method acts at the specification and testing levels while ours acts at requirements, specification and also testing levels.

Automation - T. Ohta and T. Yoneda have automatic knowledge *elicitation* but identify interactions manually (automation is considered in future work); we have automatic identification of incoherences as well as automatic test derivation, based on manual elaboration of mapping rules.

Types of FI detected - T. Ohta and T. Yoneda consider interactions caused by non-determinism and contradictions but do not seem to consider those caused by loops. We consider the three kinds.

These two approaches remain different despite their similarities. T. Ohta and T. Yoneda's method works at the specification level. Knowledge is *elicited* from the specification and feature interactions are detected in the specification. Our approach works both at requirements and specification level. It offers a filtering process that can be applied either to the requirements or to the specification, and allows the derivation of test suites that can be used to validate the specification (and even the final product).

Moreover, the filtering of incoherences is automatic and the derivation of test suites are semi-automatic, while the approach developed by T. Ohta and T. Yoneda is semi-automated in the sense that, if knowledge *elicitation* is automatic, identification of feature interactions remain manual.

6.5 In Summary

In this chapter, we illustrated the applications of our tool with examples based on the Mitel specification and the Feature Interaction Workshop. We also presented an overview of the algorithmic complexity. Results show that, even if highly complex in practice, the execution of our tool is timely reasonable. In addition, we presented a comparison with a similar approach by other authors.

Chapter 7

Conclusion & Future Work

This chapter recapitulates the thesis and lists our contributions. It also presents a number of ideas about further research and improvements of our method.

7.1 Achievements

This thesis addresses the feature interaction problem in the research area of telephony systems design. Starting with a review of this problem, we presented our definition of a feature and of a feature interaction. We enumerated and discussed a list of some known approaches with respect to their strengths and limitations. In addition to these approaches, we presented the initial approach as well as our proposed refinements.

The initial validation approach consists in deriving a Use Case Maps model from the requirements, allowing the designer to do a first validation. This model is in turn derived into a corresponding LOTOS specification. It provides an executable model and allows designers (and testers) to perform various kinds of validations (exploration or scenario based). Scenario based validation is used to perform feature interaction detection. However, this approach is based on manual validation of UCM as well as on manual derivation of test suites for feature interactions detection.

Our research mainly consists of the automation of this method. We proposed refinements that consist of a tool targeting the automatic identification of incoherences between features at the requirements stage as well as the automatic derivation of test suites. These test suites target the validation of a specification in order to check whether or not interactions corresponding to incoherences identified in requirements lead to feature interactions in the specification.

The filtering process is based on rules characterizing incoherences. Feature descriptions are formalized using information available in the requirements. Our tool performs pair-wise analysis of features and identifies pairs that satisfy one or more incoherence rules. This identification of pairs presenting incoherences allows the designer to refine the requirements and to produce a better UCM model using the knowledge available about possible interactions. Moreover, by transitivity, this also insures a better LOTOS model. Moreover, the same rules might be used for both UCM to LOTOS and test suites generation.

Our derivation of validation test suites (which follows the derivation of the LOTOS specification) can be seen as a semi-automatic process due to the fact that it requires the

designer to manually provide mapping rules to go from the formal representation of features to LOTOS scenarios. Nevertheless, the elaboration of these mapping rules is fast for a designer that knows the LOTOS model.

The rules are used to produce validation test suites corresponding to the potential interacting pairs. The test suites are then applied against the LOTOS model, performing scenario based validation, to check whether the interactions corresponding to the incoherences identified exist or not in the specification. Derivation of test suites is not specific to LOTOS and mapping rules could be derived to produce scenarios in other languages.

Mitel Corp. provided us with a UCM model from which we derived a LOTOS specification. That is to say, our filtering process could not influence the requirements or the UCM model. We refined our approach and built our tool after the derivation of the LOTOS specification and the integration of 6 of the 9 features. This gave us the opportunity to identify interactions that were not suspected by Mitel's designers.

In the future, the order of these activities should be reversed in accordance with the approach presented in by figure 2.2 in chapter 2, section 2.4.

We applied our filtering process, identified incoherences and produced (via mapping rules) validation test suites for the first 6 features. Applied against the LOTOS specification, the test suites revealed that the model was free of feature interactions, meaning that all potential interactions were avoided by Mitel designers at the UCM level. However, all incoherences identified were incoherences that could have led to interactions in a system and all the scenarios derived were successfully applied against the specification, showing that no interactions were present.

In order to extend our benchmark, we applied the filtering process to the specification of the Feature Interaction Workshop 2000 (FIW00). The application of this process gave us interesting results. Contrary to the results obtained with Mitel's specification, those obtained with the FIW00 specification showed the identification of some impossible incoherences. An analysis of the results brought us to the conclusion that these results were due to a poor formalization of the features. Such results show that our method enforces a clean design and formalization of the features, which we consider as being a strength.

7.2 Contributions

This section lists our contributions to the software engineering area and to the feature interaction problem in the research area of telephony systems design.

7.2.1 Translation of a UCM Model into a LOTOS Specification

The translation of UCM into LOTOS was first proposed by D. Amyot in [21]. Chapter 3 of the thesis presents the translation of the system provided by Mitel Corp. This system is modeled as a set of UCM composed of 110 different maps distributed in seven different levels sub-maps and contains 9 feature definitions.

We provided explanations about the analysis and decisions (chapter 3, section 3.3.2) and established general translation guidelines (chapter 3, section 3.3.2) that are applicable to

other systems. In addition, we have presented the approach used for integrating features to the specification.

These achievements show that the translation of such a large UCM model into a LOTOS specification is possible with reasonable efforts. The construction of the specification for the purpose of filtering is also easy and quick, as shown in chapter 6, section 6.3.1. In addition, it provides useful information that can be used by designers to refine requirements and to avoid interactions in the specification.

7.2.2 Incoherences Filtering Process

We proposed an approach that allows the automatic identification of incoherences between features at the requirements level. Chapter 4 presents the rules used for the identification of such incoherences. The tool implementing the automatic application of the rules is presented in chapter 5.

This improvement allows a faster identification of potential feature interactions with less human intervention than in the manual approaches.

7.2.3 Automatic Derivation of Validation Test Suites

Identifying incoherences at the requirements stage does not insure that these correspond to interactions in the specification.

The use of mapping rules, presented in chapter 5, section 5.3 enables the automatic derivation of test suites with respect to a given specification. These test suites are applied against the specification to insure that this one does not contain the feature interactions corresponding the the incoherences identified by the filtering.

This reduces the testing time in the sense that, once the mapping rules are manually specified, tests can be automatically derived for all incoherences identified without requiring further manual intervention. This process, which we demonstrated by using LOTOS, is not strictly dependent on the use of LOTOS; other specification languages could be used.

7.2.4 Scenario Based Validation of a Large LOTOS Specification

The manual validation, based on test scenarios derived from UCM, is presented in chapter 5, section 5.1.1. The automatic validation, based on test suites obtained from the filtering results via the use of mapping rules, is presented in chapter 5, section 5.3.

Both validations are done by applying the scenarios against the LOTOS specification using LOLA. This technique is presented in chapter 5, section 5.1.2.

These validations performed on the specification show that the scenario based validation of a large LOTOS specification is indeed possible.

7.2.5 Use of LOTOS in an Industrial Context

The translation of the UCM model resulted in a LOTOS specification that is 2410 lines long. This specification breaks down into 450 lines of Abstract Data Types (ADT) and 1960 lines of behavior.

By translating such a UCM model into a LOTOS specification and by performing feature interaction detection on this specification, we showed that the use of LOTOS in an industrial context is appropriate. The information about analysis and decisions as well the translation guidelines presented in chapter 3 show that building a LOTOS specification from a UCM model is not a complicated task.

7.2.6 Applications & Conclusive Results over two Case Studies

The filtering process is dedicated to the requirements and does not concern the validation of the specification except in the sense that it provides information that helps the designers to derive a better specification from requirements. Hence, the specification that is derived should not contain any of the interactions corresponding to the incoherences identified by the filtering process. Since the designers are aware of them, they are expected to derive the specification in such a way that interactions are avoided.

The only limitation is that the filtering process does not insure that all possible interactions are detected. This fact especially concerns multi-way incoherences since only pair-wise combinations are considered. This implies that the absence of incoherences does not insure the absence of interactions in the specification.

In spite of these considerations, the results obtained from the application of our approach over two different case studies showed that this methodology improves the design and validation in the specification process. These results are presented and commented in chapter 6. Moreover, we have proposed a representation of features and a filtering approach that is easy to use. This approach only requires knowledge of Prolog and of the language used for the specification to derive the mapping rules.

It should also be noted that in the industrial case study with Mitel corp., an interaction that was unknown to Mitel's designers was found. Since this case study was based on well known interactions, this gives hope that more interactions could be found in a set of unknown interactions.

7.3 Further Research

Many improvements are still possible on our method. This section presents those that are being considered.

7.3.1 Priorities

As presented in chapter 4, section 4.3, many incoherences can be avoided by using a priority mechanism. However, the way that priorities are currently represented by using contradiction pairs is not sufficient to model complex priority schemes.

In chapter 4, section 4.3.2, we provide an idea of how this problem could be solved.

7.3.2 Multi-way Incoherences

The filtering process only considers pair-wise combinations, and thus, only identifies two-way incoherences. Even if this allows the identification of a significant number of them,

identifying three-way or even n-way incoherences is an improvement that must be considered. This requires the establishment of new incoherence rules for characterizing these kinds of incoherences.

Similarly to two-way incoherences, a simple three-way incoherence that can be identified is the direct incoherence present between three features held by the same user. A Three-way rule can identify incoherences occurring when the features do not present any contradiction in their preconditions, have the same triggering events, and have different or contradictory results.

7.3.3 Distribution over a Many Hosts Network

Since our tool combines features in a pair-wise manner, the distribution of the analysis over a many hosts network could be an improvement. Distributing the analysis of features over different hosts can significantly reduce the global analysis time since the workload would be shared.

However, due to the communication cost concerning the transmissions of feature descriptions and reports, the distributed analysis of a small group of features can possibly take more time than a centralized one. Nevertheless, the distribution of the tool would increase, in most cases, its power by reducing its execution time over most feature sets.

7.3.4 Considering other methods

The automatic analysis of incoherences is based on Prolog. However, it appears that the use of a predicate for incoherences (or contradictions) is unfortunately one of the few things that one can do to handle negation in traditional Prolog. Refining the rules using newer techniques such as linear logic [45] would possibly allow to express them in a more efficient way.

Another consideration is to link our rules to the OPI (Obligation, Permission, Interdiction) [34] model. Refining rules based on this model will, for instance, allow to make a distinction between *Obliged*, *Permitted* and *Forbidden* results of a feature.

Bibliography

[1] P. Zave. http://www.research.att.com/~pamela/faq.html. FAQ Sheet on Feature Interaction.

[2] P. Zave. Architectural solutions to feature-interaction problems in telecommunications. In *Feature Interactions in Telecommunications and software systems VI*, pages 10–22. IOS Press, 1998.

[3] K. Kimbler L.G. Bouma (editors). *Feature Interactions in Telecommunications and software systems VI.* IOS Press, 1998.

[4] M. Calder E.H. Magill. *Feature Interactions in Telecommunications and software systems VI.* IOS Press, 2000.

[5] S.L. Pfleeger. *Software Engineering, Theory and Practice.* Prentice Hall, 1998.

[6] T. Bolognesi E. Briskma. Introduction to the ISO specification language LOTOS. In *The Formal Description Technique LOTOS*, pages 23–73. Elsevier Science, 1989.

[7] J. Ellsberger D. Hogrefe A. Sarma. *SDL Formal Object-Oriented Language for Communicating Systems.* Prentice Hall, 1997.

[8] K.J. Turner. *Using Formal Description Techniques.* J. Wiley & sons Ltd., 1993.

[9] R.J.A. Buhr D. Amyot M. Elammari D. Quesnel T. Gray S. Mankovski. Feature-interaction visualization and resolution in an agent environment. In *Feature Interactions in Telecommunications and software systems V*, pages 135–149. IOS Press, 1998.

[10] D. Marples E.H. Magill. The use of rollback to prevent incorrect operation of features in intelligent n etwork based systems. In *Feature Interactions in Telecommunications and software systems V*, pages 115–134. IOS Press, 1998.

[11] P. Combes S. Pickin. Formalisation of a user view of network and services for feature interaction detection. In *Feature Interaction In Telecommunications Systems*, pages 120–135. IOS press, 1994.

[12] M. Faci L. Logrippo. An algebraic framework for the feature interaction problem. In *Proc. of the 3rd AMAST Workshop on Real-Time Systems*, pages 280–294, 1996.

[13] E. Kit. *Software Testing in the Real World, Improving the Process.* Addison Wesley, 1996.

[14] G. Utas. A pattern language of feature interaction. In *Feature Interactions in Telecommunications and software systems V*, pages 98–114. IOS Press, 1998.

[15] R.J.A. Buhr R.S. Casselman. *Use Case Maps for Object-Oriented Systems.* Prentice Hall, 1996.

[16] R.J.A. Buhr. Use Case Maps as architectural entities for complex systems. In *IEEE Transactions on Software Engineering (TSE)*, pages 1131–1155, 1998.

[17] D. Amyot. http://www.usecasemaps.org/. Use Case Maps Web Page and UCM User Group.

[18] D. Amyot R.J.A Buhr T. Gray. Use Case Maps for the capture and validation of distributed systems requirements. In *Fourth International Symposium on Requirements Engineering (RE'99)*, 1999.

[19] D. Amyot L. Logrippo. Use Case Maps and LOTOS for the prototyping and validation of GPRS Group-Call. In *Computer Communications 23(8)*, 1999.

[20] D. Amyot L. Charfi N. Gorse T. Gray L. Logrippo. J. Sincennes B. Stepien T. Ware. Feature description and feature interaction analysis with Use Case Maps and LOTOS. In *Feature Interactions in Telecommunications and software systems VI*, pages 274–289. IOS Press, 2000.

[21] D. Amyot. Formalization of timethreads using LOTOS. Master's thesis, University of Ottawa, Ottawa, 1994.

[22] J. Kamoun L. Logrippo. Goal oriented feature interaction detection in the intelligent network model. In *Feature Interactions in Telecommunications and software systems V*, pages 172–186. IOS Press, 1998.

[23] A. Valmari. The state explosion problem. lectures on petri nets i: Basic models. In *Lecture Notes in Computer Science 1491*, pages 429–528. Springer-Verlag, 1998.

[24] J.E. Hopcroft J.D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison Wesley, 1979.

[25] J. L. Peterson. An introduction to petri nets. In *Proc. of the National Electronics Conference, Chicago, Illinois*, pages 144–148. National Engineering Consortium, Inc., 1978.

[26] M. Nakamura Y. Katuda T. Kikuno. Feature interaction detection using permutation symmetry. In *Feature Interactions in Telecommunications and software systems V*, pages 187–201. IOS Press, 1998.

[27] Z. Manna A. Pnueli. *The temporal logic of reactive and concurrent systems.* Springer-Verlag, 1992.

[28] A.P. Felty K.S. Namjoshi. Feature specification and automatic conflict detection. In *Feature Interactions in Telecommunications and software systems VI*, pages 179–192. IOS Press, 2000.

[29] R.H. Hardin Z. Har'el R.P. Kurshan. Cospan. In *Conference on computer aided verification*, 1996.

[30] T. Yoneda T. Ohta. A formal approach for definitions and detection of feature interactions. In *Feature Interactions in Telecommunications and software systems V*, pages 202–216. IOS Press, 1998.

[31] T. Yoneda T. Ohta. Automatic elicitation of knowledge for detecting feature interactions in telecommunications services. In *IEICE Transactions on Information and Systems*, pages 640–647. IEICE, Japan, 2000.

[32] E. Rudolph P. Graubmann J. Grabowski. Tutorial on message sequence charts. http://http://www.sdl-forum.org/MSC/index.htm.

[33] J. Grabowski A. Wiles C. Willcock D. Hogrefe. On the design of the new testing language ttcn-3. In *Testing of communicating systems tools and techniques*, pages 161–176. KAP, 2000.

[34] M. Barbuceanu T. Gray S. Mankowski. How to make your agents fulfil their obligations. In *Proceedings of PAAM-98*, 1998.

[35] S. Pavón D. Larrabeiti G. Rabay. LOtos LAboratory user manual (version 3r6). Universidad Politécnica de Madrid, 1995.

[36] Groupe d'Intelligence Artificielle, université de Marseilles, France. *Prolog: Manuel de Reference et d'Utilisation*, 1975.

[37] W.F. Clocksin C.S. Mellish. *Programming in Prolog.* Springer, Fourth Edition.

[38] A. Miga. Application of Use Case Maps to system design with tool support. Master's thesis, University of Carleton, Ottawa, 1998.

[39] C.A. Vissers G. Scollo M. van Sinderen E. Brinksma. Specification styles in distributed systems design and verification basic models. In *Lecture Notes in Computer Science 1491*, pages 179–206. Theoritical Computer Science '89, 1989.

[40] P.H. Winston. *Artificial Intelligence, Third Edition, chapter 13.* Addison Wesley, 1992.

[41] J. Wielemaker. http://swi.psy.uva.nl/projects/SWI-prolog/. SWI-Prolog.

[42] M. Kolberg E.H. Magill D. Marples S. Reiff. Second feature interaction contest, instructions. In *Feature Interactions in Telecommunications and software systems VI*, pages 293–325. IOS Press, 2000.

[43] E.H. Magill. http://www.comms.eee.strath.ac.uk/~fi/fiw00/. Feature Interaction Workshop 2000.

[44] T. Yoneda T. Ohta. Declarative language str (state transition rule). In *FIRE works Workshop on Language Construcs for Describing Features*, pages 39–54. S. Gilmore & M. Ryan, 2000.

[45] Philip Wadler. A taste of linear logic. In *Mathematical Foundations of Computing Science*. Springer Verlag LNCS 711, 1993.

Appendix A

Filtering Results for Mitel

```
• Rule #d1 -> [cfb, 1] & [ics, 1]
  Same user subscribed to two features having the
  same triggering events and different results

    • Features pre-conditions

      - Pre-conditions of [cfb, 1]
          subs(bob, cfb)
          concerns(bob, cfb)
          cfb(carole)
          busy(bob)

      - Pre-conditions of [ics, 1]
          subs(bob, ics)
          concerns(bob, ics)
          ics_list(alice)

    • Same triggering events
        call(alice, bob)

    • Different results

      - Resulting events of [cfb, 1]
          redirected(alice, carole)
          call(alice, carole)
          ring(alice, carole)

      - Resulting events of [ics, 1]
          deny_call(bob, alice)
          call_denied(bob, alice)
```

```
• Rule #d1 -> [cfb, 1] & [cw, 1]
  Same user subscribed to two features having the
  same triggering events and different results

    • Features pre-conditions

      - Pre-conditions of [cfb, 1]
          subs(bob, cfb)
          concerns(bob, cfb)
          cfb(carole)
          busy(bob)

      - Pre-conditions of [cw, 1]
          subs(bob, cw)
          concerns(bob, cw)
          busy(bob)
          talk(bob, carole)

    • Same triggering events
        call(alice, bob)

    • Different results

      - Resulting events of [cfb, 1]
          redirected(alice, carole)
          call(alice, carole)
          ring(alice, carole)

      - Resulting events of [cw, 1]
          hold(bob, alice)
          cw_notify(bob)
```

```
• Rule #d1 -> [cfb, 1] & [cw, 4]
  Same user subscribed to two features having the
  same triggering events and different results

    • Features pre-conditions

      - Pre-conditions of [cfb, 1]
          subs(bob, cfb)
          concerns(bob, cfb)
          cfb(carole)
          busy(bob)

      - Pre-conditions of [cw, 4]
          subs(bob, cw)
          concerns(bob, cw)
          busy(bob)
          talk(bob, carole)
          hold(bob, dave)

    • Same triggering events
        call(alice, bob)

    • Different results

      - Resulting events of [cfb, 1]
          redirected(alice, carole)
          call(alice, carole)
          ring(alice, carole)

      - Resulting events of [cw, 4]
          busy_ind(bob, alice)
```

```
• Rule #d1 -> [cfb, 1] & [arc, 1]
  Same user subscribed to two features having the
  same triggering events and different results

    • Features pre-conditions

      - Pre-conditions of [cfb, 1]
          subs(bob, cfb)
          concerns(bob, cfb)
          cfb(carole)
          busy(bob)

      - Pre-conditions of [arc, 1]
          subs(bob, arc)
          concerns(bob, arc)
          busy(bob)

    • Same triggering events
        call(alice, bob)

    • Different results

      - Resulting events of [cfb, 1]
          redirected(alice, carole)
          call(alice, carole)
          ring(alice, carole)

      - Resulting events of [arc, 1]
          store(bob, alice)
          arc_notify(alice)
```

- Rule #d1 -> [cfa, 1] & [ics, 1]
  Same user subscribed to two features having the same triggering events and different results
  - Features pre-conditions
    - Pre-conditions of [cfa, 1]
      subs(bob, cfa)
      concerns(bob, cfa)
      cfa(carole)
    - Pre-conditions of [ics, 1]
      subs(bob, ics)
      concerns(bob, ics)
      ics_list(alice)
  - Same triggering events
    call(alice, bob)
  - Different results
    - Resulting events of [cfa, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)
    - Resulting events of [ics, 1]
      deny_call(bob, alice)
      call_denied(bob, alice)

- Rule #d1 -> [cfa, 1] & [cw, 1]
  Same user subscribed to two features having the same triggering events and different results
  - Features pre-conditions
    - Pre-conditions of [cfa, 1]
      subs(bob, cfa)
      concerns(bob, cfa)
      cfa(carole)
    - Pre-conditions of [cw, 1]
      subs(bob, cw)
      concerns(bob, cw)
      busy(bob)
      talk(bob, carole)
  - Same triggering events
    call(alice, bob)
  - Different results
    - Resulting events of [cfa, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)
    - Resulting events of [cw, 1]
      hold(bob, alice)
      cw_notify(bob)

- Rule #d1 -> [cfa, 1] & [cw, 4]
  Same user subscribed to two features having the same triggering events and different results
  - Features pre-conditions
    - Pre-conditions of [cfa, 1]
      subs(bob, cfa)
      concerns(bob, cfa)
      cfa(carole)
    - Pre-conditions of [cw, 4]
      subs(bob, cw)
      concerns(bob, cw)
      busy(bob)
      talk(bob, carole)
      hold(bob, dave)
  - Same triggering events
    call(alice, bob)
  - Different results
    - Resulting events of [cfa, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)
    - Resulting events of [cw, 4]
      busy_ind(bob, alice)

- Rule #d1 -> [cfa, 1] & [arc, 1]
  Same user subscribed to two features having the same triggering events and different results
  - Features pre-conditions
    - Pre-conditions of [cfa, 1]
      subs(bob, cfa)
      concerns(bob, cfa)
      cfa(carole)
    - Pre-conditions of [arc, 1]
      subs(bob, arc)
      concerns(bob, arc)
      busy(bob)
  - Same triggering events
    call(alice, bob)
  - Different results
    - Resulting events of [cfa, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)
    - Resulting events of [arc, 1]
      store(bob, alice)
      arc_notify(alice)

- Rule #d1 -> [ocs, 1] & [ct, 2]
  Same user subscribed to two features having the same triggering events and different results
  - Features pre-conditions
    - Pre-conditions of [ocs, 1]
      subs(alice, ocs)
      concerns(alice, ocs)
      ocs_list(bob)
    - Pre-conditions of [ct, 2]
      subs(alice, ct)
      concerns(alice, ct)
      busy(alice)
      hold(alice, carole)
      dialtone(alice)
  - Same triggering events
    call(alice, bob)
  - Different results
    - Resulting events of [ocs, 1]
      block_call(alice, bob)
      call_blocked(alice, alice)
    - Resulting events of [ct, 2]
      talk(alice, bob)

- Rule #d1 -> [ct, 1] & [cw, 2]
  Same user subscribed to two features having the same triggering events and different results
  - Features pre-conditions
    - Pre-conditions of [ct, 1]
      subs(alice, ct)
      concerns(alice, ct)
      busy(alice)
      talk(alice, bob)
    - Pre-conditions of [cw, 2]
      subs(alice, cw)
      concerns(alice, cw)
      busy(alice)
      talk(alice, bob)
      hold(alice, carole)
  - Same triggering events
    flash(alice)
  - Different results
    - Resulting events of [ct, 1]
      hold(alice, bob)
      dialtone(alice)
    - Resulting events of [cw, 2]
      hold(alice, bob)
      talk(alice, carole)

- Rule #d1 -> [ics, 1] & [arc, 1]
  Same user subscribed to two features having the same triggering events and different results
  - Features pre-conditions
    - Pre-conditions of [ics, 1]
      subs(bob, ics)
      concerns(bob, ics)
      ics_list(alice)
    - Pre-conditions of [arc, 1]
      subs(bob, arc)
      concerns(bob, arc)
      busy(bob)
  - Same triggering events
    call(alice, bob)
  - Different results
    - Resulting events of [ics, 1]
      deny_call(bob, alice)
      call_denied(bob, alice)
    - Resulting events of [arc, 1]
      store(bob, alice)
      arc_notify(alice)

- Rule #d1 -> [cw, 1] & [arc, 1]
  Same user subscribed to two features having the same triggering events and different results
  - Features pre-conditions
    - Pre-conditions of [cw, 1]
      subs(alice, cw)
      concerns(alice, cw)
      busy(alice)
      talk(alice, bob)
    - Pre-conditions of [arc, 1]
      subs(alice, arc)
      concerns(alice, arc)
      busy(alice)
  - Same triggering events
    call(carole, alice)
  - Different results
    - Resulting events of [cw, 1]
      hold(alice, carole)
      cw_notify(alice)
    - Resulting events of [arc, 1]
      store(alice, carole)
      arc_notify(carole)

- Rule 8d1 -> [cw, 4] & [arc, 1]
  Same user subscribed to two features having the same triggering events and different results
  - Features pre-conditions
    - Pre-conditions of [cw, 4]
      subs(alice, cw)
      concerns(alice, cw)
      busy(alice)
      talk(alice, bob)
      hold(alice, carole)
    - Pre-conditions of [arc, 1]
      subs(alice, arc)
      concerns(alice, arc)
      busy(alice)
  - Same triggering events
    call(dave, alice)
  - Different results
    - Resulting events of [cw, 4]
      busy_ind(alice, dave)
    - Resulting events of [arc, 1]
      store(alice, dave)
      arc_notify(dave)

- Rule 8d2 -> [cfb, 1] & [cfa, 1]
  Same user subscribed to two features having the same triggering events and contradictory results
  - Features pre-conditions
    - Pre-conditions of [cfb, 1]
      subs(bob, cfb)
      concerns(bob, cfb)
      cfb(carole)
      busy(bob)
    - Pre-conditions of [cfa, 1]
      subs(bob, cfa)
      concerns(bob, cfa)
      cfa(dave)
  - Same triggering events
    call(alice, bob)
  - Contradictory results
    - Resulting events of [cfb, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)
    - Resulting events of [cfa, 1]
      redirected(alice, dave)
      call(alice, dave)
      ring(alice, dave)

- Rule 8d2 -> [ics, 1] & [cw, 1]
  Same user subscribed to two features having the same triggering events and contradictory results
  - Features pre-conditions
    - Pre-conditions of [ics, 1]
      subs(bob, ics)
      concerns(bob, ics)
      ics_list(alice)
    - Pre-conditions of [cw, 1]
      subs(bob, cw)
      concerns(bob, cw)
      busy(bob)
      talk(bob, carole)
  - Same triggering events
    call(alice, bob)
  - Contradictory results
    - Resulting events of [ics, 1]
      deny_call(bob, alice)
      call_denied(bob, alice)
    - Resulting events of [cw, 1]
      hold(bob, alice)
      cw_notify(bob)

- Rule 8d2 -> [ics, 1] & [cw, 4]
  Same user subscribed to two features having the same triggering events and contradictory results
  - Features pre-conditions
    - Pre-conditions of [ics, 1]
      subs(bob, ics)
      concerns(bob, ics)
      ics_list(alice)
    - Pre-conditions of [cw, 4]
      subs(bob, cw)
      concerns(bob, cw)
      busy(bob)
      talk(bob, carole)
      hold(bob, dave)
  - Same triggering events
    call(alice, bob)
  - Contradictory results
    - Resulting events of [ics, 1]
      deny_call(bob, alice)
      call_denied(bob, alice)
    - Resulting events of [cw, 4]
      busy_ind(bob, alice)

- Rule #d3 -> [cw, 3] & [cfb, 1]
  Different users subscribed to two features where feature 1 concerns subscriber of feature 2, and the features have the same triggering events and different results
  - Features pre-conditions
    - Pre-conditions of [cw, 3]
      subs(alice, cw)
      concerns(carole, cw)
      busy(alice)
      talk(alice, bob)
      hold(alice, carole)
    - Pre-conditions of [cfb, 1]
      subs(carole, cfb)
      concerns(carole, cfb)
      cfb(alice)
      busy(carole)
  - Same triggering events
    call(dave, carole)
  - Different results
    - Resulting events of [cw, 3]
      busy_ind(carole, dave)
    - Resulting events of [cfb, 1]
      redirected(dave, alice)
      call(dave, alice)
      ring(dave, alice)

- Rule #d3 -> [cw, 3] & [cfa, 1]
  Different users subscribed to two features where feature 1 concerns subscriber of feature 2, and the features have the same triggering events and different results
  - Features pre-conditions
    - Pre-conditions of [cw, 3]
      subs(alice, cw)
      concerns(carole, cw)
      busy(alice)
      talk(alice, bob)
      hold(alice, carole)
    - Pre-conditions of [cfa, 1]
      subs(carole, cfa)
      concerns(carole, cfa)
      cfa(alice)
  - Same triggering events
    call(dave, carole)
  - Different results
    - Resulting events of [cw, 3]
      busy_ind(carole, dave)
    - Resulting events of [cfa, 1]
      redirected(dave, alice)
      call(dave, alice)
      ring(dave, alice)

- Rule #d3 -> [cw, 3] & [cw, 1]
  Different users subscribed to two features where feature 1 concerns subscriber of feature 2, and the features have the same triggering events and different results
  - Features pre-conditions
    - Pre-conditions of [cw, 3]
      subs(alice, cw)
      concerns(carole, cw)
      busy(alice)
      talk(alice, bob)
      hold(alice, carole)
    - Pre-conditions of [cw, 1]
      subs(carole, cw)
      concerns(carole, cw)
      busy(carole)
      talk(carole, alice)
  - Same triggering events
    call(dave, carole)
  - Different results
    - Resulting events of [cw, 3]
      busy_ind(carole, dave)
    - Resulting events of [cw, 1]
      hold(carole, dave)
      cw_notify(carole)

- Rule #d3 -> [cw, 3] & [arc, 1]
  Different users subscribed to two features where feature 1 concerns subscriber of feature 2, and the features have the same triggering events and different results
  - Features pre-conditions
    - Pre-conditions of [cw, 3]
      subs(alice, cw)
      concerns(carole, cw)
      busy(alice)
      talk(alice, bob)
      hold(alice, carole)
    - Pre-conditions of [arc, 1]
      subs(carole, arc)
      concerns(carole, arc)
      busy(carole)
  - Same triggering events
    call(dave, carole)
  - Different results
    - Resulting events of [cw, 3]
      busy_ind(carole, dave)
    - Resulting events of [arc, 1]
      store(carole, dave)
      arc_notify(dave)

- Rule #d4 -> [cw, 3] & [ics, 1]
  Different users subscribed to two features where feature 1 concerns subscriber of feature 2, and the features have the same triggering events and contradictory results
  - Features pre-conditions
    - Pre-conditions of [cw, 3]
      subs(alice, cw)
      concerns(carole, cw)
      busy(alice)
      talk(alice, bob)
      hold(alice, carole)
    - Pre-conditions of [ics, 1]
      subs(carole, ics)
      concerns(carole, ics)
      ics_list(dave)
  - Same triggering events
    call(dave, carole)
  - Contradictory results
    - Resulting events of [cw, 3]
      busy_ind(carole, dave)
    - Resulting events of [ics, 1]
      deny_call(carole, dave)
      call_denied(carole, dave)

- Rule #t1 -> [cfb, 1] & [cfb, 1]
  The user(s) subscribed to different features and the results of F1 include the triggering events of F2 but their results are contradictory
  - Different pre-conditions
    - Pre-conditions of [cfb, 1]
      subs(bob, cfb)
      concerns(bob, cfb)
      cfb(carole)
      busy(bob)
    - Pre-conditions of [cfb, 1]
      subs(carole, cfb)
      concerns(carole, cfb)
      cfb(bob)
      busy(carole)
  - Transitivity between features
    - Triggering events of [cfb, 1]
      call(alice, bob)
    - Resulting events of [cfb, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)
    - Including triggering events of [cfb, 1]
      call(alice, carole)
  - Contradictory results
    - Resulting events of [cfb, 1]
      redirected(alice, bob)
      call(alice, bob)
      ring(alice, bob)
    - Resulting events of [cfb, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)

- Rule #t1 -> [cfb, 1] & [cfa, 1]
  The user(s) subscribed to different features and the results of F1 include the triggering events of F2 but their results are contradictory
  - Different pre-conditions
    - Pre-conditions of [cfb, 1]
      subs(bob, cfb)
      concerns(bob, cfb)
      cfb(carole)
      busy(bob)
    - Pre-conditions of [cfa, 1]
      subs(carole, cfa)
      concerns(carole, cfa)
      cfa(bob)
  - Transitivity between features
    - Triggering events of [cfb, 1]
      call(alice, bob)
    - Resulting events of [cfb, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)
    - Including triggering events of [cfa, 1]
      call(alice, carole)
  - Contradictory results
    - Resulting events of [cfa, 1]
      redirected(alice, bob)
      call(alice, bob)
      ring(alice, bob)
    - Resulting events of [cfb, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)

- Rule #t1 -> [cfb, 1] & [ocs, 1]
  The user(s) subscribed to different features and the results of F1 include the triggering events of F2 but their results are contradictory
  - Different pre-conditions
    - Pre-conditions of [cfb, 1]
      subs(bob, cfb)
      concerns(bob, cfb)
      cfb(carole)
      busy(bob)
    - Pre-conditions of [ocs, 1]
      subs(alice, ocs)
      concerns(alice, ocs)
      ocs_list(carole)
  - Transitivity between features
    - Triggering events of [cfb, 1]
      call(alice, bob)
    - Resulting events of [cfb, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)
    - Including triggering events of [ocs, 1]
      call(alice, carole)
  - Contradictory results
    - Resulting events of [ocs, 1]
      block_call(alice, carole)
      call_blocked(alice, alice)
    - Resulting events of [cfb, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)

- Rule St1 -> [cfb, 1] & [ics, 1]
  The user(s) subscribed to different features and the results of F1 include the triggering events of F2 but their results are contradictory

  + Different pre-conditions

    - Pre-conditions of [cfb, 1]
      subs(bob, cfb)
      concerns(bob, cfb)
      cfb(carole)
      busy(bob)

    - Pre-conditions of [ics, 1]
      subs(carole, ics)
      concerns(carole, ics)
      ics_list(alice)

  + Transitivity between features

    - Triggering events of [cfb, 1]
      call(alice, bob)

    - Resulting events of [cfb, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)

    - Including triggering events of [ics, 1]
      call(alice, carole)

  + Contradictory results

    - Resulting events of [ics, 1]
      deny_call(carole, alice)
      call_denied(carole, alice)

    - Resulting events of [cfb, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)

- Rule St1 -> [cfa, 1] & [cfb, 1]
  The user(s) subscribed to different features and the results of F1 include the triggering events of F2 but their results are contradictory

  + Different pre-conditions

    - Pre-conditions of [cfa, 1]
      subs(bob, cfa)
      concerns(bob, cfa)
      cfa(carole)

    - Pre-conditions of [cfb, 1]
      subs(carole, cfb)
      concerns(carole, cfb)
      cfb(bob)
      busy(carole)

  + Transitivity between features

    - Triggering events of [cfa, 1]
      call(alice, bob)

    - Resulting events of [cfa, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)

    - Including triggering events of [cfb, 1]
      call(alice, carole)

  + Contradictory results

    - Resulting events of [cfb, 1]
      redirected(alice, bob)
      call(alice, bob)
      ring(alice, bob)

    - Resulting events of [cfa, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)

- Rule St1 -> [cfa, 1] & [cfa, 1]
  The user(s) subscribed to different features and the results of F1 include the triggering events of F2 but their results are contradictory

  + Different pre-conditions

    - Pre-conditions of [cfa, 1]
      subs(bob, cfa)
      concerns(bob, cfa)
      cfa(carole)

    - Pre-conditions of [cfa, 1]
      subs(carole, cfa)
      concerns(carole, cfa)
      cfa(bob)

  + Transitivity between features

    - Triggering events of [cfa, 1]
      call(alice, bob)

    - Resulting events of [cfa, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)

    - Including triggering events of [cfa, 1]
      call(alice, carole)

  + Contradictory results

    - Resulting events of [cfa, 1]
      redirected(alice, bob)
      call(alice, bob)
      ring(alice, bob)

    - Resulting events of [cfa, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)

- Rule St1 -> [cfa, 1] & [ocs, 1]
  The user(s) subscribed to different features and the results of F1 include the triggering events of F2 but their results are contradictory

  + Different pre-conditions

    - Pre-conditions of [cfa, 1]
      subs(bob, cfa)
      concerns(bob, cfa)
      cfa(carole)

    - Pre-conditions of [ocs, 1]
      subs(alice, ocs)
      concerns(alice, ocs)
      ocs_list(carole)

  + Transitivity between features

    - Triggering events of [cfa, 1]
      call(alice, bob)

    - Resulting events of [cfa, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)

    - Including triggering events of [ocs, 1]
      call(alice, carole)

  + Contradictory results

    - Resulting events of [ocs, 1]
      block_call(alice, carole)
      call_blocked(alice, alice)

    - Resulting events of [cfa, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)

```
• Rule #t1 -> [cfa, 1] & [ics, 1]
  The user(s) subscribed to different features and the
  results of F1 include the triggering events of F2
  but their results are contradictory

    • Different pre-conditions

      - Pre-conditions of [cfa, 1]
          subs(bob, cfa)
          concerns(bob, cfa)
          cfa(carole)

      - Pre-conditions of [ics, 1]
          subs(carole, ics)
          concerns(carole, ics)
          ics_list(alice)

    • Transitivity between features

      - Triggering events of [cfa, 1]
          call(alice, bob)

      - Resulting events of [cfa, 1]
          redirected(alice, carole)
          call(alice, carole)
          ring(alice, carole)

      - Including triggering events of [ics, 1]
          call(alice, carole)

    • Contradictory results

      - Resulting events of [ics, 1]
          deny_call(carole, alice)
          call_denied(carole, alice)

      - Resulting events of [cfa, 1]
          redirected(alice, carole)
          call(alice, carole)
          ring(alice, carole)
```

```
• Rule #t1 -> [ct, 3] & [cfb, 1]
  The user(s) subscribed to different features and the
  results of F1 include the triggering events of F2
  but their results are contradictory

    • Different pre-conditions

      - Pre-conditions of [ct, 3]
          subs(alice, ct)
          concerns(alice, ct)
          busy(alice)
          hold(alice, bob)
          talk(alice, carole)

      - Pre-conditions of [cfb, 1]
          subs(carole, cfb)
          concerns(carole, cfb)
          cfb(alice)
          busy(carole)

    • Transitivity between features

      - Triggering events of [ct, 3]
          transfer(alice)

      - Resulting events of [ct, 3]
          call(bob, carole)

      - Including triggering events of [cfb, 1]
          call(bob, carole)

    • Contradictory results

      - Resulting events of [cfb, 1]
          redirected(bob, alice)
          call(bob, alice)
          ring(bob, alice)

      - Resulting events of [ct, 3]
          call(bob, carole)
```

```
• Rule #t1 -> [ct, 3] & [cfa, 1]
  The user(s) subscribed to different features and the
  results of F1 include the triggering events of F2
  but their results are contradictory

    • Different pre-conditions

      - Pre-conditions of [ct, 3]
          subs(alice, ct)
          concerns(alice, ct)
          busy(alice)
          hold(alice, bob)
          talk(alice, carole)

      - Pre-conditions of [cfa, 1]
          subs(carole, cfa)
          concerns(carole, cfa)
          cfa(alice)

    • Transitivity between features

      - Triggering events of [ct, 3]
          transfer(alice)

      - Resulting events of [ct, 3]
          call(bob, carole)

      - Including triggering events of [cfa, 1]
          call(bob, carole)

    • Contradictory results

      - Resulting events of [cfa, 1]
          redirected(bob, alice)
          call(bob, alice)
          ring(bob, alice)

      - Resulting events of [ct, 3]
          call(bob, carole)
```

```
• Rule #t1 -> [ct, 3] & [ocs, 1]
  The user(s) subscribed to different features and the
  results of F1 include the triggering events of F2
  but their results are contradictory

    • Different pre-conditions

      - Pre-conditions of [ct, 3]
          subs(alice, ct)
          concerns(alice, ct)
          busy(alice)
          hold(alice, bob)
          talk(alice, carole)

      - Pre-conditions of [ocs, 1]
          subs(bob, ocs)
          concerns(bob, ocs)
          ocs_list(carole)

    • Transitivity between features

      - Triggering events of [ct, 3]
          transfer(alice)

      - Resulting events of [ct, 3]
          call(bob, carole)

      - Including triggering events of [ocs, 1]
          call(bob, carole)

    • Contradictory results

      - Resulting events of [ocs, 1]
          block_call(bob, carole)
          call_blocked(bob, bob)

      - Resulting events of [ct, 3]
          call(bob, carole)
```

- Rule St1 -> [ct, 3] & [ics, 1]
  The user(s) subscribed to different features and the results of F1 include the triggering events of F2 but their results are contradictory

  - Different pre-conditions

    - Pre-conditions of [ct, 3]
      subs(alice, ct)
      concerns(alice, ct)
      busy(alice)
      hold(alice, bob)
      talk(alice, carole)

    - Pre-conditions of [ics, 1]
      subs(carole, ics)
      concerns(carole, ics)
      ics_list(bob)

  - Transitivity between features

    - Triggering events of [ct, 3]
      transfer(alice)

    - Resulting events of [ct, 3]
      call(bob, carole)

    - Including triggering events of [ics, 1]
      call(bob, carole)

  - Contradictory results

    - Resulting events of [ics, 1]
      deny_call(carole, bob)
      call_denied(carole, bob)

    - Resulting events of [ct, 3]
      call(bob, carole)

- Rule St1 -> [cp, 1] & [cfa, 1]
  The user(s) subscribed to different features and the results of F1 include the triggering events of F2 but their results are contradictory

  - Different pre-conditions

    - Pre-conditions of [cp, 1]
      subs(bob, cp)
      concerns(bob, cp)
      call(alice, carole)
      ring(alice, carole)
      idle(bob)

    - Pre-conditions of [cfa, 1]
      subs(bob, cfa)
      concerns(bob, cfa)
      cfa(carole)

  - Transitivity between features

    - Triggering events of [cp, 1]
      pickup(bob)

    - Resulting events of [cp, 1]
      picked_up(alice, bob)
      call(alice, bob)

    - Including triggering events of [cfa, 1]
      call(alice, bob)

  - Contradictory results

    - Resulting events of [cfa, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)

    - Resulting events of [cp, 1]
      picked_up(alice, bob)
      call(alice, bob)

- Rule St1 -> [cp, 1] & [ocs, 1]
  The user(s) subscribed to different features and the results of F1 include the triggering events of F2 but their results are contradictory

  - Different pre-conditions

    - Pre-conditions of [cp, 1]
      subs(bob, cp)
      concerns(bob, cp)
      call(alice, carole)
      ring(alice, carole)
      idle(bob)

    - Pre-conditions of [ocs, 1]
      subs(alice, ocs)
      concerns(alice, ocs)
      ocs_list(bob)

  - Transitivity between features

    - Triggering events of [cp, 1]
      pickup(bob)

    - Resulting events of [cp, 1]
      picked_up(alice, bob)
      call(alice, bob)

    - Including triggering events of [ocs, 1]
      call(alice, bob)

  - Contradictory results

    - Resulting events of [ocs, 1]
      block_call(alice, bob)
      call_blocked(alice, alice)

    - Resulting events of [cp, 1]
      picked_up(alice, bob)
      call(alice, bob)

- Rule St1 -> [cp, 1] & [ics, 1]
  The user(s) subscribed to different features and the results of F1 include the triggering events of F2 but their results are contradictory

  - Different pre-conditions

    - Pre-conditions of [cp, 1]
      subs(bob, cp)
      concerns(bob, cp)
      call(alice, carole)
      ring(alice, carole)
      idle(bob)

    - Pre-conditions of [ics, 1]
      subs(bob, ics)
      concerns(bob, ics)
      ics_list(alice)

  - Transitivity between features

    - Triggering events of [cp, 1]
      pickup(bob)

    - Resulting events of [cp, 1]
      picked_up(alice, bob)
      call(alice, bob)

    - Including triggering events of [ics, 1]
      call(alice, bob)

  - Contradictory results

    - Resulting events of [ics, 1]
      deny_call(bob, alice)
      call_denied(bob, alice)

    - Resulting events of [cp, 1]
      picked_up(alice, bob)
      call(alice, bob)

- Rule #1 -> [arc, 3] & [cfb, 1]
  The user(s) subscribed to different features and the results of F1 include the triggering events of F2 but their results are contradictory
  - Different pre-conditions
    - Pre-conditions of [arc, 3]
      subs(bob, arc)
      concerns(bob, arc)
      busy(bob)
      stored(alice)
      arc_ring(bob)
    - Pre-conditions of [cfb, 1]
      subs(alice, cfb)
      concerns(alice, cfb)
      cfb(carole)
      busy(alice)
  - Transitivity between features
    - Triggering events of [arc, 3]
      offhook(bob)
    - Resulting events of [arc, 3]
      call(bob, alice)
    - Including triggering events of [cfb, 1]
      call(bob, alice)
  - Contradictory results
    - Resulting events of [cfb, 1]
      redirected(bob, carole)
      call(bob, carole)
      ring(bob, carole)
    - Resulting events of [arc, 3]
      call(bob, alice)

- Rule #1 -> [arc, 3] & [cfa, 1]
  The user(s) subscribed to different features and the results of F1 include the triggering events of F2 but their results are contradictory
  - Different pre-conditions
    - Pre-conditions of [arc, 3]
      subs(bob, arc)
      concerns(bob, arc)
      busy(bob)
      stored(alice)
      arc_ring(bob)
    - Pre-conditions of [cfa, 1]
      subs(alice, cfa)
      concerns(alice, cfa)
      cfa(carole)
  - Transitivity between features
    - Triggering events of [arc, 3]
      offhook(bob)
    - Resulting events of [arc, 3]
      call(bob, alice)
    - Including triggering events of [cfa, 1]
      call(bob, alice)
  - Contradictory results
    - Resulting events of [cfa, 1]
      redirected(bob, carole)
      call(bob, carole)
      ring(bob, carole)
    - Resulting events of [arc, 3]
      call(bob, alice)

- Rule #1 -> [arc, 3] & [ocs, 1]
  The user(s) subscribed to different features and the results of F1 include the triggering events of F2 but their results are contradictory
  - Different pre-conditions
    - Pre-conditions of [arc, 3]
      subs(bob, arc)
      concerns(bob, arc)
      busy(bob)
      stored(alice)
      arc_ring(bob)
    - Pre-conditions of [ocs, 1]
      subs(bob, ocs)
      concerns(bob, ocs)
      ocs_list(alice)
  - Transitivity between features
    - Triggering events of [arc, 3]
      offhook(bob)
    - Resulting events of [arc, 3]
      call(bob, alice)
    - Including triggering events of [ocs, 1]
      call(bob, alice)
  - Contradictory results
    - Resulting events of [ocs, 1]
      block_call(bob, alice)
      call_blocked(bob, bob)
    - Resulting events of [arc, 3]
      call(bob, alice)

- Rule #1 -> [arc, 3] & [ics, 1]
  The user(s) subscribed to different features and the results of F1 include the triggering events of F2 but their results are contradictory
  - Different pre-conditions
    - Pre-conditions of [arc, 3]
      subs(bob, arc)
      concerns(bob, arc)
      busy(bob)
      stored(alice)
      arc_ring(bob)
    - Pre-conditions of [ics, 1]
      subs(alice, ics)
      concerns(alice, ics)
      ics_list(bob)
  - Transitivity between features
    - Triggering events of [arc, 3]
      offhook(bob)
    - Resulting events of [arc, 3]
      call(bob, alice)
    - Including triggering events of [ics, 1]
      call(bob, alice)
  - Contradictory results
    - Resulting events of [ics, 1]
      deny_call(alice, bob)
      call_denied(alice, bob)
    - Resulting events of [arc, 3]
      call(bob, alice)

- Rule st2 -> [cfb, 1] & [cfb, 1]
  The users subscribed to different features for which the resulting events of F1 include triggering events of F2 and vice-versa, which leads to a loop.
  - Different or common pre-conditions
    - Pre-conditions of [cfb, 1]
      subs(bob, cfb)
      concerns(bob, cfb)
      cfb(carole)
      busy(bob)
    - Pre-conditions of [cfb, 1]
      subs(carole, cfb)
      concerns(carole, cfb)
      cfb(bob)
      busy(carole)
  - Transitivity between features
    - Resulting events of [cfb, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)
    - Including triggering events of [cfb, 1]
      call(alice, carole)
  - Reverse transitivity, leading to a loop
    - Resulting events of [cfb, 1]
      redirected(alice, bob)
      call(alice, bob)
      ring(alice, bob)
    - Including triggering events of [cfb, 1]
      call(alice, bob)

- Rule st2 -> [cfa, 1] & [cfa, 1]
  The users subscribed to different features for which the resulting events of F1 include triggering events of F2 and vice-versa, which leads to a loop.
  - Different or common pre-conditions
    - Pre-conditions of [cfa, 1]
      subs(bob, cfa)
      concerns(bob, cfa)
      cfa(carole)
    - Pre-conditions of [cfa, 1]
      subs(carole, cfa)
      concerns(carole, cfa)
      cfa(bob)
  - Transitivity between features
    - Resulting events of [cfa, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)
    - Including triggering events of [cfa, 1]
      call(alice, carole)
  - Reverse transitivity, leading to a loop
    - Resulting events of [cfa, 1]
      redirected(alice, bob)
      call(alice, bob)
      ring(alice, bob)
    - Including triggering events of [cfa, 1]
      call(alice, bob)

- Rule st2 -> [cfb, 1] & [cfa, 1]
  The users subscribed to different features for which the resulting events of F1 include triggering events of F2 and vice-versa, which leads to a loop.
  - Different or common pre-conditions
    - Pre-conditions of [cfb, 1]
      subs(bob, cfb)
      concerns(bob, cfb)
      cfb(carole)
      busy(bob)
    - Pre-conditions of [cfa, 1]
      subs(carole, cfa)
      concerns(carole, cfa)
      cfa(bob)
  - Transitivity between features
    - Resulting events of [cfb, 1]
      redirected(alice, carole)
      call(alice, carole)
      ring(alice, carole)
    - Including triggering events of [cfa, 1]
      call(alice, carole)
  - Reverse transitivity, leading to a loop
    - Resulting events of [cfa, 1]
      redirected(alice, bob)
      call(alice, bob)
      ring(alice, bob)
    - Including triggering events of [cfb, 1]
      call(alice, bob)

Appendix B

Prolog Implementation

```
%-------------------------------------------------------------------------
% Module:     fi-lookup-rules-v15
% Version:    15
% Modified:   Wed Sep 20 10:24:08 EST 2000
% Author:     Nicolas Gorse <ngorse@site.uottawa.ca>
%
%
%  Usage: fi-lookup <command>
%       * where command must be one of the following:
%       -filtering <features> [options]
%       --> find possible incoherences
%           * where options include:
%           -g <file> --- feature sets to combine
%           -k <file> --- list of already known incoherences
%           -i <file> --- file to which reports will be written
%
%       -testiso <features> <mapping-rules> [options]
%       --> produces test suites for features in isolation
%           * where options include:
%           -t <file> --- file to which test suites will be written
%
%       -testfi <mapping-rules> <known incoherences> [options]
%       --> produces test suites for given incoherences
%           * where options include:
%           -t <file> --- file to which test suites will be written
%
%       -filttest <features> <mapping-rules> [options]
%       --> finds possible incoherences & produce test suites
%           * where options include:
%           -g <file> --- feature sets to combine
%           -k <file> --- list of already known incoherences
%           -i <file> --- file to which reports will be written
%           -t <file> --- file to which test suites will be written
```

```
%
% Comments: This version is the last stable one. It contains a small
%           interface allowing interactivity with the user using the
%           command line arguments allowing also to call it from batch
%           scripts.
%-------------------------------------------------------------------------


                    %************************************%


%----------* Basic predicate definitions *---------------------------------


% Rules predicates, basicaly used to speed up the program,
% each predicate correspond to one of the six incoherence rules.
%
rule(d1).
rule(d2).
rule(d3).
rule(d4).
rule(t1).
rule(t2).


% Predicate used to convert a list of elements to a correponding
% string. The elements are concatenated from head to tail.
%
list_to_string([], String, String).

list_to_string([Elem|List], InString, OutString):-
string_concat(InString, Elem, TmpString),
list_to_string(List, TmpString, OutString), !.


% Predicate used to write time informations about the starting time
% and stopping time of an analysis. Useful to collect statistics.
%
write_time(start):-
get_time(T), convert_time(T, Yr, Mth, Day, Hr, Min, Sec, MSec),
write('\t--- start time: '),
write(Day), write('/'), write(Mth), write('/'), write(Yr),
write(' at '), write(Hr), write(':'), write(Min), write(':'),
write(Sec), write('.'), write(MSec).
```

```
write_time(stop):-
get_time(T), convert_time(T, Yr, Mth, Day, Hr, Min, Sec, MSec),
write('\t--- stop time : '),
write(Day), write('/'), write(Mth), write('/'), write(Yr),
write(' at '), write(Hr), write(':'), write(Min), write(':'),
write(Sec), write('.'), write(MSec).


% Predicate used to output a list to the screen. It prints the
% specified "tabulation" characters followed by an element.
% The elements are printed one by one from head to tail.
%
writeSet(_, []).

writeSet(TAB, [Elem|SubSet]):-
write(TAB), write(Elem), nl,
writeSet(TAB, SubSet), !.


% Predicate used to build the sets of features to be compared.
% It takes the definitions of all features and builds sets that
% contain the complete list of features.
%
build_feature_groups:-
feature(X, _, _, _),
not(lfeature(X)),
assert(lfeature(X)),
assert(rfeature(X)),
build_feature_groups, !.

build_feature_groups:-
feature(X, _, _, _), lfeature(X).

%----------* Basic predicate definitions *---------------------------------


                    %**************************************%


%----------* Arguments predicate definitions *-----------------------------

% Predicate managing the case of empty arguments command line
%
scan_args([]):-
print_usage.
```

```
% Predicate managing the argument concerning the incoherences
% analysis without test suites generation
%
scan_args(['-filtering', Features|Options]):-
exists_file(Features),
consult(Features),
scan_fi_options(Options), (
  fi_output_file(Output),
  open(Output, write, Stream),
  set_output(Stream)
           ;
          not(fi_output_file(_)),
  stdout(Stream), set_output(Stream)
        ), (
  groups_loaded(yes), time(fi_search), halt
           ;
  not(groups_loaded(yes)),
  build_feature_groups, time(fi_search), halt
        )
 ;
not(exists_file(Features)),
write('\t--- Invalid file: '),
write(Features), nl, halt.


% Predicate managing the argument concerning the test suites
% generation for features in isolation
%
scan_args(['-testiso', Features, Mapping|Options]):-
exists_file(Features),
consult(Features), (
  exists_file(Mapping),
  consult(Mapping),
  scan_iso_options(Options), (
    iso_output_file(Output),
    open(Output, write, Stream),
    set_output(Stream)
             ;
            not(iso_output_file(_)),
    stdout(Stream), set_output(Stream)
         ), iso_gen_test, halt
   ;
  not(exists_file(Mapping)),
```

```
  write('\t--- Invalid file: '),
  write(Mapping), nl, halt
        )
 ;
not(exists_file(Features)),
write('\t--- Invalid file: '),
write(Features), nl, halt.


% Predicate managing the argument concerning the test suites
% generation without incoherence analysis
%
scan_args(['-testfi', Mapping, Known|Options]):-
exists_file(Mapping),
consult(Mapping), (
  exists_file(Known),
  consult(Known),
  scan_ts_options(Options), (
    ts_output_file(Output),
    open(Output, write, Stream),
    set_output(Stream)
             ;
            not(ts_output_file(_)),
    stdout(Stream), set_output(Stream)
         ), fi_gen_test, halt
   ;
  not(exists_file(Known)),
  write('\t--- Invalid file: '),
  write(Known), nl, halt
        )
 ;
not(exists_file(Mapping)),
write('\t--- Invalid file: '),
write(Mapping), nl, halt.


% Predicate managing the argument concerning the incoherences
% analysis followed by the test suites generation
%
scan_args(['-filtest', Features, Mapping|Options]):-
exists_file(Features),
consult(Features), (
  exists_file(Mapping),
  consult(Mapping),
  scan_ft_options(Options), (
```

```
    groups_loaded(yes)
            ;
    not(groups_loaded(yes)),
    build_feature_groups
          ), (
    fi_output_file(FiOutput),
    open(FiOutput, write, FiStream),
    set_output(FiStream)
            ;
           not(fi_output_file(_)),
    stdout(FiStream), set_output(FiStream)
          ), (
    groups_loaded(yes), nl, time(fi_search), !
            ;
    not(groups_loaded(yes)),
    build_feature_groups, nl, time(fi_search), !
          ), (
    ts_output_file(TsOutput),
    open(TsOutput, write, TsStream),
    set_output(TsStream)
            ;
           not(ts_output_file(_)),
    stdout(TsStream), set_output(TsStream)
          ),
  fi_gen_test,
  halt
           ;
  not(exists_file(Mapping)),
  write('\t--- Invalid file: '),
  write(Mapping), nl, halt
        )
 ;
not(exists_file(Features)),
write('\t--- Invalid file: '),
write(Features), nl, halt.


% Predicate managing all other cases; the bad arguments
%
scan_args(_):-
print_usage.


% Predicates managing sub-options relatives to the Incoherences
% analysis.
```

```
%
scan_fi_options([]).

scan_fi_options(['-g', Groups|Options]):-
exists_file(Groups),
consult(Groups),
assert(groups_loaded(yes)),
scan_fi_options(Options)
 ;
not(exists_file(Groups)),
write('\t--- Invalid file: '),
write(Groups), nl, halt.

scan_fi_options(['-k', Known|Options]):-
exists_file(Known),
consult(Known),
scan_fi_options(Options)
 ;
not(exists_file(Known)),
write('\t--- Invalid file: '),
write(Known), nl, halt.

scan_fi_options(['-i', Output|Options]):-
assert(fi_output_file(Output)),
scan_fi_options(Options).

scan_fi_options(_):-
print_usage.


% Predicates managing sub-options relatives to the test generation
% for features in isolation
%
scan_iso_options([]).

scan_iso_options(['-t', Output|Options]):-
assert(iso_output_file(Output)),
scan_iso_options(Options).

scan_iso_options(_):-
print_usage.


% Predicates managing sub-options relatives to the Tests Generation
% for incoherences
```

```
%
scan_ts_options([]).

scan_ts_options(['-t', Output|Options]):-
assert(ts_output_file(Output)),
scan_ts_options(Options).

scan_ts_options(_):-
print_usage.


% Predicates managing sub-options relatives to the Incoherences
% analysis and Tests Generation
%
scan_ft_options([]).

scan_ft_options(['-g', Groups|Options]):-
exists_file(Groups),
consult(Groups),
assert(groups_loaded(yes)),
scan_ft_options(Options)
 ;
not(exists_file(Groups)),
write('\t--- Invalid file: '),
write(Groups), nl, halt.

scan_ft_options(['-k', Known|Options]):-
exists_file(Known),
consult(Known),
scan_ft_options(Options)
 ;
not(exists_file(Known)),
write('\t--- Invalid file: '),
write(Known), nl, halt.

scan_ft_options(['-i', Output|Options]):-
assert(fi_output_file(Output)),
scan_ft_options(Options).

scan_ft_options(['-t', Output|Options]):-
assert(ts_output_file(Output)),
scan_ft_options(Options).

scan_ft_options(_):-
print_usage.
```

```
% Predicate printing informations about usage of the program
%
print_usage:-
write('Usage: fi-lookup <command>'), nl,
write('    * where command must be one of the following:'), nl,
nl,
write('    -filtering <features> [options]'), nl,
write('    --> find possible incoherences'), nl,
write('\t* where options include:'), nl,
write('\t-g <file> --- feature sets to compare'), nl,
write('\t-k <file> --- list of already known incoherences'), nl,
write('\t-i <file> --- file to which reports will be written'), nl,
nl,
write('    -testiso <features> <mapping-rules> [options]'), nl,
write('    --> produce test suites for features in isolation'), nl,
write('\t* where options include:'), nl,
write('\t-t <file> --- file to which test suites will be written'), nl,
       nl,
write('    -testfi <mapping-rules> <known incoherences> [options]'), nl,
write('    --> produce test suites for given incoherences'), nl,
write('\t* where options include:'), nl,
write('\t-t <file> --- file that will contain the test scenarios'), nl,
nl,
write('    -filttest <features> <mapping-rules> [options]'), nl,
write('    --> finds possible incoherences & '),
write('produces test suites'), nl,
write('\t* where options include:'), nl,
write('\t-g <file> --- feature sets to compare'), nl,
write('\t-k <file> --- list of already known incoherences'), nl,
write('\t-i <file> --- file to which reports will be written'), nl,
write('\t-t <file> --- file to which test suites will be written'),
nl, halt.

%----------* Arguments predicate definitions *----------------------------

                %**************************************%

%----------* Menu predicate definitions *---------------------------------

% Predicate main, first predicate to be called when program starts
```

```
%
main:-
dynamic(contradiction_pair(_, _)),
dynamic(lfeature(_)),
dynamic(rfeature(_)),
dynamic(fi(_, _, _, _, _, _, _, _, _)),
dynamic(test_done(_, _, _)),
dynamic(iso_test_done(_)),
dynamic(fi_output_file(_)),
dynamic(iso_output_file(_)),
dynamic(ts_output_file(_)),

current_output(Stream),
assert(stdout(Stream)),
assert(groups_loaded(no)),
unix(argv(Args)), (
  ( Args = [_|[]], print_usage )
          ;
  ( Args = [_|List], scan_args(List) )
       ).

%----------* Menu predicate definitions *---------------------------------


                    %************************************%


%----------* Rules for building the set of incoherences *-----------------

% Built-in contradiction, a user can not subscribe to the same
% feature more than one time.


% Main predicates, compare all features of the left hand set to
% all features of the right hand set.
%
fi_search:-
        (
  ( groups_loaded(no), fi_search_left_to_right )
          ;
         ( groups_loaded(yes),
           fi_search_left_to_right,
           fi_search_right_to_left
         )
        ).
```

```
% Analysis of features of the left hand set againts features of the
% right hand set
%
fi_search_left_to_right:-
rule(R),
lfeature(F1),
rfeature(F2),
not(fi(R, F1, F2, _, _, _, _, _, _)),
fi_check(R, F1, F2, PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2),
assert(fi(R, F1, F2, PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2)),
write('%---------------------------------------'),
write('---------------------------------------'), nl,
fi_infos(R, F1, F2, PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2),
nl,
portray_clause(fi(R, F1, F2, PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2)),
nl,
write('%---------------------------------------'),
write('---------------------------------------'),
nl, nl,
write('%                                  ********************'),
nl, nl,
fi_search_left_to_right, !.

fi_search_left_to_right:-
rule(R),
lfeature(F1), rfeature(F2),
(fi(R, F1, F2, _, _, _, _, _, _)), !.


% Analysis of features of the right hand set againts features of the
% left hand set
%
fi_search_right_to_left:-
rule(R),
rfeature(F1),
lfeature(F2),
not(fi(R, F1, F2, _, _, _, _, _, _)),
fi_check(R, F1, F2, PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2),
assert(fi(R, F1, F2, PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2)),
write('%---------------------------------------'),
write('---------------------------------------'), nl,
fi_infos(R, F1, F2, PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2),
nl,
```

```
portray_clause(fi(R, F1, F2, PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2)),
nl,
write('%----------------------------------------'),
write('----------------------------------------'),
nl, nl,
write('%                                  ********************'),
nl, nl,
fi_search_right_to_left, !.

fi_search_right_to_left:-
rule(R),
rfeature(F1), lfeature(F2),
(fi(R, F1, F2, _, _, _, _, _, _)), !.

%----------* Rules for building the set of incoherences *-------------------


                    %************************************%


%----------* Rules for incoherences identification *------------------------

% There is a contradiction between Set1 and Set2 if an element of Set1
% is opposed to another element of Set2 and vice versa.
% * The opposed elements must be defined in the feature descriptions.
%
contradiction(Set1, Set2):-
member(X, Set1),
member(Y, Set2),
(contradiction_pair(X, Y) ; contradiction_pair(Y, X)), !.


% *** RULE #d1 ***
%
% The goal of this rule is to identify  incoherences between
% two features held by the same user when no contradiction exists
% between pre-conditions, triggering events are the same and results
% are different.
%
% * Typical: Incoherence between Voice Mail and Call Waiting
%
fi_check(d1, [F1, Fx], [F2, Fy], PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2):-
not(fi(d1, [F2, Fy], [F1, Fx], _, _, _, _, _, _)),
feature([F1, Fx], PcnF1, TrgF1, ResF1), !,
feature([F2, Fy], PcnF2, TrgF2, ResF2),
```

```
TrgF1 = TrgF2,
ResF1 \= ResF2,
member(subs(U, F1), PcnF1), member(subs(U, F2), PcnF2),
not(contradiction(PcnF1, PcnF2)),
not(contradiction(ResF1, ResF2)), !.


% *** RULE #d2 ***
%
% The goal of this rule is to identify incoherences between
% two features held by the same user when no contradiction exists
% between pre-conditions, triggering events are the same and results
% present a contradiction.
%
% * Typical: Incoherence between Call Waiting and Incoming Call Screening
%
fi_check(d2, [F1, Fx], [F2, Fy], PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2):-
not(fi(d2, [F2, Fy], [F1, Fx], _, _, _, _, _, _)),
feature([F1, Fx], PcnF1, TrgF1, ResF1), !,
feature([F2, Fy], PcnF2, TrgF2, ResF2),

TrgF1 = TrgF2,
member(subs(U, F1), PcnF1), member(subs(U, F2), PcnF2),
not(contradiction(PcnF1, PcnF2)),
contradiction(ResF1, ResF2), !.


% *** RULE #d3 ***
%
% The goal of this rule is to identify incoherences between
% two features held by the same user when no contradiction exists
% between pre-conditions, triggering events are the same and results
% are different.
%
% * Typical: Incoherence between Voice Mail and Call Waiting
%
fi_check(d3, [F1, Fx], [F2, Fy], PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2):-
feature([F1, Fx], PcnF1, TrgF1, ResF1), !,
feature([F2, Fy], PcnF2, TrgF2, ResF2),

TrgF1 = TrgF2,
ResF1 \= ResF2,
member(subs(U1, F1), PcnF1), member(concerns(U2, F1), PcnF1),
member(subs(U2, F2), PcnF2), member(concerns(U2, F2), PcnF2),
```

```
U1 \= U2,
not(contradiction(PcnF1, PcnF2)),
not(contradiction(ResF1, ResF2)), !.


% *** RULE #d4 ***
%
% The goal of this rule is to identify incoherences between
% two features held by the same user when no contradiction exists
% between pre-conditions, triggering events are the same and results
% present a contradiction.
%
% * Typical: Incoherence between Voice Mail and Call Waiting
%
fi_check(d4, [F1, Fx], [F2, Fy], PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2):-
feature([F1, Fx], PcnF1, TrgF1, ResF1), !,
feature([F2, Fy], PcnF2, TrgF2, ResF2),

TrgF1 = TrgF2,
member(subs(U1, F1), PcnF1), member(concerns(U2, F1), PcnF1),
member(subs(U2, F2), PcnF2), member(concerns(U2, F2), PcnF2),
U1 \= U2,
not(contradiction(PcnF1, PcnF2)),
contradiction(ResF1, ResF2), !.


% *** RULE #t1 ***
%
% The goal of this rule is to identify incoherences between
% two features held by different users when no contradiction exists
% between pre-conditions, results of F1 include triggering events of F1
% and results present a contradiction.
%
% * Typical: B-(has(CFA), to(C)) & C-(has(ICS), ics_list(A)). A calls B.
% * Typical: B-(has(CFA), to(C)) & A-(has(OCS), ocs_list(C)). A calls B.
%
fi_check(t1, [F1, Fx], [F2, Fy], PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2):-
feature([F1, Fx], PcnF1, TrgF1, ResF1),
        feature([F2, Fy], PcnF2, TrgF2, ResF2),

not(contradiction(PcnF1, PcnF2)),
        subset(TrgF2, ResF1),
        contradiction(ResF1, ResF2), !.
```

```
% *** RULE #t2 ***
%
% The goal of this rule is to identify loops between features held by
% different users when no contradiction exists between pre-conditions,
% results of F1 includes triggering events of F2 and vice-versa.
%
% * Typical: B-(has(CFA), to(C)) & C-(has(CFA), to(B)). A calls B.
%
fi_check(t2, [F1, Fx], [F2, Fy], PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2):-
not(fi(d2, [F2, Fy], [F1, Fx], _, _, _, _, _, _)),
feature([F1, Fx], PcnF1, TrgF1, ResF1), !,
feature([F2, Fy], PcnF2, TrgF2, ResF2),

not(contradiction(PcnF1, PcnF2)),
subset(TrgF2, ResF1),
subset(TrgF1, ResF2), !.

%----------* Rules for incoherences identification *-----------------------


                    %************************************%


%------------* Rules for building the set of test scenarios *---------------

% Rule generating test scenarios for each of the incoherences found
% if no analysis was previously done, it will call it to first have
% the list of incoherences, then, it will build test scenarios.
%
fi_gen_test:-
fi(R, F1, F2, PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2),
not(test_done(R, F1, F2)), assert(test_done(R, F1, F2)),
fi_test(R, F1, F2, PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2),
fi_gen_test, !.


fi_gen_test:-
fi(R, F1, F2, _, _, _, _, _, _),
test_done(R, F1, F2), !.

% Rule generating test scenarios for the features in isolation
%
iso_gen_test:-
feature(F, P, T, R),
not(iso_test_done(F)), assert(iso_test_done(F)),
```

```
isolation_test(F, P, T, R),
iso_gen_test, !.

iso_gen_test:-
feature(F, _, _, _),
iso_test_done(F), !.

%------------* Rules for building the set of test scenarios *---------------


                    %************************************%


%------------* Rules for deriving test scenarios *-------------------------

% Test rule to derive test scenarios for features in isolation
%
isolation_test(F1, PcnF1, TrgF1, ResF1):-
          test_header([F1, PcnF1]),
          test_pre_conditions(PcnF1, []),
          test_triggers(TrgF1),
          test_results(ResF1),
          test_footer.


% Test rule #dx, derivation of test from Rules #d1, #d2, #d3 and #d4
%
fi_test(R, F1, F2, PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2):-
((R = d1); (R = d2);(R = d3); (R = d4)),
% test scenario case 1, feature F1 has priority
test_header([F1, PcnF1], [F2, PcnF2], F1, R),
test_pre_conditions(PcnF1, PcnF2),
test_triggers(TrgF1),
test_results(ResF1),
test_footer,

% test scenario case 2, feature F2 has priority
test_header([F1, PcnF1], [F2, PcnF2], F2, R),
test_pre_conditions(PcnF1, PcnF2),
test_triggers(TrgF2),
test_results(ResF2),
test_footer.


% Test rule #t1, derivation of test from Rule #t1
```

```
%
fi_test(t1, F1, F2, PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2):-
% test scenario case 1, feature F1 has priority
test_header([F1, PcnF1], [F2, PcnF2], F1, t1),
test_pre_conditions(PcnF1, PcnF2),
test_triggers(TrgF1),
test_results(ResF1),
test_footer,

% test scenario case 2, feature F2 has priority
test_header([F1, PcnF1], [F2, PcnF2], F2, t1),
test_pre_conditions(PcnF1, PcnF2),
test_triggers(TrgF1),
test_results(TrgF2),
test_results(ResF2),
test_footer.


% Test rule #t2, derivation of test from Rule #t2
%
fi_test(t2, F1, F2, PcnF1, TrgF1, ResF1, PcnF2, _, ResF2):-
% test scenario, features enter a loop
test_header([F1, PcnF1], [F2, PcnF2], ['a', 'loop'], t2),
test_pre_conditions(PcnF1, PcnF2),
test_triggers(TrgF1),
test_results(ResF1),
test_results(ResF2),
test_results(ResF1),
test_footer.

%------------* Rules for deriving test scenarios *-------------------------


                    %**************************************%


%------------* Rules for giving informations about an incoherence ----------

% Information about the incoherence found
%
fi_infos(d1, [F1, Fx], [F2, Fy], PcnF1, TrgF1, ResF1, PcnF2, _, ResF2):-
nl, write('% * Rule #d1'), write(' -> '),
write([F1, Fx]), write(' & '), write([F2, Fy]), nl,
write('%   Same user subscribed to two features having the'), nl,
write('%   same triggering events and different results'), nl, nl,
```

```
write('%      + Features pre-conditions'), nl, nl,
write('%         - Pre-conditions of '), write([F1, Fx]), nl,
writeSet('%              ', PcnF1), nl,
write('%         - Pre-conditions of '), write([F2, Fy]), nl,
writeSet('%              ', PcnF2), nl,

write('%      + Same triggering events'), nl,
writeSet('%           ', TrgF1), nl,

write('%      + Different results'), nl, nl,

write('%         - Resulting events of '), write([F1, Fx]), nl,
writeSet('%              ', ResF1), nl,
write('%         - Resulting events of '), write([F2, Fy]), nl,
writeSet('%              ', ResF2), nl.


% Information about the incoherence found
%
fi_infos(d2, [F1, Fx], [F2, Fy], PcnF1, TrgF1, ResF1, PcnF2, _, ResF2):-
nl, write('% * Rule #d2'), write(' -> '),
write([F1, Fx]), write(' & '), write([F2, Fy]), nl,
write('%   Same user subscribed to two features having the'), nl,
write('%   same triggering events and contradictory results'), nl, nl,

write('%      + Features pre-conditions'), nl, nl,
write('%         - Pre-conditions of '), write([F1, Fx]), nl,
writeSet('%              ', PcnF1), nl,
write('%         - Pre-conditions of '), write([F2, Fy]), nl,
writeSet('%              ', PcnF2), nl,

write('%      + Same triggering events'), nl,
writeSet('%           ', TrgF1), nl,

write('%      + Contradictory results'), nl, nl,

write('%         - Resulting events of '), write([F1, Fx]), nl,
writeSet('%              ', ResF1), nl,
write('%         - Resulting events of '), write([F2, Fy]), nl,
writeSet('%              ', ResF2), nl.


% Information about the incoherence found
%
```

```
fi_infos(d3, [F1, Fx], [F2, Fy], PcnF1, TrgF1, ResF1, PcnF2, _, ResF2):-
nl, write('% * Rule #d3'), write(' -> '),
        write([F1, Fx]), write(' & '), write([F2, Fy]), nl,
write('%   Different users subscribed to two features where'), nl,
        write('%   feature 1 concerns subscriber of feature 2, and'), nl,
        write('%   the features have the same triggering events and'), nl,
        write('%   different results'), nl, nl,

write('%      + Features pre-conditions'), nl, nl,
write('%         - Pre-conditions of '), write([F1, Fx]), nl,
writeSet('%              ', PcnF1), nl,
write('%         - Pre-conditions of '), write([F2, Fy]), nl,
writeSet('%              ', PcnF2), nl,

write('%      + Same triggering events'), nl,
writeSet('%           ', TrgF1), nl,

write('%      + Different results'), nl, nl,

write('%         - Resulting events of '), write([F1, Fx]), nl,
writeSet('%              ', ResF1), nl,
write('%         - Resulting events of '), write([F2, Fy]), nl,
writeSet('%              ', ResF2), nl.


% Information about the incoherence found
%
fi_infos(d4, [F1, Fx], [F2, Fy], PcnF1, TrgF1, ResF1, PcnF2, _, ResF2):-
nl, write('% * Rule #d4'), write(' -> '),
        write([F1, Fx]), write(' & '), write([F2, Fy]), nl,
write('%   Different users subscribed to two features where'), nl,
        write('%   feature 1 concerns subscriber of feature 2, and'), nl,
        write('%   the features have the same triggering events and'), nl,
        write('%   contradictory results'), nl, nl,

write('%      + Features pre-conditions'), nl, nl,
write('%         - Pre-conditions of '), write([F1, Fx]), nl,
writeSet('%              ', PcnF1), nl,
write('%         - Pre-conditions of '), write([F2, Fy]), nl,
writeSet('%              ', PcnF2), nl,

write('%      + Same triggering events'), nl,
writeSet('%           ', TrgF1), nl,

write('%      + Contradictory results'), nl, nl,
```

```
write('%         - Resulting events of '), write([F1, Fx]), nl,
writeSet('%              ', ResF1), nl,
write('%         - Resulting events of '), write([F2, Fy]), nl,
writeSet('%              ', ResF2), nl.


% Information about the incoherence found
%
fi_infos(t1, [F1, Fx], [F2, Fy], PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2):-
nl, write('% * Rule #t1'), write(' -> '),
write([F1, Fx]), write(' & '), write([F2, Fy]), nl,
write('%   The user(s) subscribed to different features and the'), nl,
write('%   results of F1 include the triggering events of F2'), nl,
write('%   but their results are contradictory'), nl, nl,

write('%      + Different pre-conditions'), nl, nl,
write('%         - Pre-conditions of '), write([F1, Fx]), nl,
writeSet('%              ', PcnF1), nl,
write('%         - Pre-conditions of '), write([F2, Fy]), nl,
writeSet('%              ', PcnF2), nl,

write('%      + Transitivity between features'), nl, nl,

write('%         - Triggering events of '), write([F1, Fx]), nl,
writeSet('%              ', TrgF1), nl,
write('%         - Resulting events of '), write([F1, Fx]), nl,
writeSet('%              ', ResF1), nl,

write('%         - Including triggering events of '), write([F2, Fy]),
nl, writeSet('%              ', TrgF2), nl,

write('%      + Contradictory results'), nl, nl,

write('%         - Resulting events of '), write([F2, Fy]), nl,
writeSet('%              ', ResF2), nl,
write('%         - Resulting events of '), write([F1, Fx]), nl,
writeSet('%              ', ResF1), nl.


% Information about the incoherence found
%
fi_infos(t2, [F1, Fx], [F2, Fy], PcnF1, TrgF1, ResF1, PcnF2, TrgF2, ResF2):-
nl, write('% * Rule #t2'), write(' -> '),
write([F1, Fx]), write(' & '), write([F2, Fy]), nl,
```

```
write('%   The users subscribed to different features for which'), nl,
write('%   the resulting events of F1 include triggering events'), nl,
write('%   of F2 and vice-versa, which leads to a loop.'), nl, nl,

write('%      + Different or common pre-conditions'), nl, nl,
write('%         - Pre-conditions of '), write([F1, Fx]), nl,
writeSet('%              ', PcnF1), nl,
write('%         - Pre-conditions of '), write([F2, Fy]), nl,
writeSet('%              ', PcnF2), nl,

write('%      + Transitivity between features'), nl, nl,

write('%         - Resulting events of '), write([F1, Fx]), nl,
writeSet('%              ', ResF1), nl,
write('%         - Including triggering events of '), write([F2, Fy]),
nl, writeSet('%              ', TrgF2), nl,

write('%      + Reverse transitivity, leading to a loop'), nl, nl,
write('%         - Resulting events of '), write([F2, Fy]), nl,
writeSet('%              ', ResF2), nl,
write('%         - Including triggering events of '), write([F1, Fx]),
nl, writeSet('%              ', TrgF1), nl.

%------------* Rules for giving informations about an incoherence ----------
```

MQ

66046

UMI

MICROFILMED 2002

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600



Université d'Ottawa • University of Ottawa

PERMISSION DE REPRODUIRE ET DE DISTRIBUER LA THÈSE

PERMISSION TO REPRODUCE AND DISTRIBUTE THE THESIS

NOM DE L'AUTEUR / NAME OF AUTHOR:	GUENNOUN, Mouhcine
ADRESSE POSTALE / MAILING ADDRESS:	1219-100 Hinchey Ottawa, Ontario K2Y 4L9
GRADE / DEGREE:	**ANNÉE D'OBTENTION / YEAR GRANTED**
M.A.Sc.(Electrical Engineering)	2001
TITRE DE LA THÈSE / TITLE OF THESIS:	
DESIGN AND IMPLEMENTATION OF A POLICY MANAGEMENT IN AN AGENT-BASED SYSTEM	

L'auteur permet, par la présente, la consultation et le prêt de cette thèse en conformité avec les règlements établis par le bibliothécaire en chef de l'Université d'Ottawa. L'auteur autorise aussi l'Université d'Ottawa, ses successeurs et cessionnaires, à reproduire cet exemplaire par photograhie ou photocopie pour fins de prêt ou de vente au prix coûtant aux bibliothèques ou aux chercheurs qui en feront la demande.

Les droits de publication par tout autre moyen et pour vente au public demeureront la propriété de l'auteur de la thèse sous réserve des règlements de l'Université d'Ottawa en matière de publication de thèses.

The author hereby permits the consultation and the lending of this thesis pursuant to the regulations established by the Chief Librarian of the University of Ottawa. The author also authorizes the University of Ottawa, its successors and assignees, to make reproductions of this copy by photographic means or by photocopying and to lend or sell such reproductions at cost to libraries and to scholars requesting them.

The right to publish the thesis by other means and to sell it to the public is reserved to the author, subject to the regulations of the University of Ottawa governing the publication of theses.

N.B. LE MASCULIN COMPREND EGALEMENT LE FEMININ

May 14th 2001

DATE

(AUTEUR) SIGNATURE (AUTHOR)



Université d'Ottawa · University of Ottawa



Université d'Ottawa · University of Ottawa

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

GUENNOUN, Mouhcine

AUTEUR DE LA THÈSE - AUTHOR OF THESIS

M.A.Sc. (Electrical Engineering)

GRADE - DEGREE

School of Information Technology and Engineering

FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

Design and Implementation of a Policy Management in an Agent-Based System

Ahmed Karmouch

DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

L. Orozco-Barbosa

B. Pagurek

J.-M. De Koninck, Ph.D.

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

SIGNATURE

DEAN OF THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

Design and Implementation of a Policy Management in an Agent-Based System

By
Mouhcine Guennoun

A thesis submitted to the
Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

M.A.Sc.
in

Electrical Engineering

Ottawa Carleton Institute for Electrical & Computer Engineering
School of Information and Technology Engineering (SITE)
University of Ottawa

December, 2000

©Mouhcine Guennoun

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-66046-X

Canada

Table of Contents

ABSTRACT V

ACKNOWLEDGMENT VI

ACRONYMS VII

CHAPTER 1: INTRODUCTION 1

1.1 Overview 1

1.2 Objectives 2

1.3 Main Contributions 2

1.4 Thesis Outline 3

CHAPTER 2: AGENTS & POLICIES 4

2.1 Agent Technology 4
2.1.1. Definition 4
2.1.2 Adaptive User Agents 5
2.1.3. Agent Communication Language 6
2.1.4 Multiple Agent Systems 7
2.1.4.1 Direct Communication 8
2.1.4.2 Assisted Coordination 10
2.1.5. Mobile Agents 11

2.2 Policy Overview 12
2.2.1 Deontic logic 14
2.2.2 Obligation policies 14
2.2.3 Organizational Policies 15
2.2.4 Policy Language 16
2.3 IETF Policy Framework 16
2.3.1 Introduction 16
2.3.2 Policy Schema 17
2.3.3 Policy Definition 18
2.3.4 Policy Architecture 19
2.3.5 Policy Storage & Retrieval 20

2.4 Chapter Summary 21

CHAPTER 3: POLICY MANAGEMENT: DESIGN & CONCEPTS 22

3.1. Introduction 22

3.2. Personal Mobility Management System 22
3.2.1. Introduction 22
3.2.2 Architecture 24
3.2.3. Agent Communication 26
3.2.4 Policies in PMMS 27

3.2. Policy Management 28
3.2.1 System Architecture 28
3.2.2 Policy Rules 29
3.2.2.1 Authorization 30
3.2.2.1 Obligations 32
4.2.3 Policy Editor 33
3.2.4 Policy Validation & Conflict Resolution 34
3.2.5 System Agents 35
3.2.6 Event Server 37
3.2.7 Facilitator Agent 38
3.2.8 Policy Server 39
3.2.8.1 Notification of Agent Roles 39
3.2.8.2 Policy Distribution 40
3.2.8.3 System Adaptation 41
3.2.9. Authorization Server 43

3.3 Chapter Summary 45

CHAPTER 4: POLICY MANAGEMENT IN PMMS 46

4.1 Classification of Policies 46
4.1.1 Default Policy 46
4.1.2 Security Policies 47
4.1.3 Admission Control Policy 48
4.1.4 User Profile Policies 50
4.1.5 Resource Reservation Policies 52
4.1.6 System Policies 54

4.2 Processing Policy Requests 55

4.3 Negotiation Model 60

4.4 Chapter Summary 66

CHAPTER 5: IMPLEMENTATION **67**

5.1 Technology Integration **67**
5.1.1 Java 68
5.1.2 MicMac 69
5.1.3 KQML 69
5.1.4 XML 70

5.2 XML Policy Representation **70**
5.2.1 Policy Document Type Definition of a policy 71
5.2.2 Example of an XML policy representation 72

5.3 Graphical User Interface **73**
5.3.1 Policy Editor 73
5.3.2 Constraint Editor 74
5.3.3 Policy Validation & Transformation 74

5.4 Agent Development Model **75**
5.4.1 Policy Layer 75
5.4.2 Communication Layer 77
5.4.3 Actions Layer 77
5.4.4 Knowledge Base 78

5.5. Policy Agents **78**
5.5.1 Policy Server 78
5.5.2 Authorization Server 80

CHAPTER 6: CONCLUSION **81**

6.1 Summary **81**

6.2 Future Work **82**

REFERENCES **83**

Table of Figures

FIGURE 2.1: THE CONTRACT NET PROTOCOL 9
FIGURE 2.2: FEDERATE SYSTEM ARCHITECTURE 10
FIGURE 2.3: CLASS HIERARCHY IN CORE POLICY SCHEMA 17
FIGURE 2.4: COMPOSITION OF POLICIES 18
FIGURE 2.5: IETF POLICY FRAMEWORK ARCHITECTURE 19
FIGURE 3.1: PMMS ARCHITECTURE 25
FIGURE 3.2: SITE POLICY ARCHITECTURE 28
FIGURE 3.3: EVENT NOTIFICATION MECHANISM 37
FIGURE 3.4: SYSTEM ADAPTATION 42
FIGURE 3.5: STEPS OF REQUESTING AN AUTHORIZATION 43
FIGURE 4.1: CONSTRAINT OF POLICY A010 57
FIGURE 4.2: DECORATED TREE OF POLICY A010 58
FIGURE 4.3: DECORATED TREE POLICY A011 58
FIGURE 4.4: DECORATED TREE OF POLICY A021 59
FIGURE 4.5: BELIEF OF VIDEO AGENT 60
FIGURE 4.6: BELIEF OF AUTHORIZATION SERVER 60
FIGURE 4.7: FIRST PROPOSAL FROM VIDEO AGENT 61
FIGURE 4.8: PROCESSING THE PROPOSAL 61
FIGURE 4.9: POLICIES OF VIDEO MAIL AGENT 62
FIGURE 4.10: POLICIES OF AUTHORIZATION SERVER 62
FIGURE 4.11: FIRST PROPOSAL 63
FIGURE 4.12: SECOND PROPOSAL 64
FIGURE 4.13: AGREEMENT IS REACHED 65
FIGURE 5.1: IMPLEMENTATION TOOLS 68
FIGURE 5.2: POLICY DOCUMENT TYPE DEFINITION 71
FIGURE 5.3: AGENT DEVELOPMENT MODEL 75
FIGURE 5.4: MAPPING USER PROFILE TO LOCAL POLICIES 79

Abstract

Personal Mobility Management System, a distributed application developed at the Multimedia & Mobile Agent Research Laboratory, allows users to move across a virtual network of predefined sites. Each site provides personalized services to its visitors.

The purpose of this thesis is to introduce an agent-based architecture to create and manage policies on each site involved in the virtual network. The architecture defines an administrative infrastructure that allows each site to restrict access to the local resources and to sensitive information by visiting users. These restrictions are expressed in terms of policies that are stored in a policy server at each site. Policies are also related to the privacy and ownership levels attached to resources as well as to profiles of visiting users. The thesis presents a model to detect conflicts that may occur between different agents as a result of rules imposed by different policies. These conflicts are resolved through negotiation between agents.

Acknowledgment

I would like to extend special appreciation to my thesis supervisor Professor Ahmed Karmouch, for the time and effort he had invested, and his continued guidance and support throughout this thesis work.

I also want to specially thank Tom Gray and Serge Mankovski from Mitel Corporation, Hamid Harroud, Research Assistant in MMARL, for theirs comments and criticisms that were very helpful for the achievement of this work.

Thanks are also due to the Multimedia & Mobile Agents Research Laboratory members for their help and provision of information.

A major thanks goes to my parents for their love and dedication. Also, I would like to thank my family and my friends, for the assistance and understanding during all these years of education, without which this work wouldn't have been possible.

Acronyms

ACL: Agent Communication Language.

AEE: Agent Execution Environment.

AI: Artificial Intelligence.

API: Application Programming Interface.

ARPA: Advanced Research Projects Agency.

CNP: Contract Net Protocol.

COPS: Common Open Policy Service.

IETF: Internet Engineering Task Force.

JESS: Java Expert System Shell.

JVM: Java Virtual Machine.

KIF: Knowledge Interchange Format.

KQML: Knowledge Query Manipulation Language.

KSE: Knowledge Sharing Effort.

LDAP: Lightweight Directory Access Protocol.

MAS: Multi Agent System.

PDP: Policy Decision Point.

PEP: Policy Enforcement Point.

PMMS: Personal Mobility Management System.

SNMP: Simple Network Management Protocol.

UMTS: Universal Mobile Telecommunication System.

VHE: Virtual Home Environment.

XML: eXtensible Markup Language.

Chapter 1

Introduction

1.1 Overview

Researchers have utilized the principles of multi-agent architectures in designing distributed and federated multi-enterprise systems. One example of such a system is the Personal Mobility Management System designed and implemented by the Mobile Agent Alliance, a collaborative consortium comprising the University of Ottawa, Mitel Corporation and the National Research Council of Canada. The Personal Mobility Management System (PMMS) provides personalized services and access to resources for nomadic users over a network [33]. It provides a nomadic user with a working environment similar to her home environment. In the home site, a user has a profile that describes her preferences: the services she uses, the quality of service required, the cost she is willing to pay for certain billable services, etc. A dynamic mapping between the user profile and the local policies on each site determines a set of services that a visiting site can offer to its guest. A nomadic user can be requested to pay a cost for services such as long distance calling. For example, in the instance in which the visiting site is a hotel, which has made arrangements with a guest's employers having regard to access to hotel facilities, costs that the guest incurs in the use of services can be billed directly to the employer. Thus, it is desirable for the employer to be able to institute policies on which billable services are allowable. Hence, agent negotiation techniques are used between the home site and the visiting site in order to agree on the cost or to find a service that can fit the requirements of the user with a lower cost. Negotiation is also used for resource reservation when a user starts to use a service on a visiting site.

1.2 Objectives

The use of policies in a multi-enterprise agent based systems is required to maintain privacy, accountability of proper usage of communication resources, maximize the availability of communication access for users, and provide security definitions for accessing an organization services. Very little research has been undertaken to date in the area of integrating policies and agents together [40]. With the use of agents, we believe that real automated and policy driven management can be accomplished. Policies can be used to monitor agents' behavior in specific situations and to ensure that such behavior is consistent with the global strategy of the system.

The present work aims to develop a policy platform for the control of multi-enterprise agent-based systems. Policies are defined to regulate the actions of software agents in the Inter-Sites Personal Mobility Management System [33], which is a system designed to provide personalized services and access to resources for nomadic users over a network.

1.3 Main Contributions

The main contribution of this thesis is the design and implementation of a multi-agent system for managing system policies on a site within a virtual network. System policies include authorization policies, which are actions an agent is authorized to carry out, and obligation policies for specifying actions that an agent is responsible for performing, comprising:

- A policy server for receiving and downloading obligation policies into agents representing system services and devices, for distributing authorization policies to

authorization server, and for managing system policies in accordance with changes in system modes.

- An authorization server operating in accordance with authorization policies for receiving and authenticating requests from agents. The server processes the request to decide whether the requester agent is permitted to execute a specific action.
- An event server for enabling shared communication between multiple agents.
- A mechanism to detect conflicts that may occur between agents as result of conflicting policies.
- A negotiation model to resolve conflicts.

The validity of the architectural design is demonstrated in the Personal Mobility Management System [33], which aims to provide a nomadic user with a working environment similar to her home environment.

1.4 Thesis Outline

The rest of this thesis is structured as follows. Chapter 2 introduces the agent technology, gives an overview of different concepts of policies in the literature, and presents the policy framework developed by the Internet Engineering Task Force. Chapter 3 introduces the Personal Mobility Management System, describes the architecture of the policy service and the elements involved to setup and enforce policies within each site of a virtual network. Chapter 4 illustrates the application of the policy architecture to efficiently monitor a multi-agent system such as the PMMS system. Chapter 5 describes the implementation of different modules introduced in the design section in Chapter 3. Finally, Chapter 6 presents the summary and conclusions of this work, and provides suggestions for future work.

Chapter 2

Agents & Policies

2.1 Agent Technology

2.1.1. Definition

The term 'agent' has its background in the early work on Artificial Intelligence (AI) when researchers focused on trying to create artificial entities which can mimic human abilities [4], hence, earlier definitions of agents have been more oriented toward AI. In [5], Russel defined agents as "anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors." Based on this definition, each program is an agent. Thus, to make clear distinction between programs and agents, some notions of environment, sensing, and acting should be restricted. In [6], Maes defines autonomous agents as "computational systems that inhabit some complex dynamic environment, sense and act autonomously in this environment, and by doing so realize a set of goals or tasks for which they are designed." Therefore, it restricts the agent environment to be complex and dynamic. In Maes' definition, the agents are goal oriented and have the autonomy to decide how to realize the tasks they are designed to complete. Other definitions of agents have completed the previous ones by adding new properties. Consequently, agents must perform their actions with some level of proa-ctivity and reactivity [7]. In particular, agent does not simply act in response to the environment.

Agents should be able to interact with each other and manage conflicts that may occur between them as a result of these interactions. This property has been described in literature as social ability [8]. From all these definitions, we believe that an agent is an autonomous computational entity that acts on behalf of another entity (person/agent) and exhibits some level of the key attributes of proactivity, reactivity, learning, and co-operation. Agents are software systems that reside in computers and networks, assist users with computer-based tasks. Computers are now more ubiquitous and complex than before, therefore, the gap between sophisticated computers and untrained users is becoming more apparent. Agents can reduce this gap by assisting users to complete and monitor the new generation of software. They are designed to learn the user's interests and to act autonomously on their behalf. Hence, people will be engaged in a cooperative process in which both human and computer initiate communications, monitor events and perform tasks to meet the user's goal.

2.1.3 Adaptive User Agents

One important category of agents that is relevant to this work is the adaptive user agents. Such agents try to modify their behavior to maximize the productivity of the current users interactions with the system. They assist the user to communicate her tasks to the system by presenting a simple to use interface, which hides the actual underlying system, which may be very complex. Another interesting aspect of this category of agents is their ability to learn the user interests and represent them in a custom user profile [3]. Based on their experience with the user, the Interface Agents try to build up a profile of the user. One issue in building the user profile is gathering accurate information regarding the user's interest. Many approaches have been proposed to address this

problem. In [9], Mazur proposes the learning by observation method also known as imitative learning. Agents can learn user's interest simply by watching and copying users as they display their usual behavior. With today's growth of personal communication services, the user agents has evolved to provide integrated systems in which the user can access her personalized services independently of his location or his end device (PDA, laptop...). The Universal Mobility Telecommunication System (UMTS) [39] intends to offer users the ability to access their particular set of services from any device, anywhere at any time. In [10], an agent-based framework aims to provide and deliver "personalized services across network and terminal boundaries with the same look and feel". The framework is based on the concept of Virtual Home Environment (VHE) in which the user can be constantly presented with the same personalized services, user interface customization in whatever location or end device she is using [39].

2.1.4. Agent Communication Language

The interoperation between agents that use different languages has created a need for a universal communication language. A universal language will have the advantage of eliminating incompatibilities and national variations.

The ARPA Knowledge Sharing Effort (KSE) [42] has defined an Agent Communication Language (ACL) that satisfies the following requirements:

- Sufficiently expressive to communicate information of widely varying sorts.
- Sufficiently compact to avoid excessive growth over specialized languages.

The Knowledge Query & Manipulation Language (KQML) [11], developed by KSE, can be viewed as consisting of three layers: content, message and communication layers. The content layer specifies the actual content of the message; the message layer

specifies the set of performatives for agents' communication (e.g., tell, ask-if, subscribe, etc); and the third layer consists of the communication protocol.

A message is a KQML expression in which the arguments are expressed in Knowledge Interchange Format (KIF) [12] formed from the ACL vocabulary. The vocabulary of the language is comprised of a large set of words appropriate to common application areas. Each area is expressed with a specific ontology. For instance, geometry ontology contains vocabulary for describing three-dimensional geometry in terms of polar coordinates, rectangular coordinates, and cylindrical coordinates.

However, the ACL language has an important shortcoming, which is the lack of semantics as stated in [13]. ARCOL, a language developed by France Telecom [14], overcomes this limitation by adding a semantic layer in the ACL.

2.1.5 Multiple Agent Systems

Recent works on the field of Artificial Intelligence have investigated the development of software to simulate the capabilities of human beings, such as reasoning, natural language communication, and learning; that computational agents need to be part of the systems composed of computers and humans [15]. Such systems are called Multi-Agent Systems (MAS), and can be defined as a set of independent software agents that work together to resolve problems that can not be resolved individually [15].

As stated in [43][44], Multi-Agent Systems offer the following advantages:

- Give an alternative to centralized systems, which have many weaknesses, such as performance, reliability, security, etc.
- Permit the interoperation of different legacy systems.

- Increase the reliability of the system, which can be defined as the ability to recover from the failure of individual components, with minimal performance degradation.
- Decrease the response time of the system.
- Provide efficient organization of agents to enhance collaboration.

Coordination is a key element in the performance of MAS. It has for purpose to prevent chaos in the system; to coordinate agents with different expertise to meet the global goals of the system; and to avoid dead-locks, since in a cooperative approach, an agent may have to wait for another agent to complete its tasks before it executes its own tasks [16].

There are two approaches for addressing coordination in a MAS: direct communication and assisted coordination [16][37].

2.1.5.1 Direct Communication

In direct communication, agents handle their own coordination, without relying on a third party to facilitate their communication. One example of a protocol based on the direct communication approach is the Contract Net Protocol (CNP) [16].

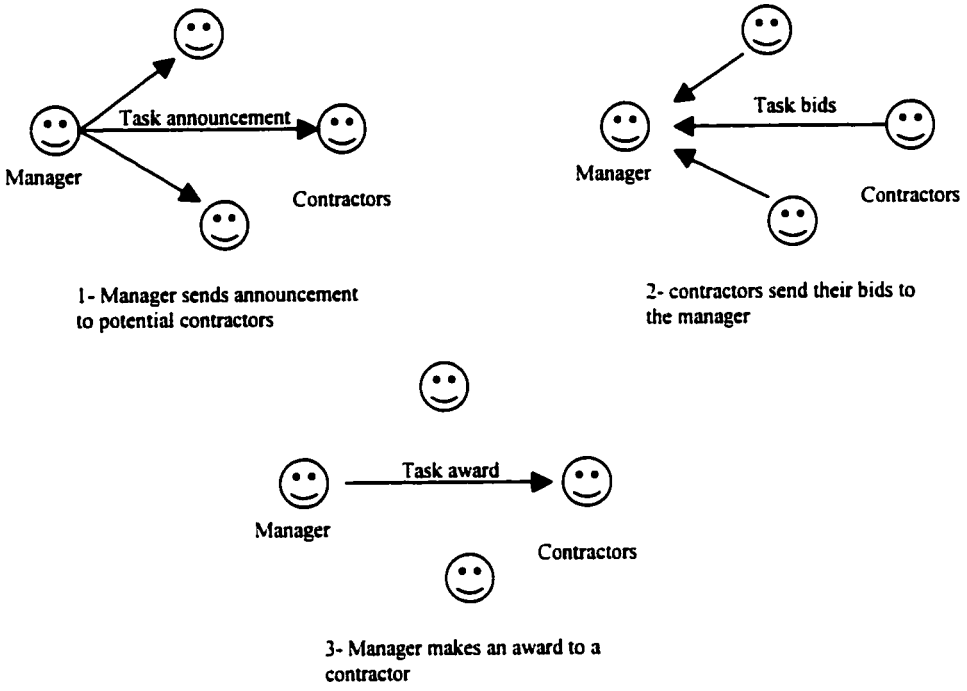


Figure 2.1: The Contract Net Protocol [16]

The protocol is in essence very simple, as it identifies two roles of agents: manager and contractor. A manager that cannot resolve a problem using local expertise has to distribute requests for proposal to other willing agents with the necessary expertise to solve the problem. Upon receipt of this announcement, the contractors submit their bids to the manager. The manager uses these bids to decide which agent has submitted the most appropriate bid, and therefore, awards it a contract. Figure 2.1 illustrates the CNP operations.

However, the CNP protocol has many disadvantages. The most obvious one is the excessive amount of communication generated. In a real world application, like an Internet-based application, it's prohibitive to broadcast and process bids of thousand of programs. Another disadvantage is that CNP assumes that there are no conflicts between

agents, and hence, doesn't detect and resolve them. The manager doesn't include its constraints in the announcement and has no possibility to bargain the bids.

2.1.5.2 Assisted Coordination

Assisted Coordination offers an alternative approach that eliminates to some extent the disadvantages of Direct Communication. In this approach, agents are organized in a federated system as illustrated in figure 2.2, in which. the communication is handled by system agents called facilitators [37].

Agents send the list of services they can offer and their needs to the Facilitator. Based on this information, facilitators build a knowledge base of system capabilities. which is used for routing future announcements. Assisted coordination has the advantage of being able to detect potential conflicts between agents plans. However, it requires that the system shares and process substantial amounts of information, and therefore, involves more computation and communication than the Direct Communication approach.

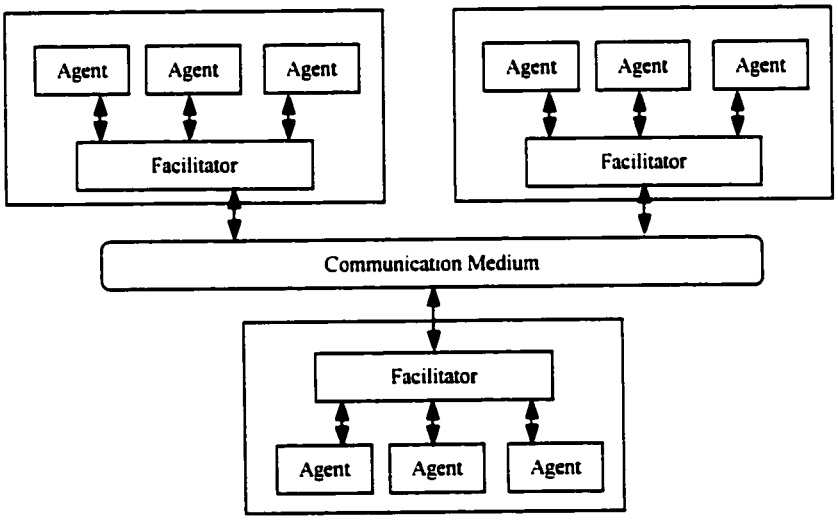


Figure 2.2: Federate System Architecture [37]

In order to resolve this issue, Assisted Coordination implementations use powerful search control techniques, developed in the Artificial Intelligence and Database communities, to enhance the processing of messages.

2.1.6. Mobile Agents

Mobile agents offer unique opportunities for structuring and implementing open distributed systems. A wide range of applications has been indicated for mobile agent technology, including electronic commerce, telecommunications services and network management [17].

Mobile agents technology was introduced to overcome the weaknesses of the traditional client/server model which are: the network overhead caused by the communication between client and the server, the latency and the unavailability of the network and the rigidity of the above model, where clients can only request a predefined set of operations from a remote server [2]. The mobile agents technology tackles these problems by enabling a program, called mobile agent, to travel to the remote server to request its services. Furthermore, the server can be customized to the need of the applications, that is, each new operation that is not supported by the server can be sent as a mobile agent . Thus, server components are no more installed by the administrators but are dynamically installed by the application itself.

The General Magic's Telescript [17] technology is the first commercial implementation of the mobile agents concept. The platform provides the basis for a complete remote programming. The concepts of this technology are [41]:

- Agent: A mobile entity that can travel over the network to complete its goal.

- Travel: Lets an agent travel from one place to another to obtain a service offered remotely and then return to its starting place.
- Meeting: Allows two agents to exchange messages on the same computer.
- Connection: Enables two remote agents to open a private communication channel to share information.
- Authority: Represents the certificate of the agent. It is a way to verify the entity of an agent. It does identify an agent as a passport identifies a human.
- Permits: The set of capabilities of an agent which indicate what operations a visitor agent can request, and what resources it can use.

The technology provides a complete, object oriented, dynamic and persistent programming language to facilitate the development of mobile agents. Furthermore, an engine, called Agent Execution Environment (AEE) [41], maintains and executes places running in its environment, as well as the agents occupying those places. A communication protocol suite enables two engines to communicate in order to transport agents between them.

2.2 Policy Overview

The term "policies" may be defined as a set of rules in a domain-managed object [23]. These rules are derived from management goals and related business strategies to define the desired behavior of distributed heterogeneous systems, networks and applications. They provide means for expressing how tasks are to be undertaken and how a system is to respond to a given situation. Policies may also be defined as a binary relationship between objects [18]. Binary relations, consisting of a subject and a target, can be used to express different types of policies. Management policies may be expressed

by a human manager in a high level manner, i.e. policy language. These policies must then be transformed and refined into scripts or rules that can be interpreted by software agents [18]. One advantage of policies is that they can be expressed independently from the agents that interpret them. This separation permits modification of the policies without changing the code of the agents [23]. Hence, agents can be reused in different networks where different policies apply.

Because of the large number of policies, classification systems have been developed to give semantics to the policies and to assist in automating the process of transforming and refining of the policies. Wies [38] has proposed a template to represent policies. The attributes of this template are:

Mode: Policies can be an obligation, permission, or prohibition.

Lifetime: A short, medium, or long term application can characterize the duration of a policy.

Trigger mode: The enforcement of a policy can be constantly active, repeated periodically for a specific period time, triggered by asynchronous events or a combination of the last two.

Activity: A policy can monitor or enforce actions on its target objects or react to a new situation that occurs in the system.

Functionality of targets: This class includes policies applicable to resources with common characteristics (e.g. printers, hubs, and routers...).

Below, we give other views and definitions of policies as described in literature.

2.2.1 Deontic logic

Many approaches to represent policies using the deontic logic, logic of permissions and obligations, have been described in literature. An early one was made by Hughes and Creswell in 1968 as described in [19].

In [20], Jonsher models the access behavior of database users by using deontic logic. The work divides the user rights into two categories: access rights, and normative rights. Access rights define the traditional users access to databases, like reading, deleting, and modifying. A positive access right is permission, and negative one is a prohibition.

The normative rights are defined as the duties of a user, i.e. the actions the user is obliged to do, whereas, the liberties are the actions the user is free to do.

Duties are seen as a way of controlling the fulfillment of some critical tasks in an organization. However, a duty may not have an authorization associated with it, and hence cannot be accomplished. Therefore, a relationship is needed to be defined between the authorizations and the obligations. A simple association is to consider an obligation to execute an action is an authorization by itself, based on the axiom "obligation implies permission"[20]. A second approach is to consider them as independent. In this case, a specific permission should be defined in order that the obligation can be completed.

2.2.2 Obligation policies

Becker [21] defines an obligation policy as the management actions a manager must perform in order to keep the network operational. A policy defines a relationship between a set of managers and managed objects specified as domains instead of particular elements. A domain is a logical grouping of objects and it can be organized in a

hierarchy of domains making the system management more scalable for large systems. All objects belonging to a specific domain will have to carry the policies that apply to this particular domain. The advantage of this approach is that the creation/deletion of an object doesn't require the alteration of predefined policies.

Koch [22] defines the management tasks as a set of production rules that are represented in an expert system model. Authorization policies are represented as preconditions to rules. When the rules' conditions are satisfied, the rules' actions are executed. Like rules of an expert system, new conditions may be satisfied when these actions are performed. Triggers of obligation policies are defined using an event definition language.

2.2.3 Organizational Policies

Many researchers have recognized the role played by the goals and objectives of an organization in defining the behavior of its information systems. Wies [23] views management policies as derived from the goals of management and related business strategies. In [24], Mcbrien proposed the External Rule Language (ERL) as a way to model organizational policy. In [25], system management is defined as the set of processes that are applied to the administration and control of information systems within an organization according to a given set of policies. Policies are to refine processes that are designed to achieve management goals, define priorities to resolve conflicts among concurrent processes, and schedules the management processes. Therefore, policies are recognized to not be actions in themselves, but mechanisms used to make the decisions.

2.2.4 Policy Language

In [26], Marzullo defines a rule-based language for programming the management layer in a system. The language called, Lomita, is a set of tools for building distributed management software. It allows the configuration of system components, the scheduling of management tasks, and the monitoring of management software. Each system component is represented as an object with three kinds of attributes: properties, which represents the states of the management application; sensors, which are functions that returns the values of the application's states and environment; and actuators, which can set new values of the application's properties. The two later attributes are similar to the set/get operations defined by the Simple Network Management Protocol (SNMP) [27]. Policies are represented in the form "if condition then action". The sensors provide a way of detecting policy conditions, in order that, the actuators perform the policy actions.

2.3 IETF Policy Framework

2.3.1 Introduction

Policies provide a mean to manage large, heterogeneous and evolving systems [28]. A traditional approach to manage such systems consists of the usage of a centralized management application that alters the behavior of each component comprising the network. The management software should have explicit knowledge of the management interfaces of each managed entity, hence, the management of standard interfaces common to the majority of the components of a network [28].

The Internet Engineering Task Force (IETF) Policy Framework [28] represents an alternative approach of network management. Instead of centralizing the network

management into a single software entity, the framework centralizes the storage of predefined rules. These rules are comprised of conditions and actions that are designed to be device and vendor-independent. Hence, the Policy Rule is the common point between entities that are involved in a policy system.

2.3.2 Policy Schema

The IETF Policy Schema [29] defines an LDAP [30] schema to represent and store policies and the characteristics of devices controlled by policy rules, as well as, the interpretation of actions and conditions represented in a policy rule. The schema consists of thirteen classes organized in two class hierarchies: structural classes and relationship classes, as illustrated in figure 2.3.

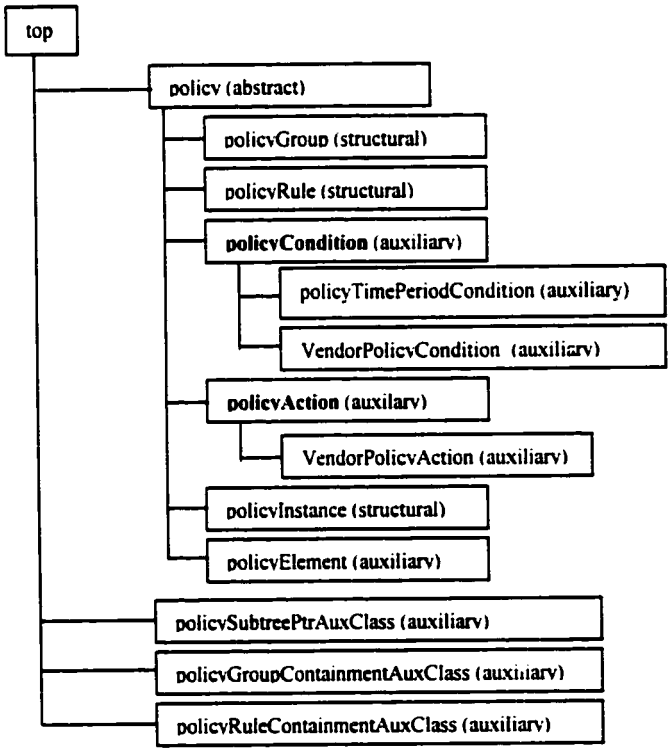


Figure 2.3: Class hierarchy in core policy schema [29]

The structural classes represent policy information and control of policies. For example, structural classes **policyAction** and **policyCondition** define the conditions and the actions of a policy.

The second type of classes represents the relationship between the instances of the structural classes. They are specific to the LDAP core schema definition. Hence, these classes are implementation-dependent and may differ if other storage technologies are used.

2.3.3 Policy Definition

In [28], Policies are defined as a set of rules, and rules are a set of conditions and actions. Policies can contain other policies in a nesting fashion, or reference a group of rules, but not both. The Policy Schema [29] defines two relationship classes: **policyGroupContainmentAuxClass** and **policyRuleContainmentAuxClass**, to represent the nesting relationship between policies and rules. Figure 2.4 illustrates the nesting relationship.

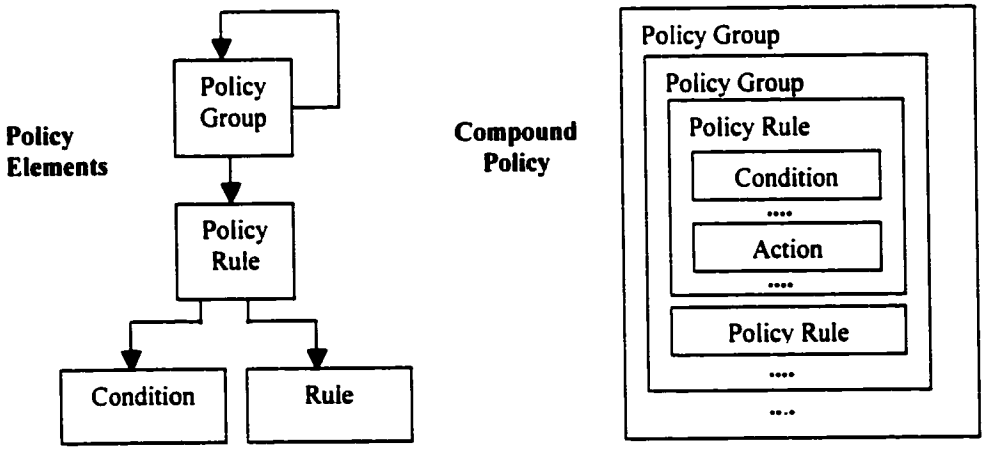


Figure 2.4: Composition of Policies [29]

Policy rule classes are structured in a hierarchical way similar to the structure of tables in SNMP's structured management format in SMI. Also, policies are specified in terms of interface functionalities, rather than, being specified explicitly for each network device.

2.3.4 Policy Architecture

The framework is comprised of two main components: the Policy Decision Point (PDP) and the Policy Enforcement Point (PEP) [28]. The PDP retrieves policies from a policy repository using the LDAP protocol or other access protocols, detects policy conflicts, translates them in a form understood by the PEPs, and then stores them in a local Policy Information Base (PIB) [31]. It sends policy elements to the PEPs based on theirs requests, or updates of policies from external entities.

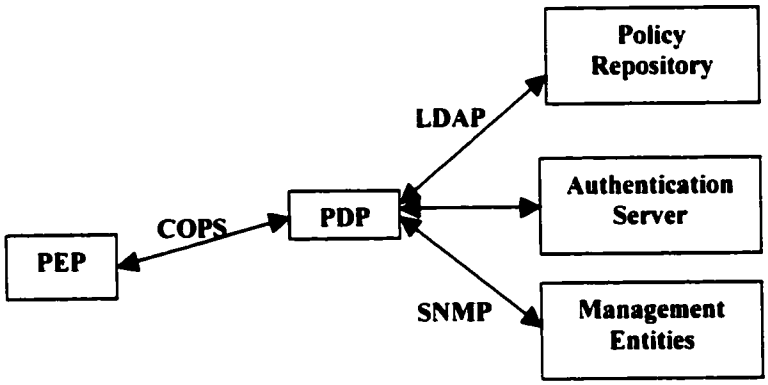


Figure 2.5: IETF Policy Framework Architecture [29]

PEP is an interface to one or more network device. It informs the Decision Point about its interfaces, and applies actions according to the PDP decisions. The separation between the PDPs and PEPs is a logical separation based on functionality. The PDP can be local to the PEP so the policy decision can be made at the level of the PEP.

PEP and PDP exchange their information using the COPS [32] protocol. The two components have to examine periodically the network elements to check whether the policies are satisfied.

2.3.5 Policy Storage & Retrieval

The framework defines a centralized repository that monitors information storage and retrieval [28]. Efficient retrieval is a key element in the performance of the decision process. For that reason, the PDP needs to retrieve, in real time, relevant policies for a specific request from a PEP. The retrieval must be done with minimum requests and has to extract the exact policy information, no more or less. The Core Policy Schema defines simple rules designed for efficient retrieval, and complex rules are designed for data efficiency, that is, allows maximum reusability of the rules. In a simple policy rule, conditions and actions are part of the rule itself, whereas, in a complex one they are part of the instances of the structural class. The policy directory can be a directory accessed using the LDAP protocol. However, LDAP cannot maintain consistency in large and dynamic systems. For this reason many policy management systems use database systems for their transactional operations.

2.4 Chapter Summary

This chapter gave an overview of the agent technology, and policy management definitions. The IETF Policy framework is an ongoing work that has the prospective to become a widely used standard in the future.

In this work, we use the assisted coordination approach to avoid the disadvantages of the direct communication approach. The Facilitator agent is a software layer on top of MicMac, which ensures communication and coordination between agents. Our policy definition and management is similar to some extent to the IETF Policy framework. We have focused on the definition of policies to manage agents rather than network entities. Policies are managed through a relational database and the storage representation doesn't follow the IETF Policy Framework LDAP Schema [29].

Chapter 3

Policy Management: Design & Concepts

3.1. Introduction

We present in this chapter an architecture of multiple agents for setting up and enforcing policies within each site of a virtual network. A policy server represents the global policies of the site and each agent manages its own policies. Policies are dynamically downloaded from the policy server into agents that carry the responsibility to enforce them. Agents propagate their policies to the policy server to detect any conflict that may rise between agents during dynamic mapping and resource reservation. A negotiation mechanism is provided to resolve such conflicts. An authorization-based mechanism is also provided such that agents must request authorization before performing any action, in response to which a ticket is delivered to the requesting agent for accountability and security reasons.

3.2. Personal Mobility Management System

3.2.1. Introduction

The development in the area of wireless data communications and mobile computers enables to a great extent mobile computing. Nowadays, the environment of mobile computing has improved considerably in terms of bandwidth, delay, computing power, and quality of display. However, mobile devices range from a very low performance equipment, e.g., personal digital assistants (PDAs), pagers, etc., up to very

high performance laptop PCs, causing a growing need for service adaptability. For example, PDAs are not designed to display high quality images, however, nomadic users will be charged for data that their equipments are unable to support. Hence, many nomadic computing platforms have been developed to address these issues and to provide the end-user with a rich set of computing and communication services in a transparent and integrated fashion [35]. Some of these systems use personal agents to represent the user's communication needs and help in the manipulation of this information to a mobile user. These personal agents provide the end user with the information about expected performance, control over the transfer operations using monitoring policies.

Such systems aim to provide an open, intelligent, secure frameworks for providing access to legacy systems, the Internet, and partners' networks that extends the enterprise network beyond existing organizational boundaries [35]. They provide manageable and secure infrastructure, along with essential applications and services, that allow nomadic users to access information across interconnected networks, while roaming from one location to another [39]. Based on the premise that the organizations manage people and their access privileges, a nomadic computing platform enables organizations to extend their network beyond traditional boundaries in a secure and manageable manner to encompass employees or associates anywhere, and build enterprises to enable mobile computing and virtual workgroups.

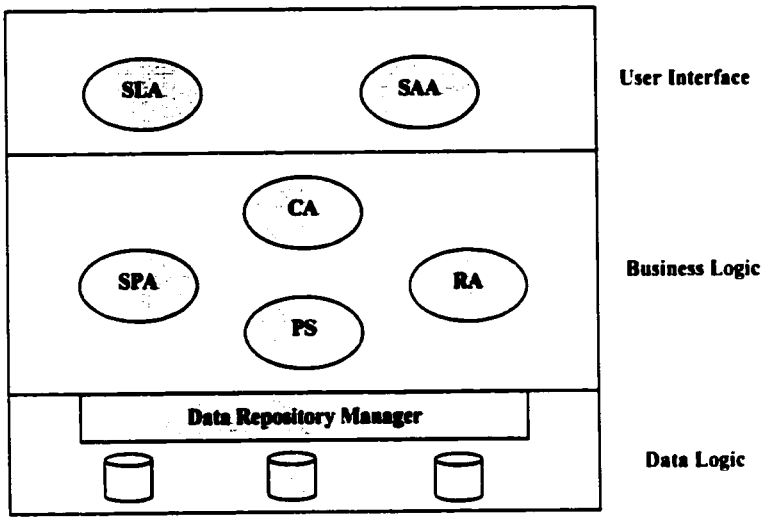
An example of a nomadic computing platform is the Personal Mobility Management System (PMMS) developed by the Multimedia and Mobile Agent Laboratory [33]. It aims to provide an integrated platform for mobile users to access different legacy systems while roaming into different locations.

3.2.2 Architecture

Personal Mobility Management System (PMMS) provides personalized services and access to resources for nomadic users over the network [33]. The application aims to offer a nomadic user a working environment similar to her home environment. In the home site, a user has a profile that describes her preferences: the services she uses, the quality of service required, the cost she is willing to pay to use some billing services, etc. A dynamic mapping between the user profile and the local policies on each site determines a set of services that a visiting site can offer to its guest. A nomadic user may be requested to pay a certain cost to use some services. In this case, agent negotiation techniques is required between the home site and the visiting site to agree on the cost or to find a service that can fit the requirements of the user with a lower cost. Negotiation is also used for resource reservation before a user starts a service on visiting site.

A prototype [33] of the system has been implemented on three sites namely: University of Ottawa, Mitel corp. and National Research Council of Canada.

As shown in Figure 3.1, the PMMS architecture is organized in three layers. The first layer is the data layer that contains all the data information of a site, such as, users' profiles, services' description, and system policies.



SLA: Site Logon Agent.
SAA: Service Access Agent.
CA: Coordinator Agent.
SPA: Site Profile Agent.
PS: Policy Server.
RA: Resource Agent.

Figure 3.1: PMMS Architecture

The second layer is comprised of four agents implementing the logic of the system. The coordinator agent manages the inter-agent communications. It's responsible to monitor agents' communication by establishing, closing, suspending, or activating communications between two agents.

The site profile agent negotiates with its remote peer (at the user's home site), the user preferences, and maps them on the local policies. The result of the mapping is a set of services that a site can offer to its visitors. Resource agent manages system resources and is responsible for the resource reservation. Moreover, the system contains a set of agents each one of them representing a service or a device. Each one of these agents reserves necessary resources for the execution of the service it represents.

The third layer defines the user interface to the system. It's comprised of two agents: User Logon Agent and Service Access Agent. The User Logon Agent is responsible for the

authentication of the user, her localization on the virtual network, and opening a new session for her. The Service Access Agent displays to the user the authorized services in the visited site and invokes the service agent when the user requests the service.

3.2.3. Agent Communication

Agents communicate using a reduced set of performatives of the Knowledge Query Manipulation Language (KQML) [11]. Messages are transported using MicMac [34], a product of Mitel Corporation. MicMac is a set of software tools, which provides a mean for communication among multiple, distributed agents. The mechanism of communication is based on the Blackboard paradigm. A tuple space, called agora in MicMac terminology, is an open area for exchanging information between agents. In PMMS, a tuple is created by the coordinator agent and can be used by any number of agents for their communication. Private agoras can also be created for private agents' communication. MicMac implements a shared memory, which stores information as tuples. Tuples are defined as unordered set of fields represented as "key-value" associations. It defines the following primitives to interact with shared memory:

- Post: Adds a new tuple to the shared memory.
- Pick: Reads and deletes a tuple from the shared memory.
- Peek: Reads a tuple from the shared memory.
- Cancel: Terminates a pending pick/peek request.

3.2.4 Policies in PMMS

A site accessed by users belonging to different organizations may be vulnerable if there is no rules to monitor their work sessions. Users always want better services, but are not always willing to pay the higher costs involved. Hence, there is a need to authenticate those that request site services and determine if they are eligible to use them or not. In [35], policies are defined as regulations that should facilitate the accountability of proper usage of communication resources, maximize the availability of communication access for users, provide security definitions for accessing a device and its corresponding services, and facilitate the creation of varied secure regions within an organization. The solution described in [35] for including policies in the system is to add certain attributes to the site related data that are stored on a Lightweight Directory Access Protocol (LDAP) [30] directory of each site. These attributes describe the services available in the site and the restricted areas that can only be assigned to privileged users. However, this approach suffers from a number of weaknesses. Firstly, no standard representation of policies is provided. Secondly, the policies are dependent on the system and co-located with the description of the system. Also, the policies are not classified according to the activities they monitor in order to facilitate their enforcement into the system. Furthermore, no mechanisms are provided to detect and resolve conflicts that are inevitable between enterprises.

The work of this thesis aims to provide a mean to define policies for each site involved in the virtual network. It provides the components to decide when they apply and how to enforce them. Policies are used to dynamically map the users' profiles and services offered by a site, to reserve resources necessary for the execution of the services

requested by the mobile users, and to monitor the actions of agents during their executions.

3.2. Policy Management

3.2.1 System Architecture

We describe in this section a policy management architecture to define and manage policies in a co-operative system, so that agent operations conform to the business goals of the organization that operates each site.

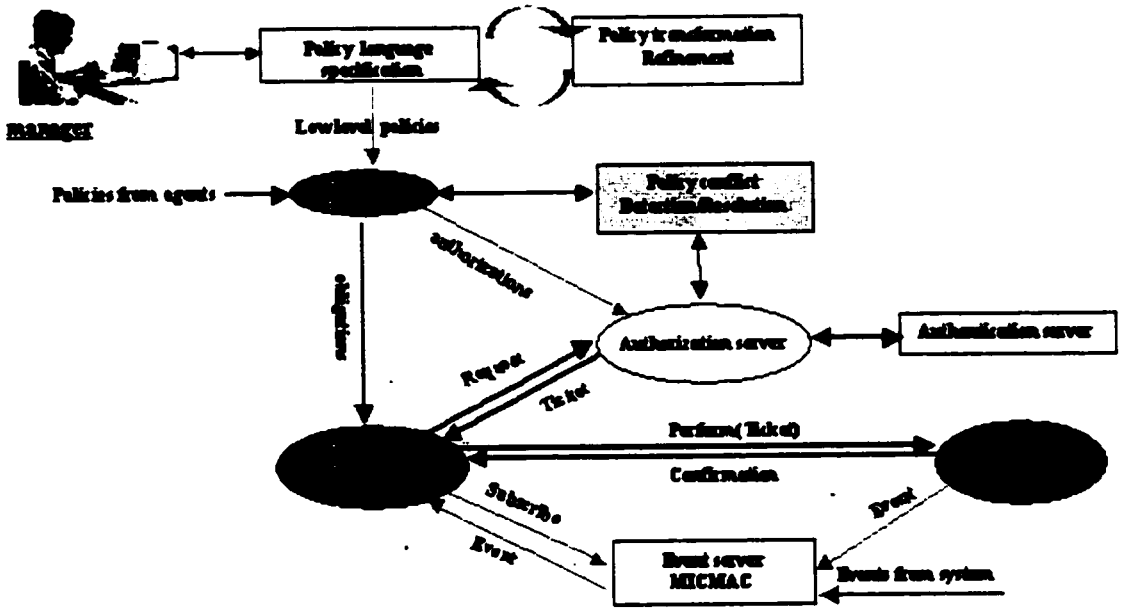


Figure 3.2: Site Policy Architecture

Agents are autonomous entities that behave according to theirs obligations with respect to the overall policies in the system. Policies are divided into two types: obligations and authorizations [18]. Obligations specify the motivations or tasks an agent is responsible to perform as part of its role. Authorizations represent the laws of the system and determine whether an agent is allowed to perform a particular action, or not.

Agents can interchangeably play two roles: requester or executor. A requester agent is an agent that requests another agent, i.e. an executor agent, to perform an action in order to fulfil one of its obligations. The executor is an agent that is designed to perform a set of actions as a part of the global role it plays in the system. For example, a printer manager agent is designed to execute printing actions, whereas, a word application agent may request the printer manager agent to print a document. Eventually, an agent could also be the requester and the executor of the same action. The division of a task into a request action and an execution action is logical division in order to ensure that no action will be performed unless it's authorized by the system policies. Hence, policies determine whether the executor should perform the action for a requester agent.

3.2.2 Policy Rules

Policy Rules are a set of actions to be initiated when a specific set of conditions are satisfied. They provide administrators the ability to manage the behavior of agents in each one of the sites involved in the virtual network. They are usually produced by the site administrators or by agents themselves and stored in the policy repository. The policy repository may be implemented as an LDAP server [30] or relational database, or any other data management system. In some cases, conflict detection and resolution may be first performed before storing the policy in the repository.

A policy rule may take one of the two following forms: authorization or an obligation.

3.2.2.1 Authorization

An authorization is a relationship between an agent, called requester, and an action of another agent called executor. Each authorization is an object with the following attributes:

Id: Uniquely identifies the authorization in the system.

Mode: The mode of an authorization can be either permission or an interdiction.

Subject: specifies the agent, also called requeter, to which the authorization applies. The subject may be specified as a role rather than an individual agent. Permission allows the subject to request that the executor agent performs an action. An interdiction prohibits the execution of the action.

Action: specifies a set of operations that the requester agent is authorized or prohibited to perform depending on the mode of the authorization. The set of operations associated with an authorization may be ordered or unordered.

Target: specifies the agent, also called executor, which executes the action. As for the subject attribute, the target attribute may specify a role rather an individual agent.

Constraint: The applicability of an authorization can be limited using a constraint. The constraint determines the conditions that must be verified in order that the authorization can be granted.

Status: A policy can have two status, enabled or disabled. Enabled policies should be enforced whenever the conditions of the constraint are satisfied. Disabled policies are not to be enforced while they are in the disabled status.

Priority: specifies the scope of the policy. A policy, with a high priority, overrides low priority policies. The priority of a policy will determine it's precedence when more than one policy is initiated by the same constraint.

Class: classifies policies according to the activity they monitor in the system. Examples of classes are security, accounting, and admission control.

System mode: A system mode represents a specific state of the system. Policies belonging to a system mode define the strategy of the system in this mode.

Creator: specifies the agent or the human manager that adds the policy to the system.

For example, to express an authorization to monitor resource reservation that may be done by a video agent providing a video service over the network, the system manager may add the following policy:

Id: A0001

Mode: Permission

Subject: Video Agent

Action: Reserve Bandwidth

Target: Resource Agent

Constraint: (Reserved Bandwidth < 10% of System Bandwidth && Cost > 12/unit)

Status: Enabled

Priority: 5

Class: Resource Reservation / Bandwidth Reservation

System Mode: Normal Resource Mode

Creator: System Manager

The Video Agent is only able to reserve 10 percent of the available bandwidth in the system and have to pay at least 12 units for each unit of the reserved bandwidth.

A label **any** can be used instead of a specific requester, action or executor. A default authorization can be set to determine the default behavior of the system. The default authorization can be:

Permission any any any: permits to any agent to request any action from any other agent in the absence of a specific interdiction.

Or

Interdiction any any any: Prohibits any agent from requesting any action from any other agent in the absence of a specific permission.

3.2.2.1 Obligations

An obligation is a duty an agent should perform once a set of conditions are satisfied. The duty can be represented as a set of ordered or unordered actions. The obligation may specify the agent that should be requested to execute the action or leave it up to the responsible agent to look for the suitable executor agent. However, the second option implies that the system has a resource discovery mechanism that allows the agents to report and knows each others capabilities. In this work, we use a facilitator agent to store all agents' actions, and acts as a broker between requesters and executors. We represent an obligation as an object with the following attributes:

Trigger: an internal or external event that triggers the obligation. The event can be a result of an action performed by another agent or a specific time, etc.

Subject: specifies the agent responsible of the execution of the obligation when it's initiated.

Action, target, constraints, class, and system mode: These attributes have the same meaning as for the authorization's attributes.

Exception: specifies an action that the subject must perform in the case it fails to accomplish the actions specified by the obligation.

Obligations and authorizations policies represent the global strategy of a site. They control behavior of the agents belonging to the site and determine how agents should react to any changes in the system. They are classified according to the activity they monitor. Four classes of policies are defined for the implementation of this work: security, admission control, user profile, and resource reservation. These classes of policies are used by the coordinator agent, site logon agent, site profile agent, and resource agent, respectively.

During its life cycle a policy may have different status depending on the requirements of the system. Hence, a policy may be enabled, that is, should be enforced whenever the conditions making its constraints are satisfied. In the contrary, a disabled policy should not be processed while it's in the disabled status.

4.2.3 Policy Editor

The policy editor is an interface to the policy system that enables the editing of a policy. Policy editing includes viewing, entering, modifying and deleting a policy rule. It may be implemented as a graphical user interface, preferred nowadays by most system administrators, or/and as a scripting interface. Policy rules may be expressed in a user friendly language. Hence, they may need to be transformed to a specific format that can be understood by the system agents. The policy translation is carried out by a policy

translation and refinement module, which can be seen as an intermediate layer between the human management and agent management. In fact, the policy translation has for purpose the mapping of the high level policy specified by human administrators to an information schema particular to each system. Thus, the refinement process may result in one or more low level policies depending on the complexity of the policy language specification. For example, an administrator who wants to secure the system may add a policy that prevents foreign users from accessing the system. However, in a big system, it's most likely that the administrator doesn't know all the specific agents carrying the responsibility to protect the system nor the actions to be executed to perform such a goal. Hence, the added policy can be expressed in a high level language such as "put system security to high". A refinement of such a policy may result in more specific policies, i.e. low-level policies, with deterministic actions that should be enforced by system agents.

3.2.4 Policy Validation & Conflict Resolution

The resulting low-level policies of the refinement process should be checked in order to not conflict with existing rules. Likely, a new entered policy stating that no user can access the system during a certain period of time, will conflict with an existing policy allowing some privileged users to access the system anytime. Hence, some mechanisms are required to detect and eventually resolve such conflicts. An obvious mechanism is to set priorities between policies, that is, a policy with a high priority will have precedence on low priority policies. A similar approach is used in [3] to resolve conflicts. In the case that two policies with the same priority conflict, precedence based on the scope of the policy may be used. For example, if a policy P1 is a subset of policy P2, i.e. P1 is always initiated whenever P2 is initiated, then policy P1 should have precedence on policy P2.

Using this conflict resolution mechanism, administrators will be able to specify a general policy denying access to all users during a certain period of time of the day, while another policy may allows some privileged users to access the system during the period prohibited by the general policy. These conflict detection rules can also be expressed as meta policies, which are policies that apply for other policies. In the case that the system cannot automatically resolve the conflict the new added policy should be rejected and a notification should be sent to the entity entering the rejected policy, whether it's a human administrator or a system agent.

However, policies may be initiated based on a specific state of the system that may not be known during the insertion of the new policy. In this case, the conflict may raise at the time of enforcement of actions specified by the policy. Hence, a dynamic conflict detection mechanism has to be implemented. Also, more elaborate conflict resolution mechanisms are needed to resolve these dynamic conflicts. Agent negotiation is one of such mechanisms that may be used, that is, agents will try to reach an agreement in order to resolve conflicts.

3.2.5 System Agents

As defined before, agents are autonomous entities behaving to fulfill their obligations. Each agent has set of defined roles in the system. For example, agents may be an interface to a device such as a router or a printer, or acts a resource broker reserving system resources for other agents. By this mean, a policy is specified for a particular role instead of an individual agent. It is a powerful mechanism that enables generation of site-wide policies independently on the existing agents in each site. Each agent should report to the policy server the roles it plays in the system. For example, if an agent is

responsible of user authentication and data encryption, then it has to report these two roles to the policy server, so the latter can distribute to this agent any obligation that applies to these two roles.

An agent can be seen as an aggregation of two different modules: a code that represents the actions the agent is designed to execute and policies that control the behavior of the agent. Thus, policies can be modified in order to change the behavior of the agent without changing the code of the agent. Furthermore, agents can be reused in different sites where different policies apply. This property is useful in the case of mobile agents, i.e. agents able to move across different network boundaries. An agent that moves from a site to another site will receive its policies from the new policy sever without need to change its code. Hence, it will continue to perform its roles according to the new site goal.

Each agent has a set of obligations it is responsible to fulfill and the associated policies will determine whether or not the agent can accomplish these obligations. When the conditions of an obligation apply, the agent responsible should perform the set of actions representing the duty of the obligation.

In PMMS system, policies represent the knowledge of the agents because they instruct them what are theirs duties, when to perform them and whether they are authorized to perform them or not.

Since agents have to collaborate to perform theirs tasks, there is a need to identify agents in the system by one or more identifiers in order facilitate communication between agents. Also, they can be assigned a group identifier depending on the roles they play in the system in order to allow for message broadcasting. For example, instead that the

policy server distributes a security policy to each agent responsible of security enforcement, it may send one message with the security identifier as receiver.

3.2.6 Event Server

A common behavior to all agents is that they are designed to report any event they internally produce, and subscribe to be notified when a new event occurs. The event server provides the mechanisms that allow agents to report and subscribe to events. It has as role to collect events from agents and eventually the system and notifies subscribed agents to these events. For example, an email forwarder agent might be interested in the email-received event. Each time an email is received, the email server will notify the event server. The latter will dispatch the event to all agents that are subscribed to it. The email forwarder agent could then accomplish one of its obligations, which consists of forwarding the received email to the current location of the receiver when she is away from the home site.

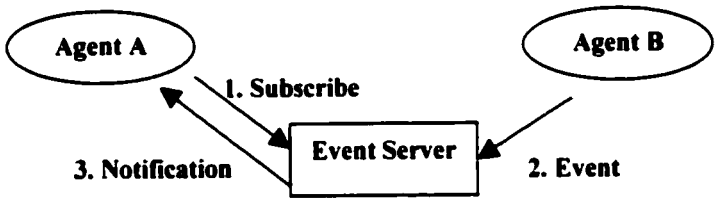


Figure 3.3: Event Notification Mechanism

Figure 3.3 shows the mechanism of events notification. In step 1, agent A subscribes to an event it needs to be notified when it occurs. The event producer, in the

figure Agent B, notifies the event server when it produces the event. The event server in step 3 dispatches it to all subscribed agents, for example agent A.

The event mechanism described above avoids the overhead generated by the polling mechanism. In a polling mechanism, agents have to periodically poll the event producers in order to track the occurrence of the event on the system.

The event server used in the implementation of this work is the product of our partner MITEL called MICMAC. MICMAC implements the blackboard paradigm for the creation of shared communication space called agora or tuple spaces. A tuple space is an open area for exchanging information [34]. It also provides the low level communication medium to the agents by enabling the creation of private and public communication channels.

3.2.7 Facilitator Agent

The Facilitator Agent is a yellow-pages that describes the services that can be provided by agents. Agents register their services with the Facilitator Agent, and can find out what services are provided by other agents on specific domains. It should keep an accurate, complete and timely list of the most current information about agents and their services in its directory. Agents need a service discovery mechanism in order to be able to request a new service from other agents. For example, if an agent would like to initiate a long distance call on behalf of the user, it needs to use the Facilitator Agent to discover the set of agents representing different carriers like Bell Canada, AT&T, Sprint Canada, etc. Hence, an agent can discover and potentially choose the cheapest, most reliable carrier able to make a particular long distance call.

3.2.8 Policy Server

The policy server is the main component of the system. It receives policies from agents or from human managers via interfaces. Obligation policies are downloaded into agents that are responsible to perform them. Authorization policies are distributed to an authorization server, which gives agents necessary authorizations to perform a set of actions. The policy server is responsible for the management of policies' status. Each policy can be, during its life cycle, enabled, disabled or deleted. Before accepting a policy, server ensures that the new policy doesn't conflict with existing policies in the system, referred as called static conflict detection. Run time conflict detection is also used to detect conflicts that may rise after a policy is enforced in the system. It's described in greater detail in chapter 4.

Policy server is notified by agents when any change in the system occurs. It reacts to these changes by enabling or disabling policies that already exist in the system, or distributing new policies. Thus, the system behavior is adapted to meet the requirements of new system states.

3.2.8.1 Notification of Agent Roles

Since agents are autonomous entities, they can change their roles without referring to the policy server. Hence, they must report to the policy server the roles they are playing on the system as part of their obligations. An agent may be responsible of several tasks and the policy server has to know about these roles in order to be able to distribute any new set of policies to the right agents that should enforce them in the system. Role notification can also be done upon request of the server as an auditing

operation in order to ensure the integrity of the system. The underlying protocol for role notification is based on two operations: Notify_Role, and Get_Roles.

- **Notify_Role(Agent_Id, Roles)** : By this operation, any agent can send to the policy server the set of roles it plays in the system. Agent_Id is a unique identifier of the agent, and Roles is the set of agent's roles. For example, an Email Agent responsible of sending and receiving users' emails may notify the server of its roles by sending the following message:

  Notify_Role("Email_Agent","Receive_Email,Send_Email")

- **Get_Role(Agent_Id)**: This operation is used by the policy server to know the roles of a specific agent identified by Agent_Id. For example, the policy server may send the following message to learn the roles of Email Agent: Get_Roles("Email_Agent"). The email agent replies by a Notify_Roles message. A broadcast message can also be sent by setting a group identifier instead of an agent identifier.

3.2.8.2 Policy Distribution

As described earlier in section 3.2.2, policies can be either an agent obligation, or a system authorization. System authorizations are automatically distributed to the authorization server that is responsible to deliver necessary authorizations to agents when they need to request an executor agent to perform a specific action. Obligation policies are downloaded to agents based on theirs roles. Once a new obligation is added to the system, policy server determines, based on the class of the policy, to which agents the policy should be sent. A policy that puts some constraints on the time when users can log on to the system is classified as an admission control policy. It should be distributed to

any agent that has as role to control the user's access to the system. Policies can also be retrieved as part of the distribution process. They may be retrieved because of many reasons such as date expiration or a change of the management requirements.

We define three operations for policy distribution:

- **Set_Policy (Agent_Id, Agent_Role, Policy)**: the policy server uses this operation to distribute any new Policy *p* to the agent identified by the Agent_Id. The role of agent to which the obligation apply is set by the Agent_Role parameter.
- **Set_Authorization (Authorization):** Used only to distribute authorizations to the authorization server.
- **Retrieve_Policy(Agent_Id, Policy_Id)**: Retrieves a policy from an agent. The policy can either be an authorization or an obligation. In the case the policy server needs to retrieve the policy from the system, an Agent Identifier referring to all agents can be specified instead of an identifier for a particular agent.

3.2.8.3 System Adaptation

The policy server is responsible to monitor the behavior of agents to meet the global system requirements. System monitoring consists of ensuring that enabled policies enforce the right behavior in each managed site. We define a set of predetermined system states, or system modes, and classify policies according to these modes. Policies belonging to a class mode are enabled when the system mode is enabled. They adapt the behavior of the system to the requirements of a new system state. For example, critical resource mode is enabled when the amount of resources available in the system is under a limit. Thus, policies belonging to this mode are enabled and will define a new resource

allocation strategy in order to prohibit agents from reserving excessive amounts of resources.

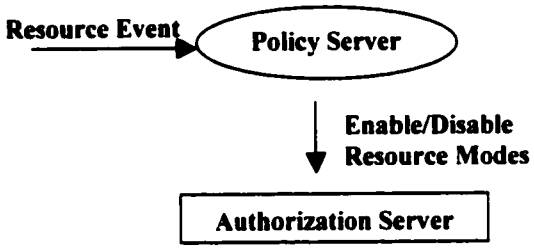


Figure 3.4: System Adaptation

Figure 3.4 illustrates the steps for the adaptation of resource modes upon the receipt of a new event from an agent responsible of the management of system resources, e.g. Resource Broker Agent. The policy server responds to this particular event by enabling policies belonging to critical resource mode and disabling policies belonging to normal resource mode.

We define two operations for enabling and disabling system modes: Enable Mode and Disable Mode.

- Enable_Mode (SM): Enables all system authorizations belonging to the system mode SM.
- Disable_Mode(SM): Disable all system authorizations belonging to the system mode SM.

We note that different system modes can be related: when a mode is enabled, other modes are automatically disabled. For example, when critical resource mode is enabled, normal resource mode is automatically disabled.

3.2.9. Authorization Server

The authorization server agent represents the laws of the system, which are expressed as authorizations. The authorization server is responsible for delivering necessary authorizations to agents requesting the execution of an action. Each agent who requests an authorization is first authenticated. Authentication is often established through third party authentication servers like Kerberos authentication server, Radius, via a certificate authority like X.509, or by sharing public and private keys. Then, the server retrieves policies that apply to the request, evaluates them and then delivers a response to the requester agent.

According to the present work, each agent must be authorized before executing any action. To illustrate how this mechanism is effected by agents requesting authorization to execute an action, an example is set forth by which a Word-Agent requests authorization to print a document using a Printer-Manager.

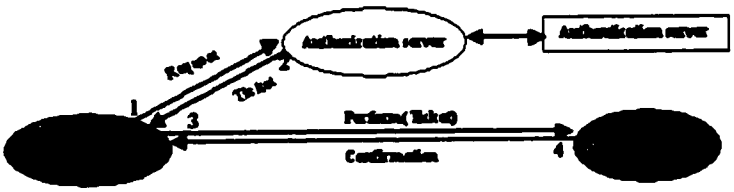


Figure 3.5: Steps of Requesting an Authorization

1- The agent requester, e.g. Word-agent, sends a request to the authorization server asking for an authorization to execute an action, e.g. print a document. The request

message includes the agent identification, the action and its parameters, and the identification of the agent executor of the action.

2- The authorization server authenticates the agent and processes the request. The authorization server then delivers a ticket that specifies the result of the processing of the request. In the case of a negative authorization, the ticket contains the reasons of failure and offers alternative choices to the requester. The requester can then modify the request to fit the conditions required by the policy server.

3- If the request is authorized, the requester sends the ticket to the agent executor of the action, e.g. printer-agent. The executor verifies the authenticity of the ticket and executes the action.

4- A confirmation is sent back to the requester agent to confirm or reports the failure of the execution of the action.

The ticket is a token that allows an agent to operate with the rights and privileges of the authorization server that granted the ticket. Naturally, it must be possible to verify that the authorization server granted the ticket. The validity of the ticket can be verified by using a public key digital signature. The ticket may include an expiration time. If a non-expiring ticket is desired, the expiration time can be set sufficiently far in the future. Furthermore, the ticket mechanism allows the agents to bill for the services they offer.
In the case the policy doesn't specify the agent executor of the action, the agent uses the Facilitator Agent to discover the set of agents providing the service. Hence the Facilitator forwards the ticket to these agents and offers will be made to the requester agent. The agents interested in this offer can then reply to the requester agent proposing to execute the action, and may specify additional constraints to execution of the action (e.g. more

cost), or provide some additional information about the execution (e.g. quality of printing). The requester chooses the best offer, or negotiates, according to the model described in next chapter, with agents that are providing the best offer.

3.3 Chapter Summary

In summary, this chapter provides a multi-agent architecture for setting up policies in a cooperative network according to different policy classes based on the activities they monitor. The advantage of this architecture is that it provides a dynamic way to adapt the behavior of agents to any changes in the system. Policies can be quickly setup to control the behavior of agents and make them behave according to the global strategy of the system. In the next chapter, we provide an example of policy definition and management in one of the sites involved in the virtual network, as well as, the negotiation model used to resolve conflicts between agents.

Chapter 4

Policy Management in PMMS

In order to better understand the concept of policy classification and the effect of system modes, this chapter describes an example of a policy management system, with reference to a working site located at the University of Ottawa. The example illustrates the classification of policies and how they control the behavior of agents belonging to the managed site.

4.1 Classification of Policies

As discussed in the previous chapter, policies dictate the global strategy of a site. Policies control the behavior of the agents and determine how they should react to any change in the system. Policies are classified according to the activity they monitor. Four classes of policies are defined for the management of the PMMS system: security, admission control, user profile, and resource reservation. These classes are used by the coordinator agent, site logon agent, site profile agent, and resource agent, respectively.

4.1.1 Default Policies

A default policy is defined with a low priority that is applies to all requests. The default policy is used to ensure that at least one policy will apply for a request. The default policy is expressed as follows:

ID	Mode	Subject	Action	Target	Constraint	Priority	System Mode
A000	Interdiction	*Any*	*Any*	*Any*		0	-

Table 4.1: Default Policy for the Site

The default policy is identified as A000, where A denotes that it is an authorization object. By default, the execution of any action in the system is forbidden. The low priority ensures that this policy will be overridden if at least one other policy applies to the request.

4.1.2 Security Policies

Inter-site communications are encrypted using different encryption algorithms such as DES and RSA [47]. Each encryption algorithm can use a different key length to encrypt data. Encryption policies specify which algorithm should be used for a specific communication between two different sites. Two modes are defined for system security: normal level and high level. Each mode has a set of policies that define the rules that the system should enforce when the mode is enabled. The following policies give an example of some security policies setup by the University of Ottawa to encrypt its communications with the other two sites.

ID	Mode	Subject	Action	Target	Constraint	Priority	System Mode
A001	Interdiction	*Any*	Encrypt	Security	Key_length<64	4	Normal security level
A002	Permission	Coordinator	Encrypt	Security	Key_length>64 and site = MITEL Or Key_length>128 and site = NRC	3	Normal security level
A003	Interdiction	Coordinator	Encrypt	Security	Key_length < 196	5	High security level

Table 4.2: Security Policies

The first policy (ID A001) applies to all requesters of an authorization to encrypt a communication. It specifies that it is forbidden for any agent to request a security agent to encrypt data with a key length less than 64 bits. This is a default security policy and sets the minimum degree of security needed for inter-site communications. The second policy (ID A002) is specific to the coordinator agent. It allows this agent to request encryption when it is establishing communication with the MITEL site using an encryption key length greater than 64 bits, or with the NRC site using a key length greater than 128 bits. University of Ottawa assumes that the connection with NRC is not secure enough. Thus, the key length used for encryption should be greater than 128 bits in length to ensure privacy of communication on top of an insecure communication channel. When the system security mode is set to high, the third policy (ID A003) sets more constraints on the key length.

4.1.3 Admission Control Policy

Admission control policies express rules, which are used to admit a new user into a visited site. They specify eligible users who are provided with access to the site and impose constraints on the login of these authorized users. These policies apply for the site login agent and apply each time a new user wishes to log onto a visited site. Some

examples of admission control policies used by the University of Ottawa site are set forth in Table 4.3, from which it will be noted that the subject and the target are the same agent.

ID	Mode	Subject	Action	Target	Constraint	Priority	System Mode
A004	Interdiction	Site logon	login	Site logon	{S} day = Sunday	4	Normal load
A005	Permission	Site logon	login	Site logon	User.Site = NRC Or (User.Site = MITEL and User.Category= Manager)	2	Normal load
A006	Interdiction	Site logon	login	Site logon		3	Heavy load

Table 4.3: Admission Control Policies

The first policy (ID A004) prohibits access to any user on Sundays. The second policy (ID A005) permits access to all NRC users but restricts access of MITEL users to MITEL managers. When the system switches to heavy load mode, the third policy (ID A006) is enabled, thereby prohibiting access to the site.

Examples of a second type of policy are shown in Table 4.4, where the ID is prefaced with an "O", which indicates an "obligation" object. The site logon agent is obliged to notify the event server (MicMac) when the number of users in the system reaches a predetermined limit. The related obligations are expressed as follow:

ID	Trigger	Mode	Subject	Action	Target	Constraint	System Mode
O001	On login	Obligation	Site logon	Notify("Heavy load")	Event server	{S} users = 7	Normal load
O002	On logout	Obligation	Site logon	Notify("Normal load")	Event server	{S} users = 4	Heavy load

Table 4.4: Site Logon Obligations

The notification of the first obligation (ID O001) is used by the policy server to switch the system mode to heavy load mode, whereas the notification of the second obligation (ID O002) is used to switch the system mode back to normal load.

4.1.4 User Profile Policies

The prototype of the PMMS presently in operation at the University of Ottawa is capable of offering three services: printing, email forwarding, and video mail service. A different quality of service (QoS) is used for each delivered service. For example, video service has three qualities of service. A high quality of service plays back an entire video for a user. Medium quality extracts video frames using a key framing application [1] and displays the resulting frames to the user. Low quality of service displays the resulting frames in black and white.

As discussed briefly in the previous chapter, the site profile agent maps the user's profile onto the local policies of a site. The following example defines some policies used by the University of Ottawa site to offer services to visiting users from MITEL and NRC sites. The example is limited to video service.

ID	Mode	Subject	Action	Target	Constraint	Priority	System Mode
A007	Permission	Site profile	Use_Video	Site Profile	QoS=1 and Cost >15 Or QoS=2 and Cost>25 Or QoS=3 and Cost>40	3	Normal resource mode
A008	Interdiction	Site profile	Use_Video	Site Profile	QoS=3 AND User.Category <> Manager	3	Normal resource mode
A009	Permission	Site profile	Use_Video	Site Profile	QoS=1 and Cost > 40 Or QoS=2 and Cost >70	4	Critical resource mode

Table 4.5: Policies for user profile mapping

The first policy (ID A007) sets a minimum cost required to use the video service. The cost depends on the quality of service to be used. For example, a high quality video service can not be used unless the user pays at least 45 units.

The second policy (ID A008) limits the use of the high quality of video service to managers from both sites.

The third policy (ID A009) is invoked during system critical resource mode. It raises the cost for using the low and medium quality of service and prohibits the use of the high quality of service.

Other policies are used to set more constraints with the use of video service, but are not described for the sake of simplicity.

4.1.5 Resource Reservation Policies

Each site offers its visitors a set of services resulting from the dynamic mapping of the user's profile and the local policies of the site. When the visitor starts a service, the agent representing the service reserves a predetermined amount of resources, CPU and bandwidth, required for the execution of the service with a specific quality, as set forth in Table 4.6:

ID	Trigger	Mode	Subject	Action	Target	Constraint	System Mode
O003	On startup	Obligation	Video agent	Reserve(CPU=20,Band=5)	Resource agent	QOS=1	Normal resource mode
O004	On startup	Obligation	Video agent	Reserve(CPU=35,Band=10)	Resource agent	QOS=2	Normal resource mode
O005	On startup	Obligation	Video agent	Reserve(CPU=75,Band=20)	Resource agent	QOS=3	Normal resource mode
O006	On startup	Obligation	Video agent	Reserve(CPU=15,Band=3)	Resource agent	QOS=1	Critical resource mode
O007	On startup	Obligation	Video agent	Reserve(CPU=25,Band=7)	Resource agent	QOS=2	Critical resource mode
O008	On startup	Obligation	Video agent	Reserve(CPU=45,Band=12)	Resource agent	QOS=3	Critical resource mode

Table 4.6: Video Agent obligations

These obligations define the amount of resources to be reserved for each quality of service in each mode. The video agent automatically adapts the requirements of the video application to the actual state of the system.

A resource reservation strategy is defined by policies that apply to controlling the reservation of resources via the resource agent, as set forth below in Table 4.7. Again, this example is limited to video service. In Table 4.7, and subsequently in this chapter, it should be noted that {S} stands for a system parameter (e.g. {S} Bandwidth is the

available bandwidth in the system whereas Bandwidth is the required bandwidth to be reserved by a resource agent).

ID	Mode	Subject	Action	Target	Constraint	Priority	System Mode
A010	Permission	Video Agent	Reserve	Resource Agent	{S} CPU > 200 and CPU < 30 Or {S} CPU > 300 and CPU < 50	5	Normal resource mode
A011	Permission	Video Agent	Reserve	Resource Agent	{S} Bandwidth > 40 and Bandwidth<5 Or {S}Bandwidth> 70 and Bandwidth<10	5	Normal resource mode
A012	Interdiction	Video Agent	Reserve	Resource Agent	CPU > 20 OR Bandwidth > 4	7	Critical resource mode

Table 4.7: Resource reservation policies

Policies that belong to normal resource mode are used to monitor the resource reservation strategy for this particular resource mode. The first and the second policies (ID A010 and ID A011) specify the maximum amount of CPU and bandwidth that the video agent can reserve depending on the available amount of CPU and Bandwidth on the site.

The third policy (ID A012) prohibits an excessive reservation of resources when the resource status is critical (the resource status can switch between two values: normal and critical). Critical status means that the available resources in the system are below a critical value.

After each reservation, the resource agent notifies the system if there has been any change in the resource status. This obligation is expressed as follow:

ID	Trigger	Mode	Subject	Action	Target	Constraint	System Mode
O03	On reservation done	Obligation	Resource agent	Notify("critical resource mode")	Event server	{S}CPU<100 OR {S} Bandwidth<40	Normal resource mode
O04	On resources released	Obligation	Resource agent	Notify("normal resource mode")	Event server	{S}CPU>200 And {S} Bandwidth>70	Critical resource mode

Table 4.8: Resource Agent Obligations

4.1.6 System Policies

Agents are designed to notify the policy server when a change in the system mode occurs. The policy server reacts to these changes by enabling and/or disabling certain policies in order to adjust the system behavior to the requirements resulting from the new state. These reactions are expressed as obligations:

ID	Trigger	Mode	Subject	Action	Target	Constraint	System Mode
O05	On heavy load mode	Obligation	Policy server	Enable("heavy load mode") Disable("normal load mode")	Policy server	-	-
O06	On critical resource mode	Obligation	Policy server	Enable("critical resource mode") Disable("normal resource mode")	Policy server	-	-

Table 4.9: Policy server obligations

The enabled policies guarantee that the system will behave to fit the requirements of the new system state. For example, the heavy load mode enables Interdiction A006, which prohibits any user from logging into the system.

4.2 Processing Policy Requests

Since policies are independent, and may be added to the system by different administrators, many policies may apply for the same request. The authorization server processes all of these policies in order to make a final decision as to which policy will prevail. First, the authorization server loads all enabled policies that apply for the request, and then it evaluates each policy to determine if the conditions specified by the policy constraint apply. The evaluation of a policy constraint involves three steps: constructing a decision tree, decorating the tree and extracting the decision related to the policy. The actual mode is then added to the attributes of the policy. The authorization server selects the applicable policies with highest priority in order to make the final decision concerning a request.

An example is provided below to illustrate the steps used in processing a request. The example relates to the reservation of system resources by a video agent, which executes a video service with a specific quality of service. The exemplary request is expressed as follow: **Video agent asks authorization to reserve (CPU=25, Bandwidth=15) by resource agent**. The following tables summarize the state of the system and the amount of resources available in the system at the time of the request is made.

Resource	Value
CPU	250
Bandwidth	100
Disc	60

Table 4.10: Available Resources

System Mode	Status
Normal resource mode policies	Enabled
Critical resource mode policies	Disabled
Heavy load policies	Disabled
Normal load policies	Enabled
High level security policies	Disabled
Low level security	Enabled

Table 4.11: System Modes

The authorization server first loads all enabled policies that apply to this request.

The following table gives the policies that apply for video agent requests.

ID	Mode	Subject	Action	Target	Constraint	Priority	Class
A000	Interdiction	Any	Any	Any		0	-
A010	Permission	Video Agent	Reserve	Resource Agent	{S} CPU > 200 and CPU < 30 Or {S} CPU > 300 and CPU < 50	5	Normal resource mode
A011	Permission	Video Agent	Reserve	Resource Agent	{S} Bandwidth > 40 and Bandwidth<5 Or {S}Bandwidth> 70 and Bandwidth<10	5	Normal resource mode
A021	Interdiction	Video Agent	Reserve	Resource Agent	Bandwidth > 12 OR CPU > 70	5	Normal resource mode

Table 4.12: Policies that apply for the video agent request

The evaluation of a policy involves an evaluation of the conditions representing the constraints of the policy. The policy A000 has no constraint; therefore, the actual mode of the policy is Interdiction. The policy A010 has a constraint to be applicable. The evaluation of the constraint determines if the policy applies or not to the request.

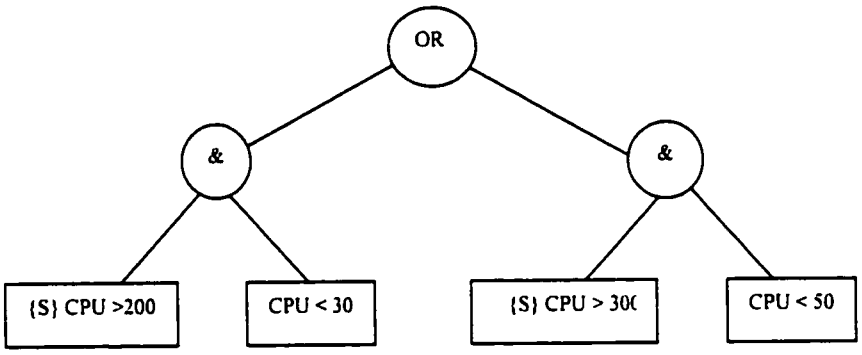


Figure 4.1: Constraint of Policy A010

The tree diagram of Figure 4.1 represents the constraint of the A010 policy. Each leaf of the tree represents a condition of the global constraints. Although the example set forth herein is characterized by a simple representation of conditions, it will be understood that more complex representations are possible. Conditions that are preceded by {S} stands for a condition that depends on the system, otherwise, the condition depends on the parameters specified by the request. The tree is decorated according to the values of parameters in the request and the actual values of system parameters (e.g. the available CPU in the system). Figure 4.2 shows the decorated tree representing the policy A010. The result of the evaluation of this policy is permission.

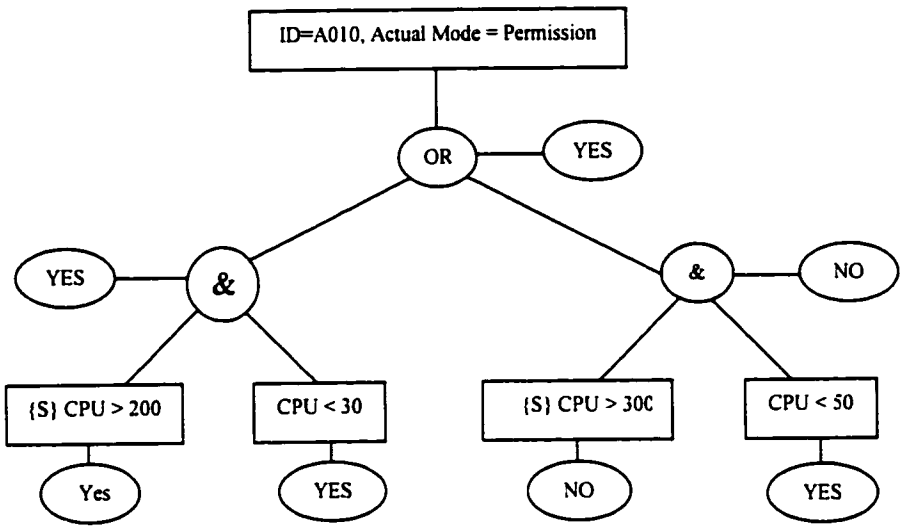


Figure 4.2: Decorated Tree of Policy A010

The policy A011 is evaluated using the same technique. The decorated tree of this policy is illustrated by Figure 4.3. This policy neither permits nor prohibits the request. Therefore, the actual mode of the policy is set to "not applicable".

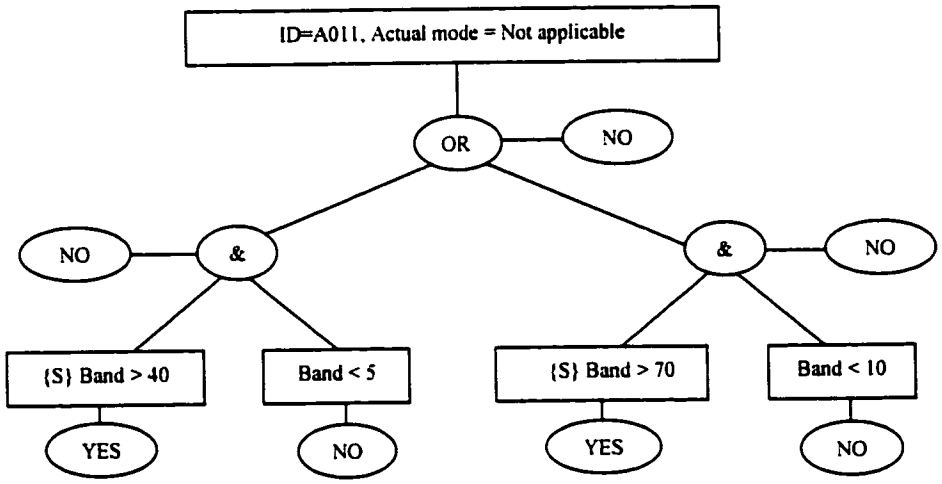


Figure 4.3: Decorated Tree Policy A011

The decorated tree of the policy A021 is illustrated by the tree diagram of Figure 4.4. The evaluation of this policy is interdiction.

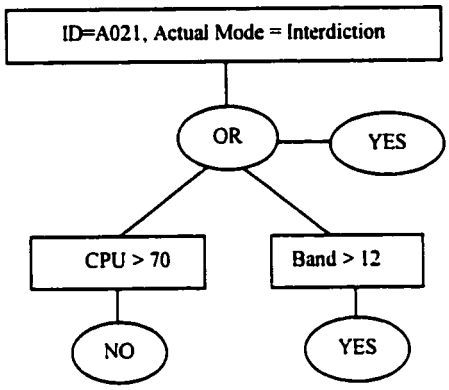


Figure 4.4: Decorated tree of policy A021

Table 4.13 summarizes the results of the evaluation of policies:

ID	Mode after evaluation	Priority
A000	Interdiction	0
A010	Permission	5
A011	Not applicable	5
A021	Interdiction	5

Table 4.13: Summary of the policy evaluation

The authorization server uses the highest applicable policies to make a decision. In this example, it uses policies A010 and A021. By default, Interdiction overrides permissions. Thus, the request is rejected.

The above example shows three different modes that result from the evaluation of policies. Another mode, conditional, is applicable if some parameters of the request are to be modified before an authorization can be granted. This type of result is used for negotiation between agents.

4.3 Negotiation Model

A request for an authorization may be refused depending on the evaluation of the authorization server due to a conflict between different policies. Therefore, the requesting agent is not able to accomplish its tasks, which may cause chaos in the system. For example, if a video agent needs to reserve an amount of bandwidth to transfer a video over the network, the request to the authorization server may be refused due to a lack of resources or insufficient cost to pay the reserved bandwidth. Hence, the conflicting agents need to start a negotiation session to resolve this conflict. The negotiation model used in this work defines a way to reach an agreement between two or more agents having different beliefs. Agreement is reached when the proposal does not violate any policy of the negotiating agents.

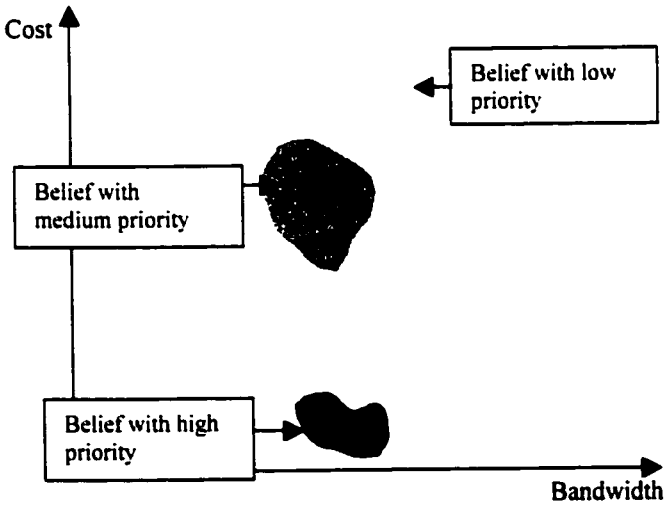


Figure 4.5: Belief of Video Agent

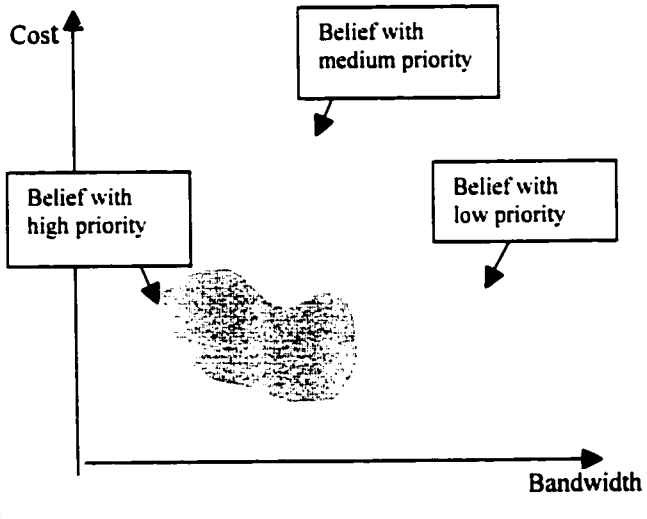


Figure 4.6: Belief of Authorization Server

We define belief of an agent as the combination of all policies monitoring that agent. Policies with the same priority define a region of belief. Depending on the priority associated with a region of belief, it can be strong, medium or weak for high, medium, and low priority respectively. As shown in Figure 4.5 and 4.6, we can represent belief of a PMMS agent as a function B that is expressed as:

$B = \sum f_p(policies)$, where each region of the graph has a priority p to express the degree of belief of the agent. For example, in Figure 4.5 the video agent strongly believes that it must reserve a certain amount of bandwidth at a small cost. On the other hand, it weakly believes (i.e. does not prefer), to reserve the same amount of bandwidth at a high cost. Also, a request, we call proposal, is represented by a point P on the graph.

A measure between a point P, i.e. a proposal, and a region of belief R_b may be defined as:

$M(P, R_b) = \min (d(P,X) + \frac{1}{p}); \quad \forall X \in R_b$; where d is the distance between two points (i.e.the Cartesian distance: $\sqrt{X^2 + Y^2}$), and p denotes the priority of the region.

In the above formula, the 1/p term expresses the tendency of an agent to prefer its strong beliefs to weaker ones.

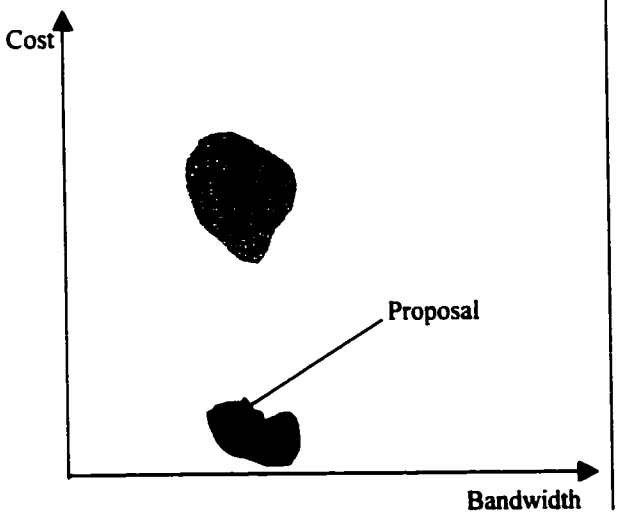


Figure 4.7: First Proposal from Video Agent

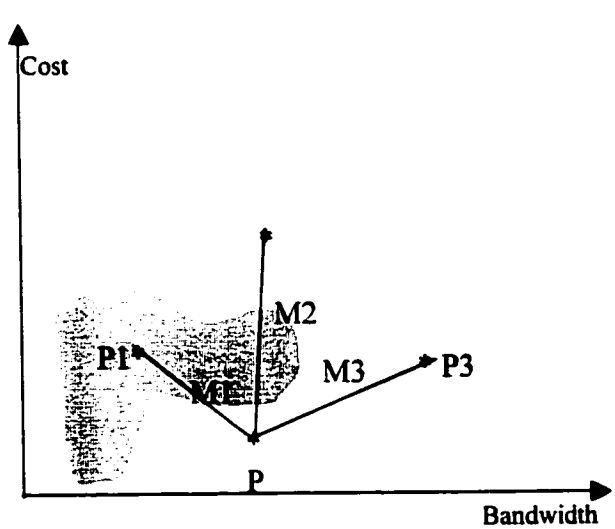


Figure 4.8: Processing the Proposal

Thus, when an agent wants to propose, or receives a proposal, it uses the formula above to make the decision. Returning to the example above, the video agent chooses a point from its strong belief region, i.e. the region with the highest priority, and sends the proposal to the authorization server as shown in Figure 4.7 and 4.8.

Using the above formula, the authorization server looks for the nearest region to the proposal. Specifically, it measures the distances M1, M2, M3 between the point P and the regions R1,R2 and R3 respectively. Since M1 is the smallest distance, i.e. R1 is the nearest region, the authorization server, chooses to guide the video agent to the first region of belief R1.

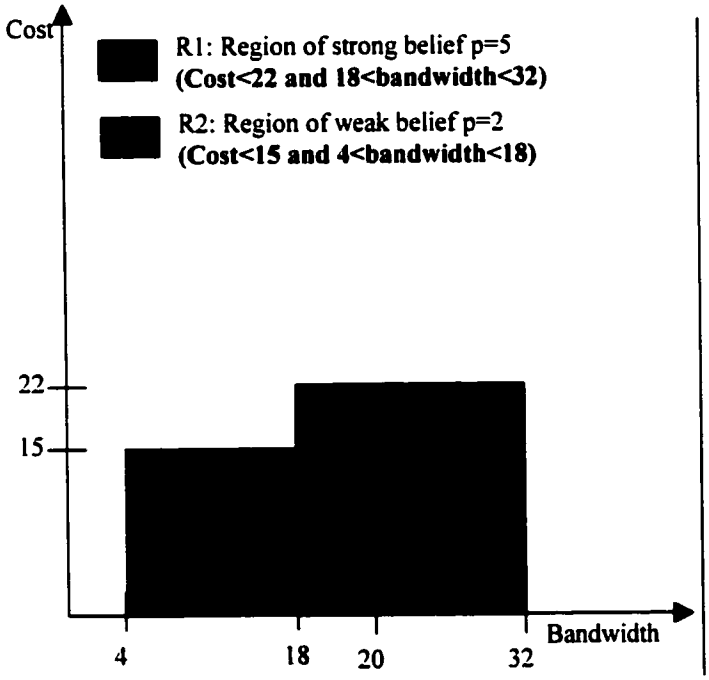


Figure 4.9: Policies of Video Mail agent

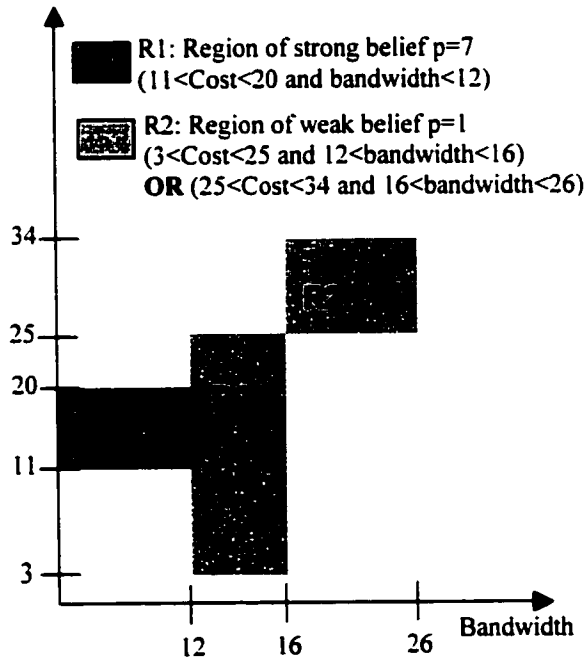


Figure 4.10: Policies of Authorization Server

Let's take a real example to show how the negotiation works. In the video agent example, the functions that represent the belief of the agents are linear, such that belief can be represented by the constraints of the policies, for example, **(cost<22 and 18<bandwidth<32 with priority p=5)** ***OR*** **(Cost<15 and 4<bandwidth<18 with priority p=2)** as belief of the video agent, and **((11<Cost<20 and bandwidth<12 with priority p=7)** ***OR*** **((3<Cost<25 and 12<bandwidth<16)** ***OR*** **(25<Cost<34 and 16<bandwidth<26) with priority=1))** as belief of the authorization server, as shown in Figures 4.9 and 4.10.

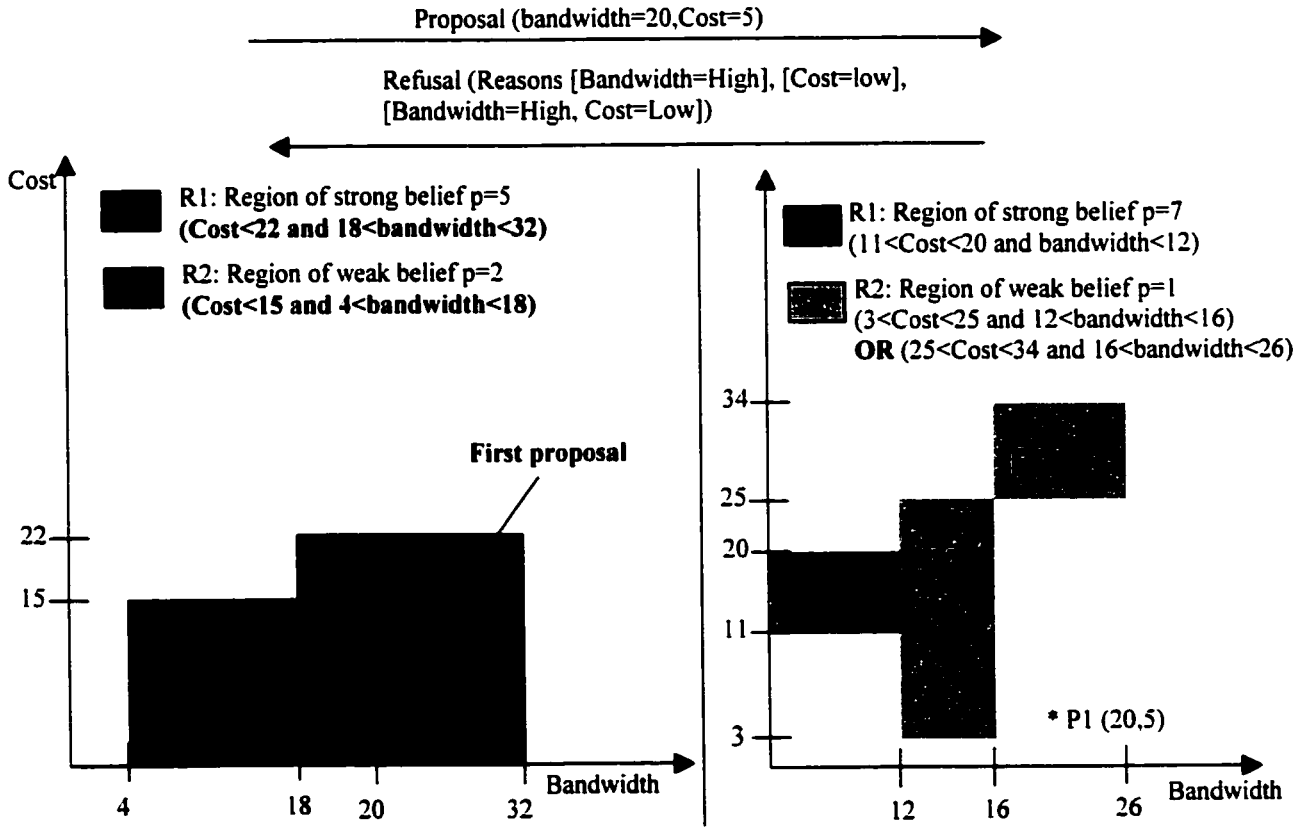


Figure 4.11: First Proposal

The belief, represented above, is used by the two agents to propose or to evaluate a proposal. The following specific formula is used in PMMS when an agent has to evaluate a proposal and deliver a counter proposal.

$$\mathbf{M(P,R)=min}\ (\sqrt{(x_p - x)^2 + (y_p - y)^2} + \frac{2}{p^3}); \quad \forall X(x,y) \in R$$

The Video Agent starts by sending a proposal to the authorization server, as discussed in greater detail below. The authorization server refuses the proposal since it violates all of its policies, however, it tries to guide the requesting agent to the nearest region (i.e. nearest in the sense of the metric M), which is R2 in the present example (Figure 4.11).

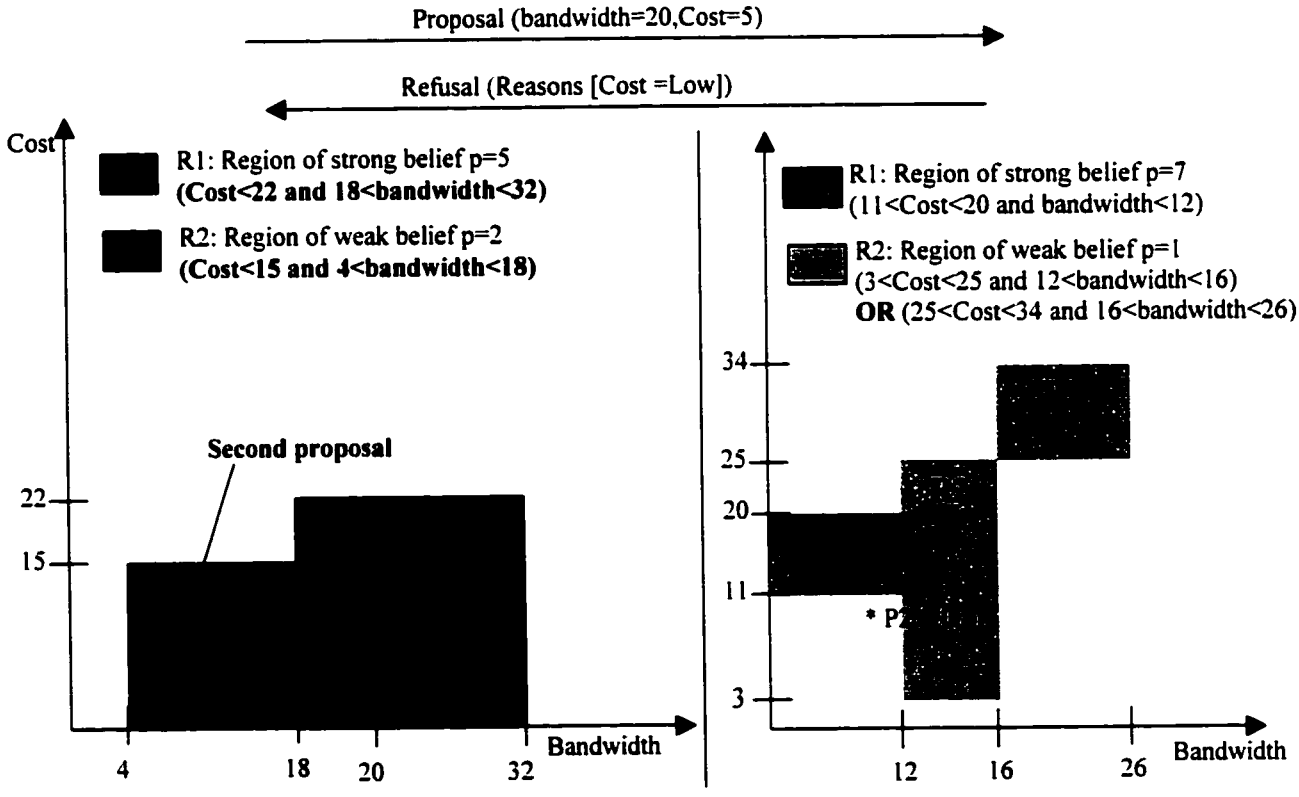


Figure 4.12: Second Proposal

Using the hints provided by the authorization server, the video agent tries to find the best proposal (according to the metric M) that is contained in its regions of belief and satisfies the conditions sent by the authorization server. The second proposal is then sent to the authorization server (Figure 4.12). Like the first proposal, the authorization server looks for the nearest region to the proposal, i.e. R1, (the use of the priority term in the formula of measurement makes the region R1 the nearest region to the proposal P2), and tries to guide it to that region. It then sends the response to the video agent.

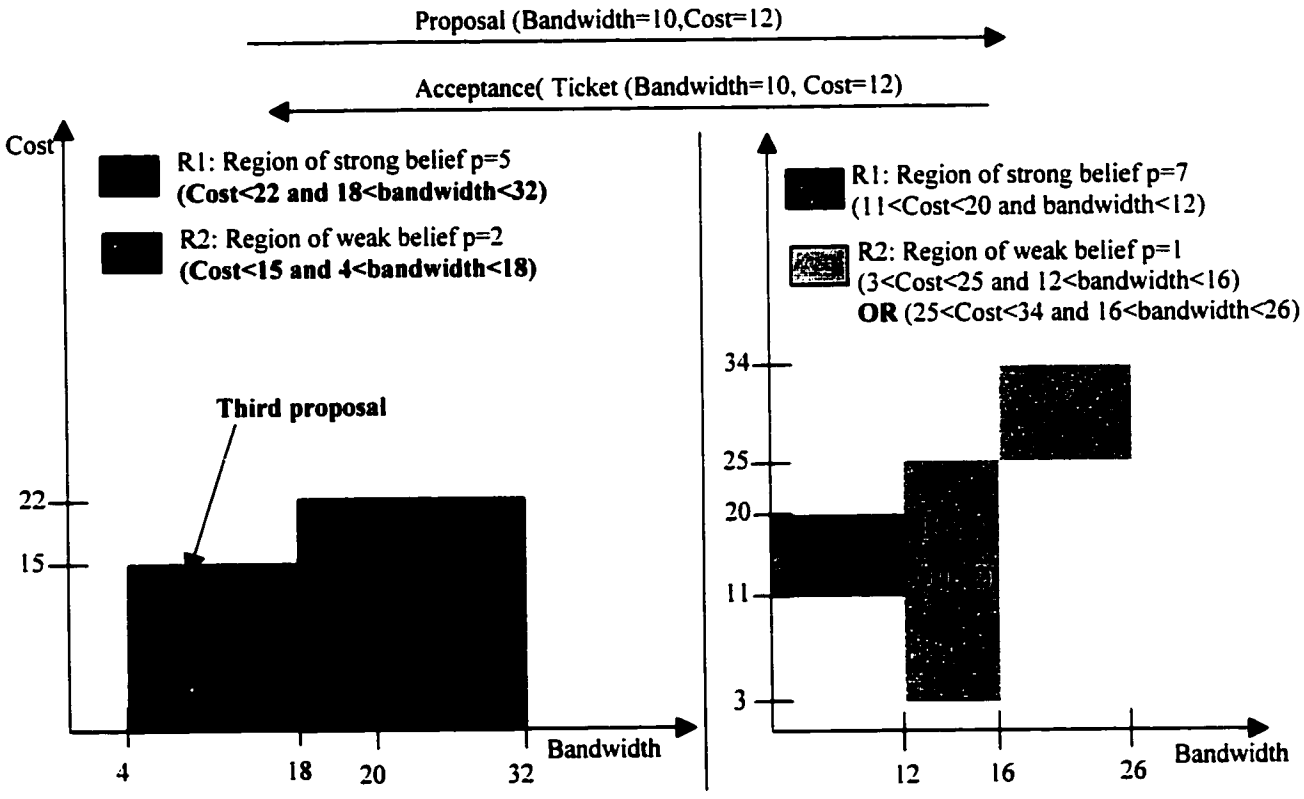


Figure 4.13: Agreement is reached

Using the last hint, the video agent looks for a third proposal (Figure 4.13). This time it sends a proposal that gets the agreement of the authorization server, since it does not violate any of its policies.

4.4 Chapter Summary

This chapter presented an example of a site management in the PMMS system. It illustrates in detail how policies are represented and classified. The advantage of this representation is that policies can be easily read and understood by system administrators. Also, we explained how policies are evaluated upon request of an authorization. The last section presented a negotiation mechanism to resolve conflicts, which are inevitable in a system fully managed by autonomous agents. Finally, we note that some of the ideas presented in chapter 3 and 4 are patent pending.

Chapter 5

Implementation

The implementation of this work consists of a multi-agent system comprised of a graphical user interface, Policy Server, Authorization Server, and an Policy Layer. All Agents in the system collaborate towards the achievement of a joint objective that is driven from the management policies. Each agent has a policy layer to analyze and take decision based on the information it receives from the policy server, event server, or other system components. Hence, the autonomy of agents is guaranteed. Cooperation between agents is strengthened by a limited communication protocol built on top of MicMac.

In what follows, we describe the representation of policies using the eXtensible Markup Language (XML) language, the communication layer built on top of MicMac, the agent development model, and the mapping of user profiles with system policies to build a session profile for a roaming user.

5.1 Technology Integration

This section briefly describes the tools used for the implementation of this system. As illustrated in figure 5.1, Java is used for the implementation of the agents, MicMac ensures the collaboration of distributed agents, KQML provides the communication layer, and XML is used for the representation of the messages content.

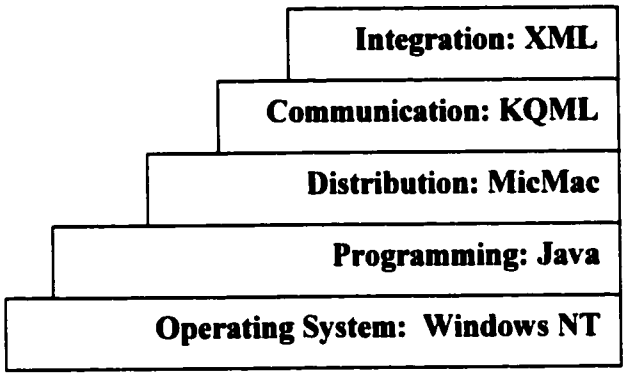


Figure 5.1: Implementation Tools

5.1.1 Java Language

The Java language has become the most used language for writing agents systems and applications. It offers the possibility to write platform independent software running on top of a Java Virtual Machine (JVM). The same software can be executed on different systems like Unix, Windows, or MAC. In addition to its platform-independent approach, Java has several advantages with respect to the development of software agents:

- Object-Oriented programming language with a syntax like C++, but cleaner code than C++.
- Classes can be dynamically loaded during execution time.
- Multi-threading is built-in to the language, with support for basic resource locking and synchronization.
- Garbage collection is handled by the Java Virtual Machine. There is no need for destructor methods to release the memory occupied by an object.

5.1.2 MicMac

MicMac is a set of software tools, which provide a common medium for the communication of multiple, distributed agents. The mechanism of communication is based on the Blackboard paradigm. A tuple space, called agora in MicMac terminology, is considered as an open area for exchanging information between agents. The communication mechanism is very simple: any agent that wants to communicate with another agent will set the receiver of the message to the latter agent and posts it in this common medium. The receiver will then pick it from the shared blackboard. Therefore, all messages are sent to the communication medium, which dispatches them to the agent listening for it. Private agoras can also be created for private agents' communication. MicMac implements a shared memory, which stores information as tuples that are unordered set of fields represented as "key-value" associations.

5.1.3 KQML

The Knowledge Query and Manipulation Language (KQML) is a language and protocol for exchanging information and knowledge [11]. It is part of the ARPA Knowledge Sharing Effort [42], which is developing methods and techniques for building large-scale knowledge bases, which can be shared and reused. KQML is both a message format and a message-handling protocol to support run-time knowledge sharing among agents. It can be used as a language for an application program to interact with an intelligent system or for two or more intelligent systems to share knowledge, in support of cooperative problem solving. It focuses on an extensible set of performatives, which defines the permissible operations that agents may attempt on each other's knowledge and

goal stores. The performatives comprise a substrate on which to develop higher-level models of inter-agent interaction such as contract nets and negotiation. In addition, KQML provides a basic architecture for knowledge sharing through a special class of agent called communication facilitators, which coordinate the interactions of other agents.

5.1.4 XML

The extensible Markup Language (XML) is an emerging standard for the representation of data for sharing and retrieval of information and efficient publishing to multiple users [46]. XML has the following advantages:

XML is flexible, making it ideal for describing any block of content at any scale, from a user bookmark, to an entire database.

- It supports a wide variety of applications for authoring, browsing and so on.
- It's simple to write programs that processes XML documents, especially with the availability of different parsers provided by companies like SUN. and IBM.
- XML documents are "human" and clear, so one can view XML source code in a text-editor and understands its meaning.

5.2 XML Policy Representation

Policies are downloaded from the Policy Server to system agents running on different platforms, therefore, they should be represented in a platform independent format that can be understood by software agents developed by different programming languages. We use XML for policy representation because it becomes a standard for the exchange of structured documents between different systems.

5.2.1 Policy Document Type Definition of a policy

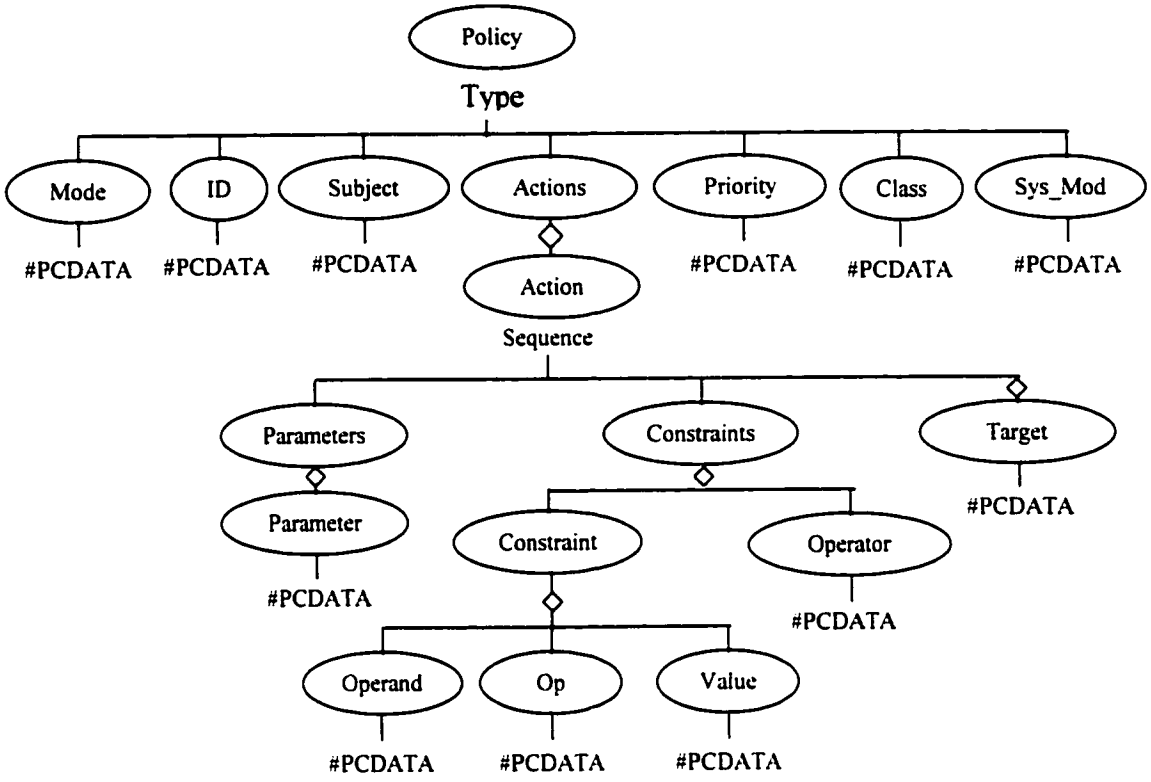


Figure 5.2: Policy Document Type Definition

As shown in Figure 5.2, each policy item has one or multiple actions associated with it. Depending on the mode of the policy, the agent is permitted or prohibited to execute these actions. Each action item has an agent executor of the action specified in the target item. Constraints of the action determine the conditions to be satisfied before the policy can be applied. Each constraint is composed of an operand, operator, and a value. An operand may be expressed using system values, like the bandwidth available at the time of the processing of the constraint; or using one of the parameters of the action. A policy may have many actions associated with it. These actions may have an order relationship expressed by the sequence item. They can be ordered, or unordered: ordered

actions have to be executed in the order specified in the policy rule; unordered actions can be executed in any order.

5.2.2 Example of an XML policy representation

Below we provide an example of a resource reservation authorization represented as an XML document.

```
<Policy type="Authorization">
<Mode> Permission </Mode>
<ID> A010 </ID>
<Subject> Video Agent </Subject>
<Actions Sequence="Ordered">
<Action>Reserve
<Parameters>
        <Parameter> CPU </Parameter>
</Parameters>
<Constraints>
<Constraint>
        <Operand type="System"> CPU <Operand>
        <Op> Greater </Operator>
        <Value> 200 </Value>
</Constraint>
<Operator>AND</Operator>
<Constraint>
        <Operand type="Parameter"> CPU <Operand>
```

<Op> Less </Operator>

<Value> 10 </Value>

</Constraint>

</Constraints>

<Target> Resource Agent </Target>

</Action>

</Actions>

<Priority> 5 </Priority>

<System_Mode> Normal Resource Mode </System_Mode>

<Class> Resource Reservation</Class>

<Creator> Administrator </Creator>

An optional element, which is not shown in the above representation, is the trigger element. The trigger is required when the type of the policy is an obligation. When triggered, the duties of the obligation have to be performed.

5.3 Graphical User Interface

The Graphical User Interface is comprised of three components: a policy editor, constraint editor, and policy validation and transformation component.

5.3.1 Policy Editor

The Policy Editor allows an administrator to perform the basic operations on policies, that is, viewing, editing, deleting, enabling and disabling policies. The policy interface has the following menus:

- View: Lists existing policies without the possibility to modify them. The viewing of policies can be restricted to a site of the virtual network, to a class of policies, or to a specific agent.
- Edit: Adds a policy to the system, or modifies an existing policy.
- Delete: Deletes an existing policy from the system.
- Enable: Puts a policy in a state where it can be processed by the authorization server, or agents when its constraints are satisfied.
- Disable: Puts a policy in a state where it is not processed even though its constraints are satisfied.
- Trigger System Mode: Puts the system in a specific state. System modes can be related, and therefore, only one system mode of related modes can be enabled at a given time.

5.3.2 Constraint Editor

Because of the complexity of the representation of the policy constraints, it was necessary to develop a constraint editor in order to ease the creation and the viewing of the constraint. The editor displays of the policy constraints as a tree. Each node of the tree represents an operator of the constraint, while leafs represents the policy operands. An administrator can also check the validity of the policy constraint by checking its syntax.

5.3.3 Policy Validation & Transformation

Once a new policy is entered, the Policy Validation and Transformation component will validate the syntax of the policy. It uses the API of the constraint viewer to determine whether the constraint is valid or not. If the policy is valid, it proceeds to the

representation of the policy using the XML syntax and then distributes it to the policy server. If the policy is not valid the editor displays a message box notifying the user of the failure of the requested operation.

5.4 Agent Development Model

Agents are implemented with Java language for the reasons given before. In our model, an agent is an aggregation of three layers: Policy, Communication, and Actions as illustrated in Figure 5.3.

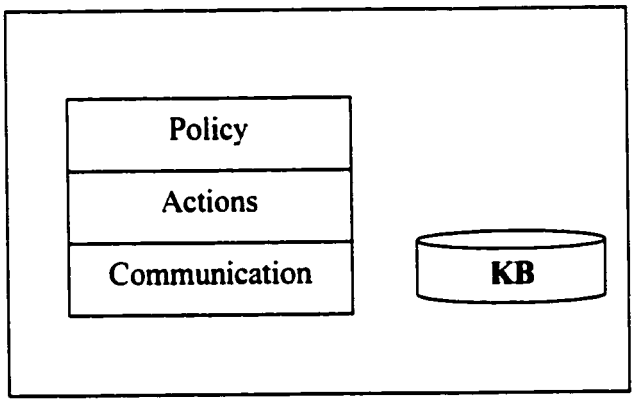


Figure 5.3: Agent Development Model

5.4.1 Policy Layer

A Policy layer is providing all the functions for receiving policies, requesting a policy, subscribing to system events, and notification policy server of its roles.

All System agents in PMMS implements the Agent Interface, which is comprised of the following methods:

- **Request**: Used by agents to request an authorization to perform an action. The method encapsulates the request in a KQML message where the content of the

message is represented using XML language, then it's sent to the authorization server via MicMac.

- **Response**: waits for the authorization to process the request and then retrieves the result from the communication channel. If the request was approved, then the response will include a ticket that gives the agent the right to execute the action or ask another agents to perform the action.
- **Ask:** Used by the agent to request another agent to perform an action. The method will include the ticket issued by the authorization server in the content of the message sent to the executor.
- **Perform:** Performs an action requested by an agent. The method checks first the validity of the ticket sent by the requester agent, such as the non expiration of the ticket, and performs the action. A confirmation of execution is sent back to the requester informing him of the success, or eventually the failure of the execution of the action.
- **Get-Policies:** Listens to the communication channel and gets any policy sent by the policy server to the agent. The trigger of the policy are then subscribed to the event server using the subscribe event method.
- **Subscribe Event:** Subscribe an event to the event server. If the event occurs, the agent will be automatically informed.
- **Unsubscribe Event:** Unsubscribe an event from the event server in case the agent is no more interested in the occurrence of this event. This can be the case of triggers of deleted or disabled policies.

- **Notify_Event**: Informs the event server of the occurrence of an event that is produced by the agent. For example, upon reception of an email, the email server will notify the event server by sending email_received event.
- **Notify_Role**: allows an agent to send to the policy server the set of roles it plays in the system.

5.4.2 Communication Layer

This layer is comprised of the MicMac client Application Programming Interface (API), which implements the following methods:

- Post: Adds a new tuple to the shared memory.
- Pick: Reads and deletes a tuple from the shared memory.
- Peek: Reads a tuple from the shared memory.
- Cancel: Terminates a pending pick/peek request.

5.4.3 Actions Layer

This layer is the code of the agent and implements its behavior. PMMS has a set of agents providing the necessary services to support user mobility.

- Site Logon Agent: authenticates users and opens new sessions.
- Coordinator Agent: manages the inter-agent communications. Its central role is to establish communication among two or more agents and control their activities when exchanging information.
- Site Profile Agent negotiates with its remote peer in the user's home site, the user preferences, and maps them to the local policies of the system. The result of the mapping is a set of services that the site can offer to its visitors.

- Resource agent manages system resources and is requested by other agents for resource reservation.
- Moreover, the system contains a set of agents each one of them representing a service or a device. Each one of them reserves necessary resources for the execution of the service it represents.

5.4.4 Knowledge Base

Upon receipt of a policy, the agent stores the element of policy in a local knowledge base. The set of actions and constrained of policies are the knowledge of the agent. This knowledge is used to perform an action upon reception of a new event from the event server, and to formulate a proposal to the authorization server when asking for an authorization.

5.5. Policy Agents

5.5.1 Policy Server

The policy server is a java standalone program that is responsible of the distribution, retrieval of policies. It maps user profiles to local policies of the system and monitors system modes, so that enabled policies meet with the requirements of the current system state.

- **Set_Policy (Agent_Id, Agent_Role, Policy)**: the policy server uses this operation to distribute any new Policy to the agent identified by the Agent_Id. The role of agent to which the obligation applies is set by the Agent_Role parameter.
- **Set_Authorization (Authorization):** Used only to distribute authorizations to the authorization server.

- **Retrieve_Policy(Agent_Id, Policy_Id)**: Retrieves a policy from an agent. The policy can either be an authorization or an obligation. In the case the policy server needs to retrieve the policy from the system, an Agent Identifier referring to all agents can be specified instead of an identifier for a particular agent.
- **Get_Role(Agent_Id)**: This operation is used by the policy server to learn the roles of a specific agent identified by Agent_Id.
- **Map_Profile:** Negotiate with Profile Agent, a session profile for a mobile user. The session profile is comprised of services the user can access during its work session at the remote site, and the quality of services associated with them.

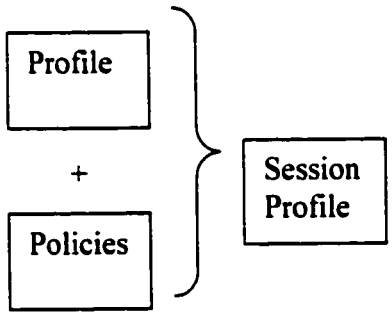


Figure 5.4: Mapping User Profile to Local Policies

- **Enable_Mode:** Puts the system in a new mode where new policies apply. PMMS supports six system modes as illustrated in table 5.1.

System Mode	Description	Category
Normal Resource Mode	Resources are available in the system	1
Critical Resource Mode	Lack of resources	1
Heavy Load Mode	Number of users is greater than a limit	2
Normal Load Mode	Number of users is under a limit	2
High Level Security Mode	System is vulnerable to external attacks	3
Low Level Security Mode	System is well protected.	3

Table 5.1: Description of System Modes

All system modes belonging to the same category are related, and therefore, one system mode of each category can be enabled at each time.

5.5.2 Authorization Server

The Authorization Server implements the engine for processing agent requests. In addition to the communication and agent interfaces, the authorization server implements the following method:

- **Process_Request:** Processes an agent request for an authorization. The server first loads all policies applicable to the request then makes a decision whether the agent shall be allowed to execute the action or not. In case of authorization, a ticket is delivered to the requester. The ticket has attributes to determine the authority that issued it, the time of its expiration, the cost of the operation, and some other useful information to be used for security and accounting purposes.

Chapter 6

Conclusion

6.1 Summary

This thesis has proposed a multi-agent architecture for enforcing policies in a cooperative network according to different classes based on the activities they monitor. The policy server is responsible of the management of all system policies (distribution, enabling, disabling, changing system modes, etc.). As part of its functions, it distributes the obligations to the agents, and the authorizations to a global authorization server that will deliver necessary authorizations to agents. The advantage of this architecture is that it provides a dynamic way to adapt the behavior of agents to any changes in the system. Policies can be quickly setup to control the behavior of agents and make them behave according to the global strategy of the system. Conflicts between agents are detected and resolved through negotiation.

A possible extension to the system is to make each agent contain it's own private authorization server that can receive policies from the global policy server and deliver necessary authorization to the agent. This approach will remove the bottleneck of a centralized common authorization server. By this way, each agent will manage its own policies and takes decision without referring to a central authority. More security mechanisms should be setup to prevent malicious agents from generating false permissions to execute some actions prohibited by the system policies.

6.2 Future Work

As future work of this research, we suggest the definition of a policy language specification to introduce policies in the system. Managers will be able to express policies in a high level language. A policy transformation and refinement agent will be responsible for the transformation of these high level policies into a set of low level policies that will be used to accomplish the goal specified by managers. The policy server will distribute the resulting low-level policies to agents that are specified as subjects of these policies. Another interesting field of research that can be explored as an extension to the current work is the usage of the Java Expert Shell System (JESS) [45] for the knowledge representation and agent reasoning. This approach has the advantage of the usage of an intelligent system widely used in the AI field, for policy storage and processing agents' requests. In this case an algorithm is needed to do the conversion of policies to JESS rules. Last, since the mobile agent technology is an emerging technology that has the potential to replace traditional client server technology, we suggest to apply the current policy management system to monitor the behavior of mobile agents.

References

[1] M. Ahmed and A. Karmouch, "Improving Video Processing Performance using Temporal Reasoning," SPIE-Applications of Digital Image Processing XXII, Denver Co, USA, July 20-23 1999.

[2] A. Karmouch, "Mobile Software Agents for Telecommunications," IEEE Communications, July 1998.

[3] P. Maes, "Intelligent Software", In Proceedings of Intelligent User Interfaces 1997, Orlando, Florida.

[4] C. Hewitt, "Viewing Control Structures as Patterns of Passing Messages," Artificial Intelligence, 8, No. 3, pp 323-364 (1977).

[5] J. Russell, and P. Norvig, "Artificial Intelligence: A Modern Approach," Englewood Cliffs, NJ: Prentice Hall, 1995.

[6] P. Maes, "Artificial Life Meets Entertainment: Life like Autonomous Agents," Communications of the ACM, 38, 11, 108-114, (1995).

[7] M. Wooldridge, and N. Jenning, "Agent Theories, Architectures, and Languages: a Survey," in Wooldridge and Jennings Eds., Intelligent Agents, Berlin: Springer-Verlag, 1-22, 1995.

[8] C. Brustoloni, "Autonomous Agents: Characterization and Requirements," Carnegie Mellon Technical Report CMU-CS-91-204, Pittsburgh: Carnegie Mellon University.

[9] J. Mazur, "Learning and Behavior," Ed. J. Mazur , Pub Prentice Hall 1994.

[10] P. FARJAMI, C. Gorg, and F. Bell "Advanced Service Provisioning Based on Mobile Agents," First International Workshop on Mobile Agents for Telecommunication Applications, MATA 99.

[11] Y. Labrou, J. Mayfield, T. Finin, "KQML as an Agent Communication Language," Proceedings of the third International Conference on Information and Knowledge Management (CIKM'94), ACM Press, Nov. 1994.

[12] M. Genesereth, and R. Fikes, et al., "Knowledge Interchange Format: Version 3.0 Reference Manual," Technical Report, Computer Science Department, Stanford University, 1992.

[13] P. Cohen, and H. Levesquem, "Communicative Actions for Artificial Agents," Proceedings for the First International Conference on Multi-Agents, San Francisco. Cambridge, AAAI Press, pages 65-72, (1995).

[14] D. Sadek, P. Bretier, and F. Panaget, "ARCOL agent communication language and MCP, CAP and SAP agent's cooperativeness protocols," Response to FIPA Call For Proposals, http://www.cselt.it/fipa/torino/cfp1/propos97_015.htm.

[15] G. O'Hare, and N. Jenings, "Foundations of Distributed Artificial Intelgence", John Wiley & Sons, 1996.

[16] R. Smith, "The Contract Net Protocol: High-Level Communication and Control in a Distributed Problem Solver," IEEE Trans. on Computers, Vol.29, No.12, pp.1104-1113, 1980.

[17] J. E. White, Telescript Technology, The Foundation of the Electronic Marketplace, General Magic White Paper, 1994.

[18] D. Marriot, and M. Sloman, "Management Policy Service for Distributed Systems," IEEE 3rd Int. Workshop on Services in Distributed and Network Environments, Macau, June 1996.

[19] G. Hughes, and M. Creswell, "An introduction to Modal Logic," Metheun, London,1968.

[20] D. Jonsher, "Extending access control with duties: Realized by active mechanisms," Proceedings of the IFIP WG 11.3 Sixth Working Conference on Database Security, IFIP WG 11.3.

[21] K. Becker et al., "Domain and Policy Service Specification," number S-SI-07-I-2-R. IDSM Deliverable D6, SysMan Deliverable MA2V2.

[22] T. Kotch et al., "Policy Definition Language for automated management of distributed systems," Proceedings of the Second International Workshop on Systems Management, IEEE Computer Society Press.

[23] R. Wies, "Policies in Integrated Network and Systems Management: Methodologies for the Definition, Transformation, and Application of Management Policies", PhD thesis, Department of Computer Science, University of Munich, 1995.

[24] P. McBrien, et al., "A rule language to capture and model business policy specifications," Proceedings of the Advanced Information Systems Engineering, Third International Conference CaiSE'91, number 498 in LNCS, Springer-Verlag, pp. 307-318.

[25] M. Massulo, et al., "Policy Management: An architecture and approach," Proceedings of the IEEE Workshop on Systems Management.

[26] K. Marzullo, et al., "Tools for distributed application management," Technical Report TR 90-1136, Cordinell University, Ithaca, New York, 1994.

[27] J. Case, M. Fedor, M. Schoffstall, and J. Davin, "Simple Network Management Protocol", RFC 1157, SNMP Research, Performance Systems International, Performance Systems International, MIT Laboratory for Computer Science, May 1990.

[28] Policy Management Working Group, http://www.ietf.org/html.charters/policy-charter.html.

[29] J. Strassner, E. Ellesson, B. Moore, "Policy Framework LDAP Schema," June 1999, <draft-ietf-policy-core-schema-04.txt>, Work in Progress.

[30] Recommendation X.500 ITU-T, Information Technology -Open Systems Interconnection- The Directory: Overview of Concepts, Models, and Services, 1993.

[31] M. Fine, et al., "Quality of Service in Policy Information Base," June 1999, <draft-mfine-cops-pib-01.txt>, Work in Progress.

[32] J. Byle, R. Cohen, and D. Durham, "The COPS (Common Open Policy Service) Protocol," August 1999, <draft-ietf-rap-cops-07.txt>, Work in Progress.

[33] H. Harroud, A. Karmouch, T. Gray, S. Mankovski, "An Agent-Based Architecture For Inter-Sites Personal Mobility Management System," Mobile Agents for Telecommunication Applications Workshop, MATA'99, 6-8 October.

[34] Mitel Corporation, Mitel-CATA (The MicMac Software Testbed), Technical Report, CITI, Adaptive Information Systems 1996.

[35] A. Hooda, "Data Management for Supporting Nomadic Users Based on LDAP and Software Agents," Master's Thesis, April 1999 University of Ottawa, Ottawa, Canada.

[36] D. Marriot, and M. Sloman, "Management Policy Service for Distributed Systems," IEEE 3rd Int. Workshop on Services in Distributed and Network Environments, Macau, June 1996.

[37] M. Genesereth, and S. Ketchpel, "Software Agents", Communications of the ACM, 37, No. 7, pages 48-53, 1994.

[38] R. Wies, "Policy Definition and Classification: Aspects, Criteria and Examples," Proceeding of the IFIP, IEEE International Workshop on Distributed Systems: Operations & Management, Toulouse, France, 10-12 October 1994.

[39] J. Hartmann et al. "Agent Technology for the UMTS VHE concept," Ericsson Eurolab, Germany 1998.

[40] R. Wies, M. Mountzia, and P. Steenekamp, "A Practical Approach Towards A Distributed and Flexible Realization Of Policies Using Intelligent Agents," Proceedings of the 8th IFIP/IEEE International Workshop on Distributed Systems: Operations & Management, October 1997.

[41] J. Bradshaw (ed.), "Software Agents" American Association for Artificial Intelligence, MIT Press 1997.

[42] R. S. Patil, and al, "The DARPA Knowledge Sharing Effort: Progress Report", Proc. of the Third International Conference on Principles of Knowledge Representation and Reasoning, Cambridge, MA, 1992.

[43] Bond, and Gasser, "Readership in Distributed Artificial Intelligence", Morgan Koffmann, San Mateo, CA, 1988.

[44] J. Rosenchein, and G. Zlotkin, "Rules of Encounter Designing Contentions for Automated Negotiation among Computers", MIT Press 1994.

[45] E. Friedman-Hill, (1997). *Jess, the Java Expert System Shell.* Sandia National Laboratories, Livermore, CA, 1997. Available at: http://herzberg.ca.sandia.gov/.

[46] Extensible Markup Language (XML) 1.0, W3C Recommendation REC-xml-19980210, World Wide Web Consortium, 1998. Available at http://www.w3.org/TR/1998/REC-xml-19980210.html.

[47] C. Zheng, "Design and Implementation of secure communication for a distributed mobile computing system," Master's Thesis, December 1999 University of Ottawa. Ottawa, Canada.

MQ

66047

UMI

MICROFILMED 2002

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

Université d'Ottawa • University of Ottawa

PERMISSION DE REPRODUIRE ET DE DISTRIBUER LA THÈSE

PERMISSION TO REPRODUCE AND DISTRIBUTE THE THESIS

NOM DE L'AUTEUR / NAME OF AUTHOR:	HARVEY, Pascal
ADRESSE POSTALE / MAILING ADDRESS:	2-486 Lisgar Ottawa, Ontario K1R 5H3
GRADE / DEGREE:	**ANNÉE D'OBTENTION / YEAR GRANTED**
M.A..(sociologie)	2001
TITRE DE LA THÈSE / TITLE OF THESIS:	
LA RATIONALISATION DE LA SUBJECTIVITÉ MODERNE	

L'auteur permet, par la présente, la consultation et le prêt de cette thèse en conformité avec les règlements établis par le bibliothécaire en chef de l'Université d'Ottawa. L'auteur autorise aussi l'Université d'Ottawa, ses successeurs et cessionnaires, à reproduire cet exemplaire par photograhie ou photocopie pour fins de prêt ou de vente au prix coûtant aux bibliothèques ou aux chercheurs qui en feront la demande.

Les droits de publication par tout autre moyen et pour vente au public demeureront la propriété de l'auteur de la thèse sous réserve des règlements de l'Université d'Ottawa en matière de publication de thèses.

The author hereby permits the consultation and the lending of this thesis pursuant to the regulations established by the Chief Librarian of the University of Ottawa. The author also authorizes the University of Ottawa, its successors and assignees, to make reproductions of this copy by photographic means or by photocopying and to lend or sell such reproductions at cost to libraries and to scholars requesting them.

The right to publish the thesis by other means and to sell it to the public is reserved to the author, subject to the regulations of the University of Ottawa governing the publication of theses.

N.B. LE MASCULIN COMPREND EGALEMENT LE FEMININ

28 février 2001

DATE

(AUTEUR) SIGNATURE (AUTHOR)



Université d'Ottawa · University of Ottawa

Université d'Ottawa • University of Ottawa

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

HARVEY, Pascal

AUTEUR DE LA THÈSE - AUTHOR OF THESIS

M.A.(sociologie)

GRADE - DEGREE

Sociologie

FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

LA RATIONALISATION DE LA SUBJECTIVITÉ MODERNE

Roberto Miguelez

DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

V. Armony

K. Mellos

J.-M. De Koninck, Ph.D.

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

SIGNATURE

DEAN OF THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

La rationalisation de la subjectivité moderne

par

Pascal Harvey

Thèse de maîtrise rédigée sous la direction du
professeur Roberto Miguelez

Soumise au Département de Sociologie
en vue de combler exigences du cours SOC 7999
et pour l'obtention du grade
M. A. Sociologie

Université d'Ottawa
Ottawa, Ontario
janvier 2001

©Copyright
2001, Pascal Harvey

National Library of Canada
Bibliothèque nationale du Canada

Acquisitions and Bibliographic Services
Acquisitions et services bibliographiques

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-66047-8

Canada

Résumé

Cette analyse propose d'envisager deux temps de l'organisation de la socialité à la lumière des enjeux qui entourent la subjectivité moderne. Il est avancé que la subjectivité moderne a parmi ses effets sociologiques importants l'individualisme. L'individualisme, alors qu'il recouvrait une dimension morale dans l'optique d'une conception supra-individuelle du Tout social au XVIIIe et XIXe siècles, ne s'inscrit plus, dans le monde contemporain, que dans les limites d'un lien social procédural et formel. Cette forme nouvelle du lien social, dont la domination technocratique de la société est partie prenante, s'inscrit dans les processus de rationalisation du monde moderne. Or, il est avancé que la subjectivité moderne prend part elle aussi à un processus de rationalisation. Aussi, en appréhendant l'organisation technocratique de la socialité contemporaine à partir du nouvel individualisme que la rationalisation de la subjectivité fait naître, il est possible de constater un changement des visées subjectives qui conduisent à des formes nouvelles d'autonomie.

Remerciements

L'accomplissement de ce qui suit est le résultat de trop de choses, de trop de coïncidences et de contingences, pour les synthétiser en quelques lignes. Tout en reconnaissant cela, je m'en voudrait toutefois de ne pas mentionner quelques personnes. D'abord, je remercie Koula Mellos et Victor Armony pour leur travail d'évaluation rigoureux. Roberto Miguelez, en m'ayant fait l'honneur d'accepter d'être mon directeur, s'est acquitté de cette tâche sous la forme d'un interlocuteur des plus éclairant, comme toujours. C'est à lui que je dois mon amour de la sociologie. Ensuite mes parents, Jacques Harvey et Éliane Martel, et ma soeur, Marie-Christine Harvey, qui, malgré le fait qu'ils se promènent entre Kingston et Stuttgart, n'ont jamais cessé d'être près de moi et de m'apporter le même support que j'aurais eu si nous avions été voisins. Je m'ennui d'eux. Des amis précieux comme Roch, Christian, Jean-Sébastien, Guillaume et Karl n'ont pas leur pareils pour bien aroser une existence. Valérie de Courville Nicol, quant à elle, ne s'est pas contenté de m'aimer et de me laisser l'aimer, elle m'a aussi permis de devenir un meilleur sociologue en m'ouvrant à des univers théoriques nouveaux. Merci beaucoup. Enfin, un mot pour mon texte qui a su lui-même me révéler les moyens pour que je puisse l'achever quand j'étais désarmé. C'est peut-être à Éleucippe Harvey que je dois ces révélations.

Table des matières

Introduction 5

Chapitre 1 : Premiers abords : appropriations conceptuelles et méthodologiques 13

1. 1. Le sujet et la subjectivité 13
1. 2. L'individu et l'individualisme : où, quand et comment ? 16
1. 3. Le politique à partir du sujet 23

Chapitre 2 : Enquêter la modernité vers un portrait du sujet en tant qu'individu 28

2. 1. Des grandes dimensions morales de la subjectivité moderne 30
2. 2. Le sujet moderne comme sujet du sens 33
2. 3. La question de soi et d'autrui dans la subjectivité moderne 39
2. 4. La convergence de l'éthique et de la morale : l'incarnation subjective 42

Chapitre 3 : L'individualisme moral des XVIIIe et XIXe siècles 55

3. 1. La liberté individuelle et l'humanisme comme fondements moraux 57
3. 2. L'individualisme entre la modernité et la tradition : l'approche scientifique 62
3. 3. La société et le progrès comme condition morale de l'individualisme 67
3. 4. Pourquoi un individualisme « moral » ? 73

Chapitre 4 : La rationalisation moderne : foyer du nouvel individualisme 77

4. 1. L'émergence d'une nouvelle socialité : le lieu du nouvel individualisme 78
4. 2. Les contours d'une nouvelle subjectivité idéale typique 81
4. 3. La rationalisation des conduites 86
4. 4. La rationalisation des techniques 93

Chapitre 5 : La dissolution de la solidarité et l'autonomie subjective 103

5. 1. La rationalisation de la société contemporaine en technique 106
5. 2. La domination technocratique et l'autonomie subjective 112

Conclusion 127

Bibliographie et références 134

Introduction

L'individualisme est un terme qui révèle une certaine ambiguïté conceptuelle. Confondu avec l'égoïsme parfois, avec l'autonomie ailleurs, ou encore envisagé comme état d'aliénation et comme liberté absolue il laisse perplexe, tant sur le terrain de son appréhension politique, qu'éthico-morale ou médiatique. Ces polémiques l'accompagnent depuis sa naissance comme concept, au début du XIXe siècle. Son apparition conceptuelle en des termes péjoratifs est l'œuvre de ses détracteurs mais le mot individualisme sera aussi adopté par ses supporteurs pour nommer le même phénomène. S'il en est ainsi, c'est qu'au-delà de sa portée doctrinaire, il suppose un caractère sociologique indéniable, un fait social.

Qu'il ait le visage de la domination ou de l'émancipation aux yeux de certains ne fait pas moins de lui la synthétisation conceptuelle d'un phénomène complexe qui, depuis l'émergence des idées humanistes, s'impose comme condition objective de la subjectivité. Autrement dit, l'individualisme est un terme privilégié pour la compréhension de l'acteur social dans sa spécificité moderne. L'acteur social moderne est le lieu de l'investissement de cadres subjectifs qui font signifier son existence à partir de son caractère individuel et distinct. Cela ne signifie en rien toutefois que l'individualisme ait eu le même visage depuis son émergence comme phénomène sociologique. Cela révèle plutôt que les points de repères de la légitimation du vivre-ensemble, de l'éthique ou de la morale prennent des assises dans l'individu, avec la naissance de ce que nous appelons le sujet humaniste.

L'individualisme correspond donc à une subjectivité particulière; un

façonnement particulier du contenu indéfini qu'est l'acteur social. Or, son existence se déploie sur un terrain particulier. Si l'on pouvait en quelques mots synthétiser les changements socio-historiques depuis la modernité, le thème de l'individualisme, aux côtés du capitalisme et de la rationalisation, seraient certainement des candidats en lice comme indicateurs principaux. Conjecture farfelue, certes, car ces trois thèmes ne peuvent certainement pas résumer le cheminement du monde moderne jusqu'au monde contemporain et rien, en bout de ligne, n'est en mesure de le faire. Or il est possible de proposer des scénarios vraisemblables qui puissent être en mesure de fournir des éclairages sur des parcelles de compréhension des transformations de la société et de certains de ses phénomènes. C'est donc principalement à travers ces trois thèmes que cette analyse sillonera. Les développements de l'individualisme, depuis le monde moderne, sont de très près liés aux impératifs de la rationalisation moderne et au lieu de sa réalisation : le mode de production capitaliste.

Notre objectif final est une exploration de la société contemporaine à partir du sujet, c'est-à-dire une tentative de sonder, à partir d'une typification idéale de l'appréhension du monde de l'acteur social, ce qui se présente comme la société si ses phénomènes constitutifs sont considérés dans leur ensemble. En soi, cela présente un défi méthodologique dans la mesure où nous ne cherchons surtout pas à nous cantonner dans une analyse qui exclue de ses paramètres les effets objectivants de la société — les institutions, la domination ou même la subjectivité. Plus précisément, c'est la question du politique qui nous intéresse dans cette analyse. À savoir les bases de la légitimation de l'organisation de la société et de la domination — au sens wébérien de la

domination charismatique, traditionnelle ou rationnelle-légale. La question du pouvoir reste central dans l'appréhension du politique mais elle ne fera pas l'objet de nos développements ici. C'est à travers une lecture de l'individualisme que nous tenterons de déterminer l'existence de deux temps différents, couplés à deux modalités différentes, de l'ambiance dans lequel le sujet humaniste se constitue et se réalise.

L'individualisme, en tant que produit de la subjectivité humaniste, est un phénomène qui confère à l'acteur social une vision du monde qui s'articule autour de l'idée de l'autonomie. Cela consitue une spécificité de l'acteur social moderne. Non pas que les acteurs sociaux pré-modernes sont par définition aliénés. Il s'agit plutôt de dire que leur existence, de façon générale, n'est justement pas envisagée subjectivement en terme d'aliénation, d'hétéronomie ou d'autonomie. Ces concepts ont une pertinence beaucoup plus grande pour l'analyse du monde moderne qu'ailleurs car on ne peut traiter de l'autonomie que dans la mesure où elle existe dans un espace de compréhension subjective. Or leur mise en place et leur appréhension par les individus dans le cadre des grandes tendances socio-historiques des sociétés modernes n'est pas la même partout. Elle diffère particulièrement si l'on prend la société contemporaine, dont nous fixons l'existence entre le milieu du vingtième siècle et aujourd'hui, pour la comparer à la société moderne des XVIIIe et XIXe siècle qui est en plein dans l'effervescence et les bouleversements des révolutions bourgeoises et démocratiques.

Notre thèse principale est que l'individualisme se manifeste de façon différente durant ces deux moments de l'histoire. Si nous regroupons

ensemble les XVIIIe et XIX siècles, c'est parce qu'ils dévoilent un individualisme moral. Nous tenterons de démontrer qu'à cette époque l'autonomie est comprise dans une conception morale de l'individu-citoyen qui fait appel à des principes supra-individuels, dont le progrès. Dans cette perspective l'individualisme à une existence légitimée par la morale. Par ailleurs nous avancerons l'idée selon laquelle l'individualisme contemporain ne réside plus dans une conception de la solidarité à l'intérieur de principes moraux universellement partagés. Nous postulerons qu'il réside plutôt dans une désubstancialisation des valeurs humanistes qui ont contribué à le créer. C'est-à-dire que le visage des valeurs constitutives de l'individualisme n'est dès lors plus que celui de la procédure, de la consistance, de la stabilité et de la performance. Nous allons proposer une lecture de ce changement qui s'inscrit dans les développements de la rationalisation du monde moderne. Ainsi, nous croyons que la conjonction des effets de la rationalisation des conduites et des techniques, sur le terrain du capitalisme, ont conduit d'une part à la dissipation de la morale pour laisser place à la gestion des moralités et d'autre part, à des formes nouvelles d'autonomie que l'on peut découvrir sous le visage *a priori* désenchantant et aliénant de la domination technocratique contemporaine. Le politique, sous le jour de la rationalisation de la subjectivité moderne, prend une allure nouvelle.

Le premier chapitre de l'analyse fera l'objet de clarifications conceptuelles et méthodologiques. Nous nous attarderons d'abord à donner les points de repères sommaires de ce que nous entendons par le sujet et la subjectivité car ces deux thèmes sont au centre de notre analyse de l'individualisme en tant qu'il émane d'une subjectivité particulière et qu'il façonne un sujet particulier.

La définition de ce que nous entendons par l'individu et l'individualisme fera l'objet de la partie suivante. Nous chercherons à situer ces thèmes dans le contexte du rapport inusité qu'ils instituent entre le sujet et le monde qui les entoure. La partie finale du chapitre constitue une clarification méthodologique et épistémologique de notre appréhension du politique à partir du sujet. Celle-ci est importante à faire dans la mesure où nous voulons garder nos réserves quant aux prétentions de cette analyse à pouvoir saisir le politique à partir de la subjectivité. Nous verrons qu'une telle entreprise n'est possible que dans le cadre de types-idéaux qui envisagent une lecture parmi d'autres de la société.

Le second chapitre sera quant à lui entièrement consacré à délimiter les contours de la subjectivité moderne. Il constituera aussi par ailleurs une discussion plus approfondie sur une phénoménologie de la subjectivité, particulièrement de la subjectivité moderne. À la suite d'une discussion d'une partie des travaux de Charles Taylor dans *Les sources du moi* (1998), nous tenterons d'expliciter les implications des valeurs humanistes dans la constitution du sujet moderne. Nous nous attarderons en premier lieu aux fondements moraux de cette subjectivité. Ce sera le rapport du sujet moderne à l'égard de lui même, contenu dans l'impératif du sens et de la dignité, et à l'égard des autres, contenu dans les paramètres de la dignité et du respect d'autrui, qui feront l'objet des parties suivantes du chapitre. Nous clorons le chapitre avec une discussion de la question de la subjectivité dans son rapport à l'éthique et la morale. Cette partie du chapitre consiste aussi en une discussion sur la question de l'aliénation dont les conclusions portent sur l'idée de l'incarnation subjective. Ce chapitre est indispensable dans la mesure où il

permet au reste de l'analyse d'entrer dans des développements nécessitant un certain bagage conceptuel et théorique précis.

Le chapitre suivant sera consacré à brosser le portrait de l'individualisme moral des XVIIIe et XIXe siècles. Ce sera surtout à travers des courants de pensée importants de la modernité que nous ferons ce portrait. C'est que la pensée théorique de cette époque est un indicateur important des angoisses et des préoccupations qui traversent la société civile révolutionnaire et la société industrielle naissante. Les thèses principales des contractualistes et de leur successeur dans la lignée des droits naturels, Emmanuel Kant, feront l'objet de la première partie du chapitre. Nous y délimiterons les grandes lignes d'une pensée qui fait de l'humanisme et de la liberté individuelle les fondements moraux de l'ordre individualiste au XVIIIe siècle. Nous verrons ensuite comment l'approche scientifique vient contribuer au débat de l'ordre moral des nouvelles sociétés démocratiques et capitalistes au XIXe siècle. Nous situons l'émergence de la sociologie dans une réfléxion socio-politique qui s'impose en Europe à cause des tensions idéologiques et des angoisses quant à la nécessité d'instituer un ordre social. Cette dernière partie servira entre autres à introduire le contexte de l'émergence de la pensée d'Émile Durkheim qui fait reposer la légitimité de l'individualisme moral sur les notions de progrès et d'industrialisation. Finalement, nous terminerons le chapitre par une courte réflexion quant à la question de l'individualisme moral qui aura comme objet de nous acheminer vers nos deux derniers chapitres.

Le quatrième chapitre sera destiné à proposer une trame compréhensive de la transition de l'individualisme moral vers un individualisme dont le

degré de rationalisation et de sécularisation sont inédits. C'est à travers le thème de la rationalisation du monde moderne que nous effectuerons cette lecture. De même, il cherchera à tracer les grandes lignes de l'individualisme contemporain : du nouvel individualisme. C'est d'abord en situant le monde contemporain dans son contexte socio-historique que nous entâmerons ce chapitre. Nous proposerons ensuite une esquisse du nouvel individualisme qui aura comme fonction principale de révéler les hypothèses sous-jacentes aux développements subséquents de l'analyse. Les deux dernières parties seront une lecture de la rationalisation du monde moderne dans deux de ses modalités, soit au niveau des conduites et au niveau des techniques. Cette partie du travail s'articulera beaucoup autour des travaux de Max Weber et de ses interprètes. Nous proposons dans ces deux parties que l'individualisme contemporain est le fruit de rationalisations qui prennent place au niveau des conduites et des techniques pour créer une atmosphère propice à une socialité dont la forme est différente de celle des XVIIIe et XIXe siècles qui auront fait l'objet du chapitre précédent.

Le chapitre final s'attardera d'abord à faire la lumière sur la question de la technocratie — sur les effets concrets de la rationalisation contemporaine des techniques. Nous y proposons, de façon générale, une société dont les impératifs d'organisation (politiques) tournent autour de la technique et dont le positionnement subjectif se déplace d'un rapport substantiel au collectif pour se cantonner dans un rapport de nature technique et procédural. Enfin, nous discuterons de l'autonomie subjective des sujets dans une société de la sorte. Alors que Weber croyait que la domination rationnelle-légale signifiait la fin de l'autonomie subjective, nous croyons que la société contemporaine nous offre

des pistes de réflexion qui ne vont pas en ce sens de façon univoque.

Chapitre 1 : Premiers abords : appropriations conceptuelles et méthodologiques

Avant d'entrer dans l'analyse proprement dit, il sera bénéfique de se lancer dans quelques remarques propédeutiques. Le premier point d'ancrage de notre analyse est le sujet contemporain. Il est donc fondamental d'introduire d'emblée les paramètres qui nous permettront de faire comprendre ce que nous entendons par le sujet contemporain. Des clarifications sur ce que nous entendons par l'individu et l'individualisme constitueront la seconde partie du chapitre. Notre appréhension de la question du politique, ou de l'organisation de la société, ainsi que la façon dont nous envisageons son rapport avec la subjectivité, pour les fins de cette analyse, nécessitera aussi quelques clarifications méthodologiques.

1. 1. Le sujet et la subjectivité

Le sujet contemporain est un sujet moderne. Il est donc un sujet issu et constitué de par l'émergence socio-historique et philosophique de la modernité. Le sujet moderne est aussi une constitution de l'agent humain en tant qu'individu. Il ne s'agit donc pas d'un individu comme entité organique; l'individu ainsi entendu ne serait pas particulier à la modernité. Nous entendons plutôt par « individu » le lieu de phénomènes sociaux nouveaux qui vont impliquer un nouveau rapport de l'agent humain vis-à-vis lui-même, vis-à-vis les autres et le monde en général. L'individu moderne est le lieu de l'investissement, dans un agent humain unique, de la possibilité d'être au centre de son appréhension du monde. Deux éléments importants de la modernité viendront rendre possible cette nouvelle réalité : la pensée rationnelle et scientifique, ainsi que la dignité — qui est au centre de ce que

nous appellerons la pensée morale humaniste au chapitre suivant — qui peuvent tous deux être rattachés de façon plus générale à l'humanisme et à la pensée politique moderne.

La notion de sujet émerge avec l'humanisme et fait référence à ce nouvel investissement, dans l'agent humain, de « la double capacité 1) d'être conscient de lui-même (l'auto-réflexion), et 2) de fonder son propre destin (la liberté comme auto-fondation). » (Renaut, 1995, p. 30) Ces deux notions font appel au rationalisme et à la dignité. La transparence du sujet vis-à-vis lui-même est une conception qui tire son explication de la tradition du rationalisme classique (pensons au *cogito* cartésien). Dans la perspective humaniste, elle est inséparable de la question de la liberté comme auto-fondation. Cette dernière fait référence à l'égalité naturelle de tous les êtres humains, conséquemment également dignes, comme agents déterminants de leur condition; de leur destin. La notion humaniste du sujet émerge en opposition à une conception de l'agent humain comme « objet » de la tradition ou d'une partie du Tout, pour reprendre les mots de Renaut. Ce sujet toutefois n'est pas celui auquel nous nous référons lorsque nous parlons du sujet contemporain.

Nous voulons distinguer trois conceptions du sujet qui se rapportent à l'agent humain, car si le terme « sujet » pour désigner l'agent humain émerge durant la modernité, pour nous, son existence est sociologiquement transculturelle — le sujet est une condition *sine qua non* de la socialité.

1) Le sujet humaniste. C'est lui qui fait l'objet de la description sommaire ci-haut. Au sens du second sujet que nous voulons distinguer toutefois, le sujet humaniste n'est qu'un type de sujet; il constitue une

subjectivité particulière.

2) Le sujet sociologique. Lorsque nous parlons du « sujet contemporain », du « sujet moderne » ou du « sujet » tout court, nous voulons signifier un sujet sociologique qui existe dans un contexte précis. Il est le propre de l'agent humain social dans la mesure où il est le sujet — c'est-à-dire constitué et façonné — de processus de socialisation. Par « subjectivité », on entend ainsi la constitution de l'agent humain comme sujet de la socialisation. Nous verrons que le sujet humaniste, l'archétype philosophique du sujet moderne, est au fondement de la subjectivité contemporaine de l'individu.

3) Le sujet du pouvoir. C'est le sujet en tant qu'agent humain soumis ou asservi à une quelconque hiérarchie, ou à une quelconque relation de pouvoir. Il s'agit d'une domination lorsque l'effet de ce pouvoir donne lieu à un sentiment d'aliénation et/ou d'impuissance.

Restons-en aux deux premiers sujets que nous avons distingués. Si le sujet contemporain est moderne, c'est parce qu'il puise les sources de sa subjectivité dans la conception humaniste de l'agent humain. L'individu doit son émergence au sujet humaniste car, en instituant d'une part l'idée de l'agent humain comme transparent à lui-même, et d'autre part, l'idée de son égale dignité par rapport aux autres, il fallait « que le réel fût conçu comme ensemble d'individualités; mais il était surtout requis que l'individu apparût comme *princeps* et que la totalité, conçue désormais exclusivement à partir de l'individualité ou à travers elle, lui fût subordonnée. » (Renaut, 1995, p. 28) Le sujet humaniste, en effet, a donné lieu à l'individu; soit la fragmentation du social en la somme des individus qui se perçoivent comme autonomes par

rapport à une totalité quelconque. Mais comme le précise Renaut, il a aussi donné lieu à l'individualisme.

1. 2. L'individu et l'individualisme : où, quand et comment ?

L'individualisme est un concept auquel on peut prêter plusieurs significations. Il peut être compris au sens d'une première distinction établie par Pierre Birnbaum et Jean Leca. « L'individualisme peut être un élément d'un processus de *caractérisation* des institutions et des comportements sociaux. C'est en ce sens que l'on parle d'individualisme sociologique, d'individualisme économique ou d'individualisme juridique. » (Birnbaum et Leca, 1986, p. 13) Dans cette lecture de l'individualisme, ce sont les conditions objectives de la société qui sont au cœur de l'explication du phénomène. C'est-à-dire que les structures institutionnelles (la famille, l'école, le droit, l'économie, etc.) de la société fonctionnent selon une conception de l'agent humain comme individu. Or, les mêmes auteurs vont d'une seconde distinction.

> Il peut aussi s'inscrire dans un processus intentionnel de *légitimation* plus ou moins systématique (« doctrinale ») des institutions et des normes et valeurs, notamment politiques [...] L'individualisme éthique et, dans une moindre mesure, l'individualisme « philosophique » entrent dans ce débat ainsi que l'individualisme « politique » si, par exemple, on pose que le « contractualisme » est la seule base rationnelle logiquement concevable de la justification de l'autorité politique, ou que l'agrégation des préférences individuelles est le moyen le moins coûteux (pour la raison et l'éthique) de parvenir à des choix collectifs (Ibid., 1986, p. 13).

Toutefois, cette seconde distinction des auteurs n'est pas étrangère à la première. Car si les conditions objectives favorables à l'existence d'un individualisme sociologique se sont instituées dans la modernité, c'est en partie

à cause de l'émergence d'idées « individualistes » sur le droit, la société, le politique, les institutions ou la subjectivité. Il faut certes distinguer, à l'instar de Birnbaum et Leca, la prise de position « pour » un individualisme politique ou éthique du fait de l'existence sociologique de l'individualisme. « Il est en effet tout autant un mode de vie courante, pas le moins du monde théorisé, pratiqué marginalement ou par toute une société — qu'une doctrine intellectuelle très élaborée fixant les principes d'organisation globale de la société ou n'ayant que valeur d'éthique personnelle. » (Laurent, 1993, p. 8) Mais le chercheur doit garder à l'esprit le fait que le constat sociologique de l'individualisme n'est rendu possible que parce que des conditions sociales, historiques, politiques, économiques, juridiques, toutes ancrées dans la subjectivité — qui ne sont donc pas à l'abri de la « doctrine » — , ont permis l'institutionalisation de l'individualisme et par conséquent, son existence sociologique.

Ainsi, le second individualisme distingué par Birnbaum et Leca relève d'un caractère sociologique plus important qu'ils ne le laissent l'entendre dans la mesure où le premier n'est envisageable que si l'on peut saisir toute la portée du second. Quand nous proposions plus tôt, à la suite de Renaut, que le sujet humaniste donne lieu à l'individu mais aussi à l'individualisme, nous envisagions ce dernier comme un phénomène objectif des conditions de la socialité moderne qui s'inscrit dans une histoire du politique, de l'économique, du juridique, bref, qui est ancré dans la subjectivité, les idées, l'idéologie, la doctrine.

Nous tenterons de montrer, dans les chapitres ultérieurs, comment

l'individu et l'individualisme émergent avec la modernité et l'humanisme mais que le déploiement de l'individualisme, de par des phénomènes sociaux précis, changera pour donner lieu à une socialité nouvelle que nous appelons la société contemporaine. Pour nous, cette socialité, dans ses structures institutionnelles et la subjectivité qui l'accompagne, est de plus en plus marquée par l'existence de l'individu comme *princeps*.

Par société contemporaine nous entendons les sociétés occidentales de l'après deuxième guerre mondiale qui seront éventuellement marquées par la chute des grandes idéologies comme la désacralisation du marxisme surtout, par l'ébranlement des valeurs traditionnelles, par le déploiement de la démocratie, du libéralisme économique, de la société de consommation et de la gestion techno-scientifique du social. La deuxième guerre mondiale marque un point décisif des transformations importantes de la société, mais nous pouvons dire que les changements sociaux significatifs culmineront dans les années soixante[1]. Cette société, dans son appellation, fait l'objet d'un débat vieux de plus de vingt ans. Pour les uns, elle marque une rupture définitive avec la modernité, entrant dans un stade postmoderne, alors que pour d'autres la société contemporaine reste moderne. Notre appellation de « société contemporaine » est pratique pour éviter ces considérations épistémologiquement lourdes de conséquences. Néanmoins, nous croyons qu'il n'est pas possible de comprendre cette société sans scruter les conditions

1. En ce qui concerne l'individualisme contemporain, nous verrons que de façon plus générale, l'Amérique du Nord se démarque considérablement de ce qui eût lieu en Europe. Quoique nous puissions postuler que dans les années 1980 les deux continents vivaient des situaions similaires sur le plan de l'individualisme, l'Amérique du Nord a été beaucoup plus précoce à développer ce que l'on a appelé le « nouvel individualisme ». Il n'en demeure pas moins toutefois que ce fût en Europe et en Amérique du Nord, dans les années soixante, que les conditions objectives du développement de la société et de la subjectivité contemporaines se sont instituées.

d'émergence de la modernité. De même que le terme de postmodernité est équivoque comme le signale Gilles Lipovetsky, un tenant de l'école postmoderniste (Lipovetsky, 1993, p. 113), postuler que nous sommes toujours dans la modernité sans accompagner cela d'explications théoriques solides, le serait tout autant. Qu'il soit question de société contemporaine, de société de consommation, de société postindustrielle ou postmoderne, la modernité reste incontournable pour comprendre sociologiquement ce moment de l'histoire et c'est ce à quoi nous voulons nous en tenir.

Si l'individu contemporain est moderne on ne peut toutefois pas le comprendre, dans sa manifestation spécifique, comme calque de celui qui émerge à l'époque des Lumières et qui va se consolider lors des Révolutions bourgeoises d'Europe et des États-Unis. Si les fondements de la subjectivité de l'individu contemporain sont modernes, les siècles qui séparent son émergence de sa condition actuelle ont provoqué des changements fondamentaux sur ce que l'on pouvait constater de l'individu auparavant, et sur ce qu'on peut constater aujourd'hui. Nous chercherons à soutenir la thèse selon laquelle la grande transformation historique de l'individualisme moderne, jusqu'à aujourd'hui, réside dans le fait que le sujet contemporain, contrairement au sujet moderne, n'existe pas dans un espace où la moralité a des référents sur la collectivité, la solidarité et les visions d'avenir.

Mais même si ce phénomène caractérise la société contemporaine, les constatations sociologiques sur l'individualisation croissante ne datent pas d'hier. Alexis de Tocqueville, au XIXe siècle, dans un chapitre de *De la démocratie en Amérique* (1961) intitulé *Influence de la démocratie sur les*

sentiments des Américains, fait part de ses craintes par rapport au cantonnement des individus dans des « petites sociétés à leur usage » (Tocqueville, 1961, t. 2, p. 143). La crainte de Tocqueville sur le repli sur soi à l'âge démocratique était liée aux effets de la liberté politique, amenée par l'égalité des conditions, qui reposait sur une certaine homogénéité éthique et morale (le puritanisme) des colons américains. Si l'on assiste à cette époque à un mouvement d'individualisation croissant et à un repli sur soi des individus, ce n'est toutefois pas sans que la population partage les mêmes valeurs morales et entretienne un esprit de solidarité. Qui plus est, ce n'est pas non plus sans qu'ils participent activement aux délibérations politiques. Or, le monde contemporain est porteur d'un individualisme qui se manifeste par un repli sur soi mais parallèlement, il témoigne d'une explosion des différences individuelles, à l'évanouissement de l'idée de collectivité et au désengagement politique.

Ainsi, on peut tenter de comprendre le monde contemporain comme celui du temps des tribus (Michel Maffesoli, *Le temps des tribus* (1998)), de « petites sociétés ». On ne peut pas réfuter la place importante du communautaire, des groupes et des sous-cultures diverses dans la société contemporaine; soient-elles axées sur l'ethnie, la culture, le mode de vie, la religion, le quartier de résidence, les intérêts de lobby ou la consommation. Il n'en demeure pas moins toutefois qu'il faille à cet effet retenir au moins trois choses.

D'abord, que ces groupes ne font plus l'objet d'une allégeance[2] « éternelle » ou complètement traditionnelle (comme dans le cas des religions par exemple) de la part des individus. Ces derniers étant très versatiles, on ne peut plus dire des groupes qu'ils composent, qu'ils sont à même de renouveler, avec le même degré de cohérence, le caractère holiste retrouvé dans les sociétés traditionnelles. L'autre argument est que les sociétés occidentales et démocratiques, structurellement et institutionnellement, évaluent toutes choses en termes individuels; les grandes institutions, comme les lois, n'en ont que pour l'individu. Les diverses revendications des groupes sont évaluées par la magistrature, les média, bref, par le discours et les institutions des sociétés occidentales, en termes de droits individuels. Si par ailleurs, c'est le groupe qui est visé par des mesures répressives ou émancipatrices du système institutionnel, l'argument juridique tournera toujours autour de la liberté individuelle. La troisième chose à retenir est que l'individualité est aujourd'hui non seulement objectivement incontournable, mais aussi subjectivement incontournable, dans la mesure où l'individu idéal-typique ne peut que difficilement échapper à la socialisation du moi. S'il fait partie d'un groupe, on peut postuler que c'est autour de la compréhension qu'il a de ce groupe que ce dernier trouve sa pertinence.

Dans ces circonstances, et en dépit de toute allégeance au groupe, l'acteur social contemporain est, en vertu des conditions objectives de la société, un individu avant tout. Nous retrouvons un fondement symbolique de cet état de choses dans la première phrase de La Déclaration universelle des Droits de

2. Nous disons « allégeance » mais l'on pourrait remettre en cause, comme nous l'aborderons au chapitre suivant, le fait que les religions comme cosmologie des sociétés traditionnelles, constituaient des choix ou des allégeances.

l'homme de 1948 : « Tous les êtres humains naissent libres et égaux en dignité et en droits. »[3] Celle-ci implique non seulement la liberté individuelle mais aussi l'évaluation juridique des choses en termes individuels. Les structures dont se sont dotées les sociétés occidentales — celle du Droit tout particulièrement — reposent sur une perspective individualisée de la société. L'individualisation est certes ancrée dans la socialisation des acteurs sociaux contemporains, mais elle est aussi une condition objective de ces mêmes sociétés; une condition qui se situe dans ses institutions et dans ses propres limites.

À cet effet, nous verrons que l'individualisme contemporain peut être compris comme une contrainte si on l'appréhende d'un certain angle, mais qu'il peut aussi être compris comme faisant partie d'un processus de rationalisation ancré dans la subjectivité. Autrement dit, nous croyons possible d'envisager les transformations de la société moderne comme un phénomène subjectif. Si l'expression wébérienne de « cage de fer » peut être employée sans qu'elle soit imprégnée de jugements péjoratifs sur la contrainte, nous pourrions dire que l'individualisme contemporain est une cage de fer — vu les conditions objectives de la société — mais que cela ne constitue pas nécessairement la vision du sujet sur son (incontournable ?) individualité. Nous postulerons que les transformations de la subjectivité moderne peuvent être comprises comme étant au fondement des transformations de l'appareil institutionnel de la société contemporaine et notamment, du changement du statut de la morale dans l'appréhension sociologique du politique. Nous entendons le politique comme organisation de la collectivité et comme une

3. Cette phrase, tirée de la Révolution française, est intéressante dans sa version onusienne dans la mesure où elle survient après la deuxième-guerre mondiale et qu'elle tente d'embrasser comme fait accompli cette valeur pour toute la Terre.

forme de domination — la domination étant entendu au sens wébérien de la domination charismatique, traditionnelle ou rationnelle-légale.

1. 3. Le politique à partir du sujet

Alain Renaut postule avec justesse que « c'est incontestablement dans le registre de la pensée politique des Modernes que l'affirmation de l'individu comme valeur a laissé apparaître ses véritables contours. » (Renaut, 1995, p. 28) Nous verrons, de fait, que l'individu dépend de conditions objectives favorables à son existence et à l'instar de Renaut, nous croyons que c'est à travers la pensée politique moderne que la constitution d'une subjectivité de l'individu se révèle le mieux. Même si nous voulons appréhender la question du politique à partir du sujet, nous devons quand même prendre en considération le fait que le sujet contemporain n'existe pas *ex nihilo*. Autrement dit, c'est en grande partie dans l'horizon d'une conception particulière de la justice et du politique que l'individu a pu prendre forme dans la modernité.

Ce qui nous permet toutefois de détacher l'individu d'un éventuel « déterminisme politique » est que le phénomène de l'individu, dans sa version humaniste, implique une liberté (dont la révélation la plus importante est l'autonomie) individuelle ancrée dans la subjectivité et vécue (typiquement) comme telle par les acteurs sociaux; une autonomie subjective. Ainsi, l'appréhension de la question du politique — ou de la légitimation de l'organisation de la société — à partir des visées des sujets, paradoxalement constitués originellement en grande partie par et dans le politique, est plus

facilement concevable avec la modernité comme toile de fond. Dans cette perspective, nous croyons que l'organisation politique (scientificisée et technocratisée) qui caractérise le monde contemporain ne peut être réduite à une stratégie de domination en vue d'aliéner « les masses », mais qu'elle répond à des besoins subjectifs qui ont su trouver leur voie dans les transformations de la société moderne. Cela mérite quelques clarifications méthodologiques et épistémologiques.

La subjectivité contemporaine est en grande partie celle de l'individualité — la manifestation phénoménologique de l'individu —, ce qui résulte en des différences dans les façons de penser, d'agir, de juger, etc. En dernière instance, c'est l'individu, dans le monde moderne et contemporain, qui accorde une cohérence à sa vision du monde. C'est lui qui est l'agent principal du sens de sa vie. Comme le postule Renaut, l'émergence du sujet humaniste amène des conditions où la réalité ne peut être conçue qu'à partir de l'individualité. Est-ce à dire que pour faire un portrait du sujet en tant qu'individu, il faille scruter l'intentionalité des acteurs sociaux dans sa multitude ?

Dans le quotidien — le monde de la société en général — les activités sociales et les jugements qui y sont liés sont intuitifs. Ils se manifestent en vertu d'une connaissance non médiatisée, non intellectualisée, des phénomènes qui se posent devant les acteurs sociaux. Les activités sociales se réalisent généralement sous fond d'un *world as taken for granted*, pour employer la terminologie schützienne. Il est vrai toutefois, comme le démontre une tradition sociologique qui prend racine chez Max Weber, que

dans les sociétés modernes, l'accès au monde est de moins en moins spontané. Dans le monde moderne il y a plus de choix éthiques et moraux posés devant l'acteur social, dans la mesure où son destin n'est pas prédéterminé par la tradition ou la cosmologie. Ainsi l'individu est aux prises avec des décisions fondamentales qu'il doit rationaliser. Mais cette rationalisation se fait sous fond d'un horizon cognitif familier et reste intuitive, même si elle n'est pas spontanée. Ses choix éthiques et moraux, quoiqu'ils peuvent être calculés et pondérés, ne se font pas pour autant à l'extérieur d'éléments de choix immédiatement accessibles à l'intuition.

Nous voulons ancrer nos définitions de la morale et de l'éthique ailleurs que dans des distinctions qualitatives — celles-ci appartenant plutôt au monde de la subjectivité qu'à celui de la sociologie. Ainsi, nous entendons par éthique les enjeux entourant la conduite des agents humains ou, au sens plus large, leurs façons d'être. La morale quant à elle constitue les jugements qui orientent ou sous-tendent les actions éthiques. Nous tâcherons de démontrer que dans le monde moderne, en vertu de la subjectivité humaniste naissante, la morale comme code substantiel et conventionnel qui auréole de façon plutôt uniforme les intuitions, l'éthique, les façons d'être et les activités des sujets, reste pertinente dans l'analyse sociologique. C'est ainsi que les puritains américains qui font l'objet des études de Tocqueville existent dans une atmosphère politique où l'individualisme est moral. Par ailleurs, notre cheminement sociologique dans l'histoire de l'individualisme et de la pensée politique nous amènera a postuler que le monde contemporain devient un terrain d'activités sociales dont la portée morale s'estompe pour faire place à une dimension plutôt instrumentale. Il en résulte, parmi d'autres conséquences, une

différenciation de plus en plus grande entre les individus, une explosion des moralités au sein d'un système procédural.

Mais notre objectif n'est pas de saisir les tableaux particuliers des intuitions des acteurs sociaux, ni leurs intentionalités multiples. Nous voulons voir comment la modernité a façonné et constitué un sujet particulier (contemporain) et comment ce qui est une subjectivité généralisée de l'individu, en dépit des différences individuelles — justement issues de cette subjectivité —, peut être comprise. Notre intention n'est donc pas de comprendre ce qui apparaît aux yeux du sujet comme relevant de l'éthique ou de la morale par exemple. Nos définitions de ces termes ne vont pas, d'ailleurs, dans le sens d'une distinction qualitative de l'un et l'autre, mais plutôt d'une distinction conceptuelle qui nous permettra de mieux se rapprocher de notre objectif sociologique : comprendre le sujet contemporain en tant qu'individu et son rapport au politique, à la domination.

Le sociologue est un peu laissé au dépourvu quant il s'agit de donner un verdict objectif et vrai de l'activité sociale. À cet effet, le travail de Max Weber sur la sociologie compréhensive est révélateur. Dans l'optique wébérienne, l'activité sociale n'est saisissable que dans la mesure où on l'évalue par rapport à des types idéaux — constructions théoriques du chercheur qui viennent proposer une explication plus vraisemblable que vraie. Ainsi, la connaissance vraie de ce qui concerne l'activité sociale est réduite à sa compréhension dans le cadre d'intellectualisations idéales-typiques. Le travail d'Alfred Schütz, lui aussi, va en ce sens; il faut typifier les activités pour pouvoir les saisir. Mais ce n'est que de façon typique qu'elles peuvent être saisies. Nous voulons inscrire

notre entreprise de traiter du sujet moderne dans une telle perspective. Nous n'avons pas l'intention de traiter de tous les phénomènes sociaux modernes et contemporains qui tournent autour de l'individu. Nous voulons plutôt dresser un portrait vraisemblable du sujet en tant qu'individu qui saura nous être utile et révélateur dans l'appréhension de notre objet d'étude.

Chapitre 2 : Enquêter la modernité vers un portrait du sujet en tant qu'individu

Si ce chapitre sillonnera presque exclusivement dans la pensée théorique de Charles Taylor, c'est parce que le travail qu'il a entrepris au niveau de la compréhension de la subjectivité moderne, surtout dans *Les sources du moi* (1998), est à même de donner un constat qui nous semble assez juste des éléments fondamentaux qui la constitue. Certes, nous n'entreprenons pas dans ce chapitre de résumer les thèses de Taylor; certaines seront omises, d'autres nuancées, d'autres discutées, d'autre sociologisées. Il n'en demeure pas moins que le travail de Taylor a été une inspiration pour l'objet de ce chapitre — fondamental pour le reste de notre analyse — tout comme le moyen de saisir les grands enjeux qui concernent le sujet moderne et, nous le verrons, le sujet contemporain.

Aussi, la perspective adoptée par Taylor se rapproche beaucoup de la façon dont nous cherchons à envisager le sujet moderne. Il nous semble qu'en ayant une perspective de la subjectivité moderne non médiatisée *a priori* par le politique, nous serons à même de mieux saisir par la suite les enjeux du politique. Or, Taylor offre la possibilité d'initier une réflexion de ce genre. Ce chapitre aboutira à une réflexion autour des thèmes de l'éthique et de la morale et nous permettra d'approfondir la définition provisoire du sujet que nous avons élaboré au dernier chapitre. Nos définitions de l'éthique et de la morale, essentielles au développement de ce chapitre, sont d'ailleurs fortement inspirées du travail de Taylor, conceptions à leur tour fortement influencées par une perspective qui prend comme point de départ le sujet.

Jürgen Habermas propose qu'il faille établir une différenciation cognitive

et ontologique entre la morale et l'éthique. On doit selon lui distinguer l'éthique de la morale dans la mesure où l'éthique relève de la vie bonne, alors que la moralité relève d'une justice universellement valide. Ainsi, pour Habermas, il est possible de reconnaître — essentiellement par voie de la communication rationnelle — ce qui ne relève que du bien (éthique) et ce qui relève d'une justice universellement valide (morale). Pour Taylor, la dichotomie théorique habermasienne d'association entre le bien et l'éthique d'une part, et la justice et la moralité d'une autre, est questionnable :

> Cette philosophie morale [contemporaine] a eu tendance à se pencher sur ce qu'il est juste de faire plutôt que sur ce qu'il est bon d'être, à définir le contenu de l'obligation plutôt que la nature de la vie bonne; en outre, elle n'a concédé aucun espace conceptuel à une notion du bien en tant qu'objet d'amour ou d'allégeance [...] (Taylor, 1988, p. 15).

Il n'y a pas, selon Taylor, de distinction théorique forte qui puisse être envisagée entre vie morale et vie éthique. Il y a plutôt une convergence des deux termes en une seule et même manifestation[4] et son projet est en partie de réhabiliter cette conception dans la compréhension de l'acteur social. « Il se trouve que le moi et le bien, autrement dit, le moi et la morale, s'entremêlent de façon inextricable. » (Ibid., p. 15) L'éthique individuelle pour Taylor est inextricablement morale. De là émerge notre postulat de la conjonction inévitable entre les conduites (éthique) et les jugements (morale). Ainsi la moralité pour Taylor, contrairement à ce que laisse supposer la thèse habermassienne — où la morale est un critère de justice, donc de régulation de la réciprocité —, dépasse les seuls impératifs de respect de son prochain ou d'entente sur la justesse morale ou normative des rapports qu'il est permis ou

4. Il est à noter toutefois que Taylor emploi le terme éthique quand il est question de traiter d'une façon d'être, de la question des conduites; le « comment devrions-nous vivre ? » (Taylor, 1998, p. 80) Cela ne compromet en rien le fait qu'il considère cette éthique comme étant inextricablement morale.

non d'entretenir avec autrui. Pour Taylor donc, le jugement moral, qu'il envisage dans l'appelation de « discriminations qualitatives fortes », est impliqué dans le juste autant que dans le bien : le moi cherchant à orienter sa vie vers un bien l'oriente inévitablement vers des « discriminations qualitatives fortes ». La morale ainsi entendue embrasse en une seule considération l'éthique et la morale habermassienne. Il écrit : « Pour comprendre notre univers moral, nous devons comprendre quelles idées et quelles images fondent notre sens du respect d'autrui, mais aussi quelles idées et quelles images sous-tendent nos conceptions d'une vie pleine. [...] celles-ci ne constituent pas deux ordres complètement séparés. » (Ibid., p. 29 - 30) L'entreprise de ce chapitre est de clarifier les implications de cette proposition théorique importante sur une conception idéale-typique du sujet en tant qu'il se constitue dans la modernité comme individu.

2. 1. Des grandes dimensions morales de la subjectivité moderne

Taylor postule que la vie humaine ne peut être envisagée par des scissions ontologiques entre une sphère morale et une sphère éthique qui opéreraient (sinon immédiatement, du moins potentiellement) à différents niveaux de l'appréhension du monde du sujet. Il identifie « trois axes de la pensée morale ». Selon lui, les trois axes sont le lieu privilégié de la manifestation du jugement qualitatif de la vie humaine, et par extension, trois axes privilégiés de la vie humaine tout court. Nous nous appliquerons à les présenter d'abord. Nous verrons toutefois que ces catégories ne sont pas sans poser quelques problèmes que nous tâcherons d'identifier. Elles sont; le respect d'autrui; la plénitude de la vie; et la dignité. Pour lui ces trois axes sont moraux

mais contribuent à une orientation éthique de la vie.

Le premier renvoie à la question de la réciprocité, donc à la façon d'envisager autrui en général. C'est ce qui constitue le fait que nous croyons avoir certains devoirs à l'égard d'autrui, ou du moins, certains droits à lui concéder. Le second renvoie à la notion de la vie bonne; viser la plénitude de la vie c'est viser une vie bonne. Le troisième axe identifié par Taylor, la dignité, renvoie quant à lui à l'image que nous nous forgeons de nous-même en fonction d'autrui. Cela peut être compris à partir de l'attitude respectueuse ou non que nous croyons susciter auprès des autres. La dignité est donc selon Taylor « [...] le sentiment que nous avons de nous-mêmes en tant que nous imposons le respect (d'attitude). » (Ibid., p. 30) La dignité est un respect qui est plus substantiel que le simple respect relatif à la réciprocité dans la mesure où il fait référence à plus qu'un simple respect formel mais bien à des valeurs relatives à toute la complexité de notre pensée morale : « Elle peut traduire notre pouvoir, le sentiment que nous avons de dominer l'espace public; notre immunité au pouvoir; ou notre indépendance envers lui, parce que notre vie possède son propre centre; ou l'affection et l'attention que nous portent les autres. » (Ibid., p. 31)

Ainsi, pour Taylor, sont aussi morales les unes que les autres les questions de notre respect envers les autres, les bases sur lesquelles on juge bien ou mal d'interagir avec autrui de certaines façons; notre façon de concevoir que notre vie est vécue pleinement ou non; ou encore le sentiment que nous avons de susciter le respect de ce que nous sommes dans notre entièreté, dans notre façon d'être; la dignité. Nous reviendrons plus loin sur ce qui fait de ces

dimensions des dimensions morales. De ces trois axes que nous propose Taylor, on doit toutefois s'interroger sur certaines choses. On doit remettre en cause la validité du postulat selon lequel : « Ces trois axes existent peu ou prou dans toutes les cultures. » (Ibid., p. 31) Nous verrons que ces catégories puisent toutes leur pertinence dans la modernité plutôt que dans l'universalité.

Dans *La politique de reconnaissance* (1994), tout comme dans *Grandeur et misère de la modernité* (1991), Charles Taylor insiste sur une distinction fondamentale à faire entre les sociétés modernes et les sociétés pré-modernes. Les premières fonctionnent sur le mode de la reconnaissance de la dignité de tout un chacun, un corollaire de la notion moderne (politique) de l'égalité. Les secondes quant à elles fonctionnent d'après le mode de l'honneur, qui justifie les inégalités et qui rend légitime le pouvoir (aristocratique, monarchique, féodal, ecclésiastique, esclavagiste, etc.). Pour Taylor, de toute évidence, la notion de dignité ainsi comprise n'est pas tout à fait similaire à celle de la « dignité de la pensée morale » retrouvée dans *Les sources du moi*. La première est plutôt objective, faisant référence à un changement survenu dans les structures sociales; un passage de l'honneur à la dignité dans le discours et les institutions sociales. L'autre, quant à elle, est un besoin transcendant de l'humanité : la recherche du respect d'attitude serait, dans la perspective taylorienne, un besoin humain vital et universel. Pour Taylor, une société fondée sur l'honneur n'exclut pas le besoin, voire l'exigence, qu'est la dignité, pour chacun de ses membres. Le « respect d'attitude » devient en quelque sorte un besoin de reconnaissance universel.

Nous voulons garder la notion de dignité comprise en tant que

contrepartie à l'honneur : nous n'abondons pas dans le sens du double emploi du terme que laisse supposer Taylor. Nous croyons que la dignité — non le mot, qui date du XIe siècle, mais sa compréhension théorique contemporaine que Taylor éclaire bien, soit la reconnaissance, le respect d'attitude que l'on cherche à obtenir — prend racine dans la modernité et dans le sujet humaniste et dans leur conséquente valorisation de l'individu. Selon nous, une conception de la dignité qui se manifesterait comme une quête quelconque n'est possible que si l'agent humain a l'espace nécessaire pour le faire. Nous tenterons de démontrer que le respect d'autrui, la vie bonne et la dignité sont trois catégories de la pensée morale spécifiquement moderne et qu'elles existent toutes trois en fonction de la même exigence « épicentrique » qui permet le développement d'un espace nécessaire à leur existence : la constitution du sujet moderne. La vie bonne a selon nous une étroite relation avec la question du sens, spécifique à la modernité, et est manifestement liée au développement du besoin de reconnaissance moderne. Par ailleurs, nous croyons que la question du respect d'autrui est un thème résolument moderne qui puise ses sources dans la pensée politique et qui révèle aussi les contours de la reconnaissance, de la dignité.

2. 2. Le sujet moderne comme sujet du sens

Selon Taylor, les questions de la vie bonne et pleine sont transculturelles. Pour lui, « la renommée transmise par la mémoire et le chant de la tribu, l'appel de Dieu qu'explicite la révélation, ou, pour prendre un autre exemple, l'ordre hiérarchique de l'être dans l'univers », sont toutes des façons dans l'histoire et dans les cultures par lesquelles « on évalue sa vie et on

mesure, pour ainsi dire, sa plénitude ou son vide. » (Ibid., p. 32)

À cet effet, peut-on assurément postuler que toute vie est évaluée par son sujet en terme de vide ou de plénitude ? Selon nous, un tel postulat, quoiqu'intuitivement vraisemblable, n'est pas à même de saisir toute la portée de la spécificité sociologique de la subjectivité moderne. Si d'ailleurs, intuitivement, nous sommes capables d'envisager cette possibilité, c'est parce que nous nous situons à une époque où la thématisation de la vie bonne au quotidien, par tout un chacun, est centrale; nous sommes traversés, et en partie constitués, par cette préoccupation angoissante, comme Christopher Lasch le dépeint dans *The culture of Narcissism* (1979). Nous tenterons de démontrer que les questions qui portent sur la plénitude de la vie sont inséparables de la problématisation du sens de la vie. Taylor fait la remarque selon laquelle l'impératif du sens dans la modernité est une modalité nouvelle de la plénitude de la vie. Pour nous, l'un et l'autre n'ont pas d'existence indépendante et ne sont concevables que pris ensemble.

> L'une des manières les plus nettes par laquelle notre époque tranche avec les époques antérieures concerne le deuxième axe. Tout un ensemble de questions ont du sens pour nous aujourd'hui, qui ont trait à la signification de la vie, qu'on n'aurait pas pu comprendre auparavant. Il arrive souvent aux modernes de se demander avec anxiété si la vie a un sens et en quoi il consiste (Ibid., p. 31 - 32).

Taylor semble postuler que notre époque tranche d'avec les autres par la forme que prend le second axe. Selon nous, cette soi-disant « forme moderne du second axe » — le sens de la vie — est intrinsèquement liée à la question de la vie bonne. Taylor a raison de postuler que c'est dans la modernité que la question du sens de la vie émerge. Cela fait en sorte qu'elle constitue l'essentiel

des interrogations — implicites ou explicites — sur la vie bonne. Mais surtout, l'émergence du sens de la vie pose les conditions de la possibilité d'une interrogation sur la vie bonne. Donner un sens à sa vie et l'évaluer en terme de vide ou de plénitude, n'était pas un problème, ni même une possibilité, pour un membre d'une société traditionnelle.

Dans ces sociétés, la question du sens n'était pas présente à cause surtout d'une absence subjective de transparence à soi et de réflexivité. Ce qui était présent — mais qui ne se faisait pas sentir comme le sens se fait sentir chez les modernes — était une cosmologie unique et toute-puissante qui n'impliquait aucun questionnement. Il ne s'agit pas de dire que tout allait toujours bien. Il s'agit plutôt de dire que les événements de la vie étaient tributaires — « malheureux » comme « heureux » — d'une intégration à quelque chose qui les dépassait et qui dispensait le sujet de l'investir lui-même d'un sens quelconque. Les préoccupations sur le sens et sur la vie bonne n'étaient pas, comme dans la modernité, au centre des activités sociales. Dans ces circonstances le questionnement sur la vie bonne, tout comme celui sur le sens, n'a pas de pertinence sociologique — c'est-à-dire comme facteur déterminant de la socialité — à l'extérieur du monde moderne. Quoique le thème de la vie bonne ait été thématisé dès l'âge classique, sa portée ubiquiste comme préoccupation surtout quotidienne est spécifiquement moderne.

Si Weber postule que les sociétés pré-modernes bénéficiaient d'un enchantement du monde c'était à cause de la cohérence cosmologique dont elles étaient tributaires. Nous pouvons postuler que d'évaluer cela en terme d'enchantement est typiquement moderne, parce que symptomatique de

l'incertitude et du questionnement que constitue l'existence d'un individu confronté au sens de la vie et à sa plénitude. On peut mettre en doute le postulat selon lequel le chant de la tribu, l'appel de Dieu ou l'ordre hiérarchique de l'être dans l'univers, pour reprendre les exemples de Taylor, étaient enchantants et vécus comme tels : ils étaient des façons incontournables d'envisager le monde, ils étaient des nécessités structurelles et institutionnelles de la société. Pour le sujet moderne toutefois — et c'est de plus en plus vrai aujourd'hui —, confronté à l'incertitude, au scepticisme, à une multitude de choix, une telle chose peut apparaître comme un état d'enchantement.

Pour conceptualiser l'idée du sens de la vie, il faut nécessairement comprendre l'éventualité du non sens et dans nos sociétés, nous y sommes tous confrontés : c'est ainsi que le sens fait sentir sa lourde présence. Dans cette perspective, l'individu moderne est un sujet du sens. Pour prendre l'exemple du Québec de la dernière décennie où le taux de chômage et de suicide élevé chez les jeunes sont parmi les grands thèmes médiatiques : dans la façon dont elles sont traitées, ces questions font immédiatement référence, implicitement ou explicitement, à une crise de sens (d'identité[5]). Quand ce malaise social est médiatisé ou discuté, il trouve sont point d'ancrage principal dans un questionnement sur le sens. Dans les sociétés contemporaines, on parle de perte de sens, de crise de sens, de trouver un sens et ce que cela révèle est la présence incontournable du sens comme enjeu individuel, communautaire, ou social.

5. Dans les média c'est plus souvent d'identité dont il est question. Notons qu'elle est immédiatement rattachée à la question du sens car c'est elle qui la révèle : en faisant manifester son identité, on parle du sens de sa vie.

Si pour Taylor la vie bonne en tant que second axe de la pensée morale n'est rattachée au sens de la vie que dans la modernité, pour nous l'un et l'autre, la vie bonne et le sens, doivent nécessairement s'accompagner. Autrement dit, on ne peut prétendre à l'existence d'une préoccupation sur la vie bonne et pleine sans que la question du sens soit en son centre. Ainsi, le second axe de Taylor est typiquement moderne. L'individu comme phénomène sociologique est une chose moderne. Or, il a rendu possible l'existence d'une conscience particulière de l'agent humain à l'égard de lui-même. Il s'agit du moi moderne, imprégné des valeurs humanistes de transparence à soi, de la rationalité comme source de l'entendement humain et par conséquent comme « donneur de sens ». Ainsi, la révolution moderne de la façon d'appréhender le monde — une révolution humaniste et scientifique qui a mis au centre de l'appréhension du monde l'agent humain et son entendement en faisant table rase du cosmos et de la divinité — a aussi mis en place les conditions qui allaient faire de l'individu le donneur de sens du monde, de « son » monde.

C'est aussi en grande partie dans le fait que le quotidien, au lendemain de l'ébranlement de la dimension sacrée de la socialité divine ou cosmologique, s'est vu investi d'une importance fondamentale pour les modernes, que la nécessité de donner du sens à pris de l'importance. Dans ces circonstances, la question du sens a imposé sa présence au centre du moi moderne. Mais le non-sens et l'absurde aussi ont émergé à cause de l'impératif de l'explication scientifique et du rejet des explications cosmologiques qui pouvaient intégrer tous les phénomènes de la vie; de là les thèses du désenchantement et l'enjeu fondamental qu'est le sens.

Les thèses de Max Weber sur la rationalisation sont révélatrices sur ce thème. En effet, pour lui les rationalisations du monde antérieures à la modernité se sont distinguées dans leur façon de freiner un plus grand développement de la rationalisation par le recours à la divinité ou au cosmos : tout faisait l'objet d'une intégration cosmologique. La modernité, en abolissant ces freins, est entrée dans une rationalisation poussée à son paroxysme : il parle de la rationalisation moderne comme d'une intellectualisation, c'est-à-dire l'instrumentalisation et la désacralisation de l'appréhension du monde. En proposant une appréhension du monde (scientifique) qui saisirait son sens réel et positif, la modernité se bute au paradoxe des conséquences : un désenchantement du monde qui va puiser ses sources dans un excès (parce que non-stop et sans limites) de sens (rationnel) attribué aux choses. La thèse de Weber abonde implicitement dans notre direction. Elle implique que la question du sens ne soit possible et envisageable que dans la modernité où le sens côtoie le non-sens.

Émile Durkheim faisait une grande distinction quant à l'individualisation qui sous-tend l'indépendance de l'individu à l'égard du groupe dans les sociétés modernes et la faible individualisation des sociétés traditionnelles. Ces dernières n'existaient que dans la mesure où le « moi » et autrui étaient confondus dans une cosmologie englobante de tous les phénomènes du monde de la vie. L'émergence de l'individualisation est ce qui a permis au moi moderne et au sens d'émerger, tout comme à l'idée de la vie bonne comme un bien. Aussi, la question du sens et de la vie bonne n'est pas étrangère à celle de la dignité. Dans le contexte moderne, l'importance grandissante que prend le sens au centre de la vie des individus, tout

particulièrement dans le quotidien, est relié à la question de la dignité. Ainsi l'individu moderne puise-t-il sa dignité dans ses aspirations, ses valeurs, son statut, bref, ce qui *à ses yeux* constitue la plénitude et le sens de sa vie.

Les répercussions de cette subjectivité d'origine typiquement moderne sur le monde contemporain sont évidentes. Le pluralisme des sociétés occidentales n'est envisageable que si l'on peut comprendre les fondements socio-historiques et philosophiques qui ont permis à l'individu de devenir l'élément déterminant du sens de sa vie et de sa quête d'authenticité et de reconnaissance en tant que moi particulier. Mais si le monde contemporain se caratérise justement par une grande individualisation, ou par ce que certains définissent comme un repli sur soi, il n'en demeure pas moins que la nature du rapport à autrui a lui aussi été l'objet de transformations. Si la dignité et la vie bonne ont des conséquences importantes sur l'univers personnel et quotidien du sujet contemporain, l'existence de cette nouvelle subjectivité dépend aussi de l'institution d'un nouveau rapport à l'autre.

2. 3. La question de soi et d'autrui dans la subjectivité moderne

La dignité concerne d'une part, le rapport du moi à l'égard de lui-même : le sens et la vie pleine, dans le monde moderne, sont une affaire individuelle, tout comme l'impression de dignité (ou d'authenticité et de reconnaissance subjective) qui s'en dégage. Mais par ailleurs, la dignité moderne ne peut être sans un rapport avec les enjeux qui relèvent d'autrui : c'est à partir du premier axe de la pensée morale moderne, le respect d'autrui, que l'on peut comprendre cet autre aspect de la dignité.

Aucune idéologie, aucune prise de position sur le monde, depuis la modernité, ne peut être cohérente sans une justification qui relève du respect d'autrui et de la dignité tant ils sont institutionnellement forts et ancrés dans la subjectivité. Le racisme est un bon exemple. Le fait qu'il puisse bénéficier d'une légitimité relève d'un discours qui place certaines personnes dans une situation « objective » d'indignité. Ainsi, la question du respect d'autrui ne se pose pas : c'est en partie pourquoi le racisme n'est jamais sans fondements « scientifiques » ou « empiriques » (soient-elles reliés à la « race » ou à des explications conservatrices sur la santé économique de l'État, par exemple).

La dignité est l'image que nous nous faisons de nous-même. Si elle est aussi ce que l'on croit représenter aux yeux des autres, il n'en demeure pas moins qu'elle part de nos principes et de nos préférences, car nous ne sommes pas en quête de respect de la part de ceux que l'on répudie. Ainsi, la dignité joue un rôle fondamental dans la reconnaissance du statut égalitaire ou non d'autrui. Ce rapport entre dignité et égalité est caractéristique de la façon moderne d'appréhender autrui. Il s'inscrit dans la distinction entre l'honneur et la dignité que nous avons brièvement rapporté plus haut, dans la mesure où il présuppose tout un chacun comme étant égaux. Il est difficile de séparer cette conception morale d'avec celle de la dignité faisant le pont entre la valorisation de l'individu « comme chose » (le respect d'autrui) et du moi « comme lieu » (avec un droit à la vie bonne). Dans cette perspective, il y a lieu de comprendre les trois axes de la pensée morale que nous avons rapportés comme les piliers de la subjectivité moderne, et nous le verrons, contemporaine. Ces trois dimensions s'articulent et se consolident les unes et les autres.

Les sociétés occidentales issues de la modernité, c'est-à-dire les sociétés de droit, comportent implicitement en leur sein les idées de respect d'autrui, de vie bonne et de dignité. Ainsi, les « droits naturels » accordent à tout un chacun une égalité légale leur permettant de mettre à jour une individualité. Or, cela n'a de sens que si l'on postule que tout un chacun est digne de bénéficier de ce respect. Comme nous le disions plus tôt, et à l'instar de Renaut, c'est effectivement dans le registre de la pensée politique moderne que le dévoilement de la valorisation de l'individu prend son sens. L'idée des droits naturels et celle de la dignité sont des critères moraux certes, mais des critères moraux institués d'abord et avant tout comme critères politiques, d'organisation des rapports formels entre les individus. Et dans la logique de cette rationalisation séculière du monde, le droit à la mise en valeur de l'individualité devait être aussi acquis : la création de l'individu créait le lieu d'épanouissement d'une culture individualiste.

Il est intéressant de regarder les fameux principes de la Révolution française énoncés dans la Déclaration des droits de l'Homme, *Liberté, Égalité, Fraternité*, qui « ramasse tout l'héritage européen des Lumières » (Laurent, 1993, p. 42), pour tenter de les mettre en parallèle avec les trois catégories de la pensée morale moderne. L'égalité fait penser au respect formel d'autrui, soit ses droits et le devoir que nous avons de ne pas les entraver; la liberté fait penser à la possibilité pour l'individu de mettre à jour son individualité; la fraternité quant à elle peut renvoyer à la dignité, dans la mesure où elle met de l'avant l'idée selon laquelle nous sommes tous différents mais également dignes de respect parce que fondamentalement humains. Cette interprétation est très littéraire mais on ne peut rejeter le parallèle entre ce grand principe

politique moderne et libéral et les trois axes de la pensée morale moderne. À cet effet Laurent écrit que « [...] l'individualisme libéral ne se conçoit pas sans la forte adhésion à une éthique exigeante de la liberté, fondée sur les vertus d'initiative et de responsabilité individuelle, de confiance en soi et de respect tolérant des autres. » (Ibid., p. 50)

Il va sans dire que les catégories tayloriennes sont révélatrices d'une subjectivité proprement moderne et que c'est là qu'elles puisent toute leur pertinence : autour de chacun des axes de Taylor les questions du moi, de l'individualisme et des droits naturels sont inévitables. Or, ces dimensions ne sont saisissables que dans l'horizon de l'émergence philosophique et socio-historique de la modernité. Nous sommes à même de constater avec ces axes de la pensée morale les fondations du sujet moderne et elles nous permettrons de comprendre les changements de cette subjectivité dans le cadre de la société contemporaine. Toutefois, ce thème de la transformation de la société moderne, puisque nous cherchons à l'aborder dans l'optique de la subjectivité, doit être précédé d'une meilleure compréhension de la question du sujet.

2. 4. La convergence de l'éthique et de la morale : l'incarnation subjective

> *La reconnaissance de l'égale dignité de tous les hommes dans la vie sociale — que traduit, dans la vie politique, le principe de citoyenneté — me paraît le seul fondement légitime, à la fois moralement et politiquement, de la vie collective [...]. Je suis consciente du fait que cette conviction est liée à mon destin particulier dans une société particulière et à l'intériorisation des valeurs démocratiques.*
> *Dominique Schnapper, La relation à l'Autre, p. 28*

En dépit de nos interrogations sur les catégories tayloriennes qui remettraient en cause leur valeur ontologique et transculturelle pour les

ramener sur un terrain sociologique et localisé, ces dernières restent très pertinentes dans l'horizon du monde moderne et contemporain. Elles constituent des types idéaux révélateurs de la subjectivité socio-historique qui nous intéresse. Mais il y a autre chose à laquelle s'attarde Taylor pour faire avancer son entreprise et qui, cette fois, bénéficie selon nous d'une existence ontologique; ce sont les cadres. Qui plus est, ce que révèle ce concept est indéniable dans la mesure où l'on veut se positionner sociologiquement. Les cadres, tels que Taylor les définit et les explique, sont ce à quoi la sociologie fait référence pour tenter de comprendre l'activité sociale, la manifestation du social. Les cadres sont aussi ce qui fait de nous, sociologiquement, des sujets; c'est-à-dire des agents proprement sociaux dont la détermination est issue des mécanismes de socialisation. Les cadres orientent donc les modalités particulières de la subjectivation en tant que dénominateur commun de la vie humaine proprement sociale. Dans ce passage de Taylor nous reconnaissons une définition générale des cadres en fonction de la subjectivité moderne idéale-typique que nous avons mise de l'avant plus haut.

> Les cadres fournissent le contexte, explicite ou implicite, de nos intuitions, réactions ou jugements moraux [...]. Définir un cadre, c'est expliciter ce qui donne un sens à nos réactions morales. Autrement dit, lorsque nous essayons d'expliquer ce que nous présupposons quand nous jugeons qu'il vaut vraiment la peine de mener une certaine forme de vie, lorsque nous attachons notre dignité à une certaine réussite ou à un certain statut ou lorsque nous définissons nos obligations morales d'une certaine façon, nous nous trouvons à expliciter entre autres choses ce que j'ai appelé des « cadres » (Taylor, 1998, p. 44).

Les cadres sont inévitables puisqu'il est impossible pour un agent humain social — agissant, jugeant, vivant — de s'en passer : l'acteur social, puisqu'il est un sujet, n'est possible que dans l'horizon de cadres quelconques. Pour le sujet moderne, les cadres sont à l'origine d'une « subjectivité de

l'individu » dont une des modalités est essentiellement la conscience d'être un individu bénéficiant d'une liberté relative parmi d'autres individus et d'un pouvoir d'introspection quant aux modalités de son existence; le moi. Nous avons vu aussi trois axes idéaux-typiques de la pensée morale moderne — corollaire à la subjectivité particulière du moi moderne — qui ont leur ancrage dans le discours philosophique, social et juridico-politique et qui accompagnent l'émergence socio-historique de l'individu dans l'Occident moderne.

Les trois axes — mais le second et le troisième plus spécifiquement —, contribuent aux impératifs qui entourent la question du sens : l'individu, en tant que moi, oriente sa vie en fonction d'un sens qui représente pour lui un bien-être (vie bonne) et qui lui fait bénéficier d'une reconnaissance (dignité) à ses yeux. Dans le monde moderne, les cadres sont donc les assises générales de l'individu qui se comprend comme tel mais aussi, à un autre niveau, les assises individualisées du sens que prend son individualité (la manifestation pragmatique de l'individu). Mais si l'individu moderne a émergé, c'est conjointement à la conscience (et au respect) de l'existence d'autrui (d'autres moi). L'axe premier de la pensée morale moderne, l'axe du respect d'autrui donc, dans la forme qu'il prend depuis la modernité (droits naturels et rationnels), est une condition idéelle et matérielle de l'émergence du moi moderne. Si le sujet moderne est libre de donner un sens à sa vie (en vertu de cadres individualisés ancrés dans la famille, la religion, l'éducation, etc.), il n'en demeure pas moins qu'il partage, collectivement, une subjectivité du moi qui a ses fondations légitimantes dans le respect d'autrui et la dignité — la dignité, nous l'avons vu, qui renvoie aussi bien à soi qu'à l'autre. Aussi le moi moderne est-il une subjectivité dont l'existence repose sur l'autonomie du

sujet.

Il est important de souligner toutefois que cette compréhension de la subjectivité moderne et contemporaine est un type idéal. Rien n'exclut la possibilité d'une communauté — ou d'un groupe, ou d'une personne — qui serait complètement en dehors de ce type de subjectivité. Toutefois, la réalité institutionnelle, discursive, légale, politique et sociale de l'Occident, voire même du monde entier aujourd'hui, nous confronte tous à ce « modèle » en vertu de « conditions objectives ». Ne serait-ce que la façon d'envisager les litiges sur les droits de certains groupes sociaux : ces derniers sont traités en fonction d'une compréhension de l'individu libre, mais soumis aux impératifs de la collectivité dans l'esprit des droits naturels et rationnels issus de la modernité. Légalement, le groupe est *une* personne morale et ce n'est qu'en vertu des individualités qui le composent et qui l'entourent qu'on lui concède ou non des droits.

D'un point de vue donné donc, les conditions objectives de la société moderne et contemporaine surtout — puisque la gestion rationnelle-légale de la société se consolide, comme nous le verrons au quatrième et cinquième chapitres — confrontent nécessairement l'acteur social à l'individualité. Nous verrons que dans le cas de la modernité, c'est plutôt une moralité politique et collective qui vient légitimer l'individualisme, alors que dans le cas de la société contemporaine, c'est la performativité et la consistence du système qui jouent ce rôle. Dans les deux cas, l'individu et ce qui s'y rattache est une contrainte institutionnelle et structurelle inévitable. Or, cela laisse-t-il droit de cité au concept de l'aliénation comme phénomène objectif ? À savoir que les

institutions priveraient l'acteur social de quelque chose qui lui revient par nature, comme l'autonomie ? Ou, par ailleurs, ces deux conditions socio-politiques sont-elles plutôt des cadres institutionnels différents dont les sujets se sont prévalus pour la mise à jour de leur autonomie ? À savoir une autonomie qui n'existe que dans les limites de la subjectivité ? Cette interrogation est importante dans la mesure où nous tentons de réfléchir l'organisation des rapports sociaux à partir de leur appréhension subjective. Les thèses tayloriennes sur les cadres peuvent nous fournir des pistes de réflexion à cet effet.

Les conceptions de la moralité qui la définissent comme une justice universellement valide présupposent implicitement la possibilité d'une aliénation objective du sujet, soit une aliénation indépendante de sa reconnaissance subjective. Dans le cadre des catégories morales modernes, cette aliénation serait définie par le manque d'autonomie. Dans l'optique toutefois d'une conception de la morale comme véhicule subjectif des jugements accompagnant toutes les activités sociales, une telle conception de l'aliénation semble plus difficile à soutenir. Aussi l'autonomie du sujet que cette conception de la morale présuppose prend un visage différent : c'est le moi moderne et le sens individualisé qui devient le lieu unique de l'autonomie.

Les cadres, dans l'optique taylorienne, sont au fondement de toutes nos intuitions sur le monde. Ces intuitions ne peuvent être comprises que dans l'optique de la moralité, c'est-à-dire que nos intuitions, qui régissent notre appréhension et notre compréhension du monde, sont des « discriminations qualitatives fortes ». Que ce soit une question de justice ou une question de

préférence esthétique qui concerne les jugements ou les actions des acteurs sociaux, les deux se situent, selon Taylor, dans un horizon de moralité, de jugement qualitatif. L'éthique individuelle, le bien pour soi, selon Taylor, ne peut pas exister à l'extérieur ou indépendamment de l'espace moral. Toute prétention à pouvoir le faire faillirait à reconnaître la réalité sociologique des sujets, mais aussi à reconnaître que l'éthique est empreinte d'un cadre, c'est-à-dire qu'elle constitue une subjectivité particulière.

Pour Taylor, les cadres impliquent que les moralités sont ontologiques. Autrement dit, les assises (cadres) sur lesquelles sont fondés les jugements moraux des acteurs ne relèvent pas d'une fiction[6], d'un sens de la vie compris comme tel ou même envisagé comme tel, mais d'un sens de la vie qui, pour eux — c'est-à-dire dans leur incontournable subjectivité —, bénéficie du critère de réalité, de fait. Or, cette réalité aux yeux du sujet est ontologique dans la mesure où la vérité sur le monde ne peut être que celle qui est reconnue comme telle par les sujets : l'histoire et l'existence de la société étant inextricablement liées à l'histoire de la subjectivité et de la socialisation. C'est-à-dire qu'il n'existe pas de réalité objective et transculturelle qui transcenderait les différences puisqu'elle impliquerait la désincarnation de la subjectivité. Or, l'impératif de la socialisation, pour ce qui est de l'existence de l'agent proprement social, et toute la charge incontournable de subjectivation qu'elle implique, ne permet pas d'envisager l'existence d'un agent social non-sujet. Les propos de Taylor l'emmènent à postuler que toute appréhension du monde est

6. Les cadres sont peut-être des fictions, ou du moins sont-ils des fruits de l'imaginaire social des cultures, des civilisations, des individus. Si Taylor postule qu'ils ne sont pas des fictions c'est parce qu'ils n'ont pas de contrepartie « réalité ». La pensée morale qui est issue des cadres est donc ontologique, d'après lui, dans la mesure où tout ce qui relève des activités, des intuitions, des perceptions ou des jugements sociaux des sujets trouve dans les cadres une réalité à leurs yeux.

empreinte inévitablement de « discriminations qualitatives fortes », puisque l'existence de l'acteur social, de ses jugements et de ses actions — et, par extension, de toute formation sociale — reste inextricablement ancrée dans la subjectivité.

C'est dans cette optique que Taylor adresse une critique importante aux thèses naturalistes et réductrices, comme celles qui dédoublent le sujet en une part rationnelle et universelle et une autre part, culturelle, subjective et localisée. Selon Taylor, les thèses naturalistes et réductrices résistent à l'idée de reconnaître la valeur ontologique des moralités en prétendant « [...] que le problème du sens est une pseudo-question et [en soutenant] que les divers cadres à l'intérieur desquels ils (les sujets) trouvent une réponse sont de pures fictions » (Ibid., p. 35) et aussi en se voulant de « rejeter toute distinction qualitative et d'interpréter tous les objectifs de l'homme sur un même plan, de considérer qu'ils sont susceptibles par conséquent de quantification et de calcul selon une quelconque « unité commune de compte ». » (Ibid., p. 39)

Ainsi, toute perspective projetée sur le monde est une perspective soumise à des cadres quelconques — celle de sujets donc — et ne peut bénéficier du statut de supérieur, de vérité ou d'universel qu'aux yeux du sujet lui-même (c'est-à-dire en tant que « discrimination qualitative forte ») et non en vertu d'une rationalité qui trouverait des fondements transcendants dans la nature des humains ou de la société (une unité commune de compte). C'est dans cette perspective que Taylor n'appuie pas la thèse de la séparation théorique de la moralité et de l'éthique, notamment.

D'abord, cette séparation fait implicitement référence à une part naturelle et rationnelle et à une part culturelle d'un sujet dédoublé qui renverraient respectivement à la justice (universelle) et au bien (culturel). Ensuite, elle n'est pas en mesure de rendre compte de la réalité ontologique des acteurs sociaux : c'est-à-dire que le monde tel qu'envisagé par les sujets n'est pas fictif mais bien ontologique puisqu'il est le seul véhicule d'interprétation du monde possible. Dans cette perspective, la dichotomie entre éthique et moralité est soumise elle-même à des impératifs de subjectivité issue de cadres et constitue une « discrimination qualitative forte » projetée sur le monde. Pour Taylor, la question des cadres ne peut être évitée et on ne peut même pas supposer un sujet qui en serait « libéré ».

> En outre, cela (les cadres) ne s'entend pas seulement comme une contingence de la psychologie humaine, qui pourrait s'avérer, un jour, sans pertinence pour un individu exceptionnel ou pour un homme d'un type nouveau, surhomme de l'objectivation désengagée. Je soutiens au contraire que vivre à l'intérieur de tels horizons fortement déterminés constitue une des caractéristiques de l'agent humain, que transgresser ces limites reviendrait à transgresser les limites de ce que nous reconnaissons comme personne humaine intégrale, c'est-à-dire non aliénée (Ibid., p. 45).

Or, le fait de supposer que les cadres (ce qui fait de l'humain un sujet donc) sont inaliénables de l'existence humaine ne fait-elle pas de la question traditionnelle de l'aliénation une interrogation inutile ? Dans la conception moderne et humaniste de l'individu rationnel et autonome, c'est-à-dire non aliéné, c'est justement de l'émancipation des cadres dont il est question par la mise de l'avant d'une conception d'un « surhomme de l'objectivation désengagée »; une conception, de surcroît, à tendance naturalisante. Donc, dans l'optique tayiorienne, l'aliénation de l'individu vis-à-vis une certaine nature qui lui serait immanente est renvoyée au rancart, confinant la condition de

l'individu à celle d'un sujet de la socialisation « toujours et en tous lieux ». L'aliénation ne peut plus prendre dès lors de sens que dans la mesure où nous parlons de conditions matérielles de domination ou d'injustice[7] qu'il revient toutefois au sujet de reconnaître comme telles[8]. C'est là que le troisième type de sujet que nous avions proposé devient pertinent pour l'analyse sociologique : le sujet du pouvoir.

Ainsi pour Taylor, une vision naturaliste qui tenterait de réduire l'humanité à une « unité commune de compte », comme l'autonomie en vertu d'un entendement « rationnel » intrinsèque aux individus, verse dans ce qu'elle dénonce — l'existence des cadres — en fondant des « discriminations qualitatives fortes » à partir de la validité prétendument scientifique de ses postulats. C'est d'ailleurs dans leur caractère scientifique que se situe toute la conviction des thèses selon lesquelles les cadres sont de pures fictions. En ce sens, les « discriminations qualitatives fortes » des tenants de cette école ne le sont pas puisqu'elles bénéficient du statut de vérité. Taylor fait une critique épistémologique de ce point de vue en observant des critères de validité de la connaissance scientifique pour les mesurer aux impératifs de l'appréhension du moi.

Selon lui, (1) on ne peut observer un moi « dans l'absolu », en vertu de

7. Il y a chez Marx un empiétement entre ces « deux types d'aliénation ». L'aliénation des prolétaires de leur travail est une domination matérielle : dépossession du fruit du travail. Mais elle est aussi une contrainte à un état de justice qui trouverait ses fondements dans la positivité de la raison.

8. Le sujet, en effet, doit être celui qui reconnaît sa propre domination (son assujettissement par le pouvoir) si l'on veut s'affranchir de l'idée d'une aliénation naturalisante. Définir objectivement ce qu'est une situtation de domination revient à transporter sur les conduites et les situations des jugements et pourrait constituer en soi une domination (pensons à l'Inquisition). D'un côté plus pragmatique, cela ne devrait pas exclure l'interrogation sur ce qui semble inacceptable...

critères d'objectivité indépendants de la signification que le moi revêt pour un sujet; conséquemment, (2) le moi ne peut rester indépendant d'une description ou d'une interprétation subjective (Ibid., p. 54). D'autre part, (3) on ne peut donner une description parfaitement explicite des interprétations du moi (Ibid., p. 55) — ni même d'un objet de science — en vertu des limites sémantiques du langage. De même qu'il n'est pas possible de (4) comprendre le moi sans référence à son environnement (nous en avons longuement traité avec la question des cadres, de la socialisation et de la subjectivation). Si certains courants à tendance scientificisante ont cherché ou cherchent à mettre au point une moralité qui n'est pas subjective mais objective — ce qui revient à faire de la moralité substantive une pure illusion —, selon Ross Poole, ils faillissent à la tâche.

> Where science provides the standard of objective knowledge, morality can only be a self-deluding subjectivity. But there it (morality) has had its revenge : the pretensions of science to secure objective truth are themselves exposed to reveal the same subjectivity as had been revealed in morality and religion (Poole, 1991, p. 69).

Pour Taylor, les obstacles à l'appréhension objective du moi (dans le cas du sujet moderne), relèvent de l'incontournable modalité de sujet sociologique des acteurs sociaux. En dépit de types-idéaux que l'on peut construire et mettre de l'avant afin de comprendre sociologiquement l'acteur social, en dernière analyse, sa « vraie réalité » se constitue à travers une subjectivité qui lui est propre. Pour l'acteur social moderne, cela est d'autant plus vrai que la typification idéale de l'action sociale se bute à des différenciations individuelles multiples et justement, à l'avènement du moi.

> Ce que je suis en tant que moi, mon identité, se définit essentiellement par la manière dont les choses ont une signification pour moi. [...] ces choses ont du sens pour moi, et le problème de mon identité prend

> forme seulement grâce à un langage d'interprétation que j'ai fini par accepter comme formulation valable de ces problèmes (Taylor, 1998, p. 54).

C'est surtout en vertu de cette problématique importante, qui pourrait nous faire sombrer dans un relativisme futile, que notre troisième partie du premier chapitre a voulu mettre en garde le lecteur. Sans que nous ayons la prétention de saisir le monde moderne, mais surtout contemporain, dans sa vérité totale, nous nous proposons, comme c'est l'objet de la sociologie de le faire, de former des types-idéaux d'une socialité que l'on pourrait généraliser comme tendance dominante du monde moderne et contemporain. Le monde moderne fera l'objet de notre troisième chapitre, alors que la société contemporaine sera explorée aux chapitres quatre et cinq.

Parmi les conséquences les plus importantes et les plus discutées du moi moderne, est celle de l'individualisme — doctrinaire, comme sociologique. Le sujet moderne humaniste a contribué à façonner un acteur social qui se comprend comme unique, parmi d'autres individus uniques, et qui exerce une réfléxivité sur sa condition : il vise notamment l'autonomie par cette réflexivité. C'est à travers l'individualisme que nous voulons continuer notre cheminement, maintenant que nous avons bien établi les bases de notre analyse.

Nous avons circonscrit dans ce chapitre les catégories morales qui caractérisaient la subjectivité moderne, le respect d'autrui, la vie bonne, et la dignité et nous avons cherché à révéler quels étaient les liens qui les unissaient et comment ils se légitimaient les uns et les autres. En dernière partie, notre

objet a été de scruter les dessous de la subjectivité, et de la subjectivité moderne plus spécifiquement. Ce que nous avons appelé l'incarnation subjective tend vers une critique à l'égard de « l'aliénation objective ». Dans nos vues, l'aliénation ne peut être comprise que dans le cadre de ce que nous avons appelé au chapitre précédent le « sujet du pouvoir », soit d'une reconnaissance subjective de quelque chose « d'aliénant ». Nous avons aussi postulé que le visage de l'autonomie, une valeur profondément ancrée dans la subjectivité moderne et humaniste, ne pouvait être « objectif » puisque l'activité social ne peut jamais échapper à la subjectivité.

Cela n'empêche en rien toutefois le fait que l'individu moderne ne peut être compris à l'extérieur d'un certain projet philosophique et juridico-politique d'une collectivité unie et ordonnée par la morale, ce sera notre dessein de le démontrer au chapitre troisième. Ainsi l'individualisme aux XVIIIe et XIXe siècles relève d'une conception morale partagée par les citoyens en vue d'une meilleure citoyenneté, de plus d'autonomie, d'une plus grande justice, d'une solidarité sociale et d'une vision de progrès. La moralité moderne n'est pas complètement « atomisée » et une conception forte du collectif demeure au centre de la socialité. Mais la légitimation de l'ordre social dans l'arène de la subjectivité humaniste moderne doit provenir « du bas » parce que le sujet humaniste est intrinsèquement lié à la reconnaissance subjective de son autonomie. Nous croyons donc que si l'organisation politique moderne est morale c'est qu'elle répond aux exigences subjectives de la population dans des conditions historiques bien précises que nous tâcherons d'éclairer dès le chapitre troisième.

Cela nous amènera à constater que la société contemporaine est le terrain d'un individualisme qui se consolide sur les bases de la performativité, plutôt que de la morale, cela, en fonction du type de rationalisation propre aux sociétés modernes. Autrement dit, que les conditions objectives de la société contemporaines sont en grande partie le fruit de la rationalisation des visées subjectives des acteurs sociaux contemporains. Nous élaborerons cette thèse plus spécifique à partir des développements du quatrième chapitre qui porteront sur la rationalisation.

Chapitre 3 : L'individualisme moral des XVIIIe et XIXe siècles

> *Dans le schéma conceptuel qui s'est imposé au XIXe siècle [...] une valeur était reconnue comme bonne, et l'objectif de l'intellectuel était de faire en sorte que cet universel devienne force de loi. C'est cela la perspective politico-morale.*
> *Michel Maffesoli, Le temps des tribus, p. 233*

Alors que l'humanisme allait déposer chez les acteurs sociaux le germe de l'individualisme, le moi moderne, il n'en demeure pas moins que la modernité, dans sa conception de l'individu autonome et rationnel, ne plaçait pas l'individu dans un vide. Autrement dit, la modernité institue de toute évidence une rupture d'avec les grands ensembles divins et cosmologiques mais ne se dissocie pas elle-même de l'idée d'une unité fondée toutefois sur la transparence et la raison. De partie du Tout, l'individu devenait plutôt membre et acteur rationnel du Tout. Les théories politiques modernes en témoignent; de Rousseau, en passant par Kant et Durkheim, jusqu'à Habermas[9].

L'articulation de la solidarité avec l'individualisme (ou de la liberté individuelle et de l'autonomie) n'est pas chose impossible dans le registre de la pensée politique moderne. Au contraire, la solidarité permet la liberté individuelle dans la mesure où la raison (les théoriciens du contrat et Kant), la société (Durkheim) ou la communication (Habermas) sont capables de fournir les critères vrais et justes de la solidarité, garants de l'autonomie du sujet. La solidarité est donc plus qu'accessible à l'individu, elle y est ancrée : il est doté du bagage nécessaire pour l'assimiler, mais c'est un bagage toutefois qui repose sur une capacité de son entendement (les théoriciens du contrat et Kant), du fait de

9. Le cas de Habermas s'avère très singulier. Par le fait d'abord, qu'il se situe à l'époque contemporaine et que l'on peut le rattacher sans l'ombre d'un doute, de l'aveu même de ses allégeances, à l'humanisme moderne. Mais c'est chez Habermas une tentative fort intéressante de rallier certains faits sur la subjectivité contemporaine avec les développements de la pensée morale moderne (ce que nous nous proposerons d'observer ici) qui constitue le plus grand intérêt pour nous.

son existence proprement sociale (Durkheim), ou des deux (Habermas). Dans cette perspective, l'individu moderne n'est pas largué dans le vide; il est plutôt institué et envisagé dans un nouvel ordre de l'ensemble social. L'individualisme, de ce point de vue, répond à des critères de moralité.

Nous voulons retracer cette perspective moderne à travers deux paradigmes. Le premier à une teneur politique et cherche à promouvoir la liberté individuelle en la faisant reposer sur une entente entre des individus naturellement libres et rationnels et ayant des droits naturels. La moralité trouve donc son compte dans le caractère rationnel de l'individu libre. Nous comprendrons que dans ce modèle, ce n'est pas l'individualisme qui est problématisé (l'emploi du mot date du XIXe siècle), mais bien la liberté du sujet humaniste dans le cadre d'une réflexion sur la solidarité (morale) d'individus libres, égaux et rationnels. Enfin, le deuxième paradigme que nous proposons de regarder se détache de l'autre du fait que la raison n'est pas en son centre. Il s'agit d'une sociologisation du caractère moral de l'individualisme et on peut attribuer sa différence profonde du « déterminisme de la raison » à l'émergence des sciences humaines.

Ainsi, le premier paradigme peut être associé à des penseurs qui précèdent l'avènement des sciences humaines, aux heures de gloire des Lumières. Leurs grands représentants sont les théoriciens du contrat et Emmanuel Kant. Le second paradigme quant à lui, a son plus remarquable représentant en Émile Durkheim. Ils sont différents, certes, mais ils contribuent tous à renforcer l'aspect moral de l'individualisme moderne.

3. 1. La liberté individuelle et l'humanisme comme fondements moraux

Il y a des différences entre tous les théoriciens que l'on dit être « du contrat ». Or, ils ont en commun le fait qu'ils entourent de près, les ayant inspirés, les premiers balbutiemments de l'institutionnalisation de l'humanisme; soit les sociétés issues de révolutions bourgeoises. On ne peut toutefois pas placer à tous les égards Rousseau, Hobbes, Locke et Beccaria (pour ne nommer que certains des principaux) dans le même panier. On peut toutefois les rapprocher, de manière très générale, sur le thème de ce que nous avons appelé l'individualisme moral. Ce rapprochement est d'abord possible du fait que chacun d'entre eux cherche à faire reposer leur théorie politique sur des conjectures anthropologiques dont l'un des effets est de poser les exigences morales nécessaires pour que l'homme puisse vivre le plus librement possible en société, la clé se trouvant dans l'exercice du politique selon des motivations rationnelles qui trouvent leurs sources dans l'individu. Des exigences morales qui tournent autour de l'institutionalisation par les individus de l'État, organe politique à qui l'on cède certains droits (ceux de la répression surtout) pour optimaliser la possibilité d'être libre dans un état social inéluctable et permettre la sécurité et la liberté individuelle.

L'association politique, dans sa version hobbesienne, exige la sublimation rationnelle et stratégique des pulsions naturelles du primitif en nous et ce, par la voie de la rationalité stratégique qui mène à une organisation politique solidaire et morale. La version rousseauiste quant à elle est moins « contre-nature », dans la mesure où la volonté générale a des fondements rationnels émancipateurs plutôt que stratégiques. Les deux façons d'envisager

le contrat font néanmoins référence au dépassement d'un ordre primitif pour le substituer à un ordre civilisé et moral. Dans le *Soupçon gothique* (1999), Valérie de Courville Nicol pose le dédoublement d'un sujet qui se gouverne de l'intérieur par la peur de sa propre nature barbare comme une condition nécessaire de la mise en place d'un ordre politique contractuel permettant à la liberté individuelle et à l'égalité d'être mises à jour. À la nature égoïste et violente (ou chez Rousseau, naïve et peureuse) de l'homme pré-politique s'élève donc rationnellement ou stratégiquement l'individu moral.

La solidarité politique chez les contractualistes propose un État dont la légitimité n'est ni divine ni naturelle mais rationnelle, en vertu de l'entendement stratégique et/ou rationnel d'individus désirant optimaliser la possibilité de mettre à jour leur liberté individuelle dans un climat de sécurité. Les conjectures de l'état de nature ou de guerre, chez les contractualistes, constituent des mises en garde à l'égard du peuple voulant s'émanciper du pouvoir autoritaire et terrorisant de la monarchie. Pour entrer dans une ère rationnelle de justice, d'égalité et de liberté — pour s'émanciper de sa condition de sujet du pouvoir et devenir un citoyen-individu libre et autonome — il ne suffit pas de mettre à terre l'ancien régime. Pour être un individu civilisé, l'individu révolutionnaire de l'Europe doit être un individu moral qui sait se gouverner pour ses propres bienfaits. Bienfaits individuels qui renvoient, dans la logique du premier axe de la pensée morale moderne vu au chapitre second, aux bienfaits d'autrui.

> D'être civilisé, c'est donc d'être à la fois sujet qui gouverne et objet gouverné. L'individu, être humain, est à la fois le sujet qui connaît la chose [la nature de l'humain[10]] et cette chose elle-même, le sujet qui maîtrise et l'objet sujet à la maîtrise. La rationalisation de la conduite

10. Notre ajoût.

> politique, ce qui la rend d'ailleurs pensable, aura comme effet de déplacer le site du pouvoir pour qu'il soit intérieur à lui-même. Le sujet moderne en sera le producteur et le produit, un produit de lui-même (de Courville Nicol, 1999, p. 169).

Ainsi l'État, garant de la liberté et de la sécurité individuelle, de même que les individus qui en sont tributaires, après l'avoir institué rationnellement et librement, reposent sur certaines exigences morales qui permettent à l'autonomie individuelle d'être effective. Fût-elle plus conservatrice, comme chez Hobbes, ou plutôt révolutionnaire comme chez Rousseau, il n'en demeure pas moins que les théories contractualistes mettent au centre de leur analyse une détermination rationnelle du politique qui puise ses sources dans la liberté et l'autonomie individuelle humanistes et le caractère moral de la civilité. Il est aussi important de retenir que c'est dans l'individu et sa raison que réside la possibilité du contrat et que c'est lui qui donne naissance à l'État.

Kant poursuivra dans la lignée des penseurs du droit naturel et du contrat social. Il y apportera toutefois des changements importants. De façon générale, on peut dire qu'une dimension plutôt métaphysique et philosophique de l'individualisme moral et des droits naturels est proposée par Kant à la place d'une dimension plus politique et utilitariste comme chez les contractualistes. À la notion du droit est ajoutée, par Kant, celles de la vertu et du devoir. Il renforce du coup la conception de l'individualisme moral qui situe la moralité dans l'individu, plutôt que dans le politique ou le droit — chose qui est tout de même déjà perceptible chez les contractualistes qui posent comme condition de la sécurité et de la paix le gouvernement de soi par soi avant le gouvernement politique.

> Faut-il [...] rappeler à quel point à été forte à la fin du XVIIIe siècle la contribution de Kant à la promotion d'un individualisme rationnel et humaniste. Non seulement dans la *Critique de la raison pratique* il pose l'homme en sujet radicalement rendu indépendant par sa raison propre, source de « l'autonomie de la volonté » qui l'élève individuellement à la dignité de « fin en soi » dont personne ne peut disposer comme « moyen » — mais il développe dans la *Doctrine du droit* (1790) des thèses légitimant l'interprétation individualiste du droit naturel. (Laurent, 1993, p. 44).

Chez les contractualistes donc, l'autonomie et l'individualisme — le gouvernement de soi-même par soi-même et la liberté individuelle — sont garantis par la raison dans le cadre d'un exercice politique visant à protéger les individus (les droits naturels) et à les émanciper (du pouvoir arbitraire et de leur nature barbare). Chez Kant, les mêmes thèmes sont centraux dans son impératif catégorique qui inspire toute sa conception du droit mais qui indique aussi le lieu transcendental de la moralité humaine (« Agis selon une maxime, qui puisse valoir en même temps comme une loi universelle. »). L'autonomie et la liberté sont accessibles chez Kant par la raison, mais alors que les contractualistes les font reposer sur une exigence politique et utilitariste, Kant les fait reposer sur un discours monologique de l'individu rationnel et moral ayant en arrière-pensée son semblable. Dans la même lignée que le droit naturel des contractualistes donc, la doctrine kantienne a cela de particulier qu'elle situe les fondements du droit naturel encore plus définitivement dans l'entendement des individus.

> La théorie kantienne du droit se veut purement rationnelle, « métaphysique », édifiée toute a priori : alors que l'École du Droit Naturel, quelle pût être sa tendance au rationalisame, cherchait encore un point d'appui pour ses constructions dans un reliquat d'expérience, une expérience d'ailleurs réduite à l'observation prétendue de la « nature de l'homme » isolé, abstrait de son contexte social (Villey, 1971, p. 10).

Ce qu'il importe de retenir dans les perspectives des contractualistes et de Kant est d'abord le fait qu'elles fondent la liberté individuelle sur le caractère moral de l'entendement rationnel de l'humain. Mais c'est aussi le fait que ces idées du XVIIIe siècle qui ont précédé ou suivi de peu la Révolution française ont contribué à rendre à l'individualisme naissant sa dimension morale dans la perception de la population. L'individualisme moral s'oppose aux régimes tyraniques et est une condition, en vertu de son caractère rationnel, de l'autonomie et de la liberté individuelle. C'est ainsi que le citoyen des nouveaux États bourgeois vit son individualisme. C'est dans le partage commun de certaines valeurs morales de solidarité qu'il peut se différencier des autres et forger son identité personnelle, comme il était nécessaire pour les citoyens américains de Tocqueville d'être placés sur des bases communes au niveau des valeurs collectives (le puritanisme et l'*entrepreneurship*) pour devenir individualistes (un individualisme donc, qui n'était pas sans dimension morale).

Qui plus est, l'individualisme rationnel caractéristique des États bourgeois, démocratiques et libéraux devient, dans l'absolu, un mode supérieur d'existence, favorisant la liberté et l'autonomie, fût-il le seul moyen de contrer la bête en nous ou tout simplement, fût-il le fait de la raison.

> Le XVIIIe siècle voit d'autant plus l'Europe occidentale (sauf la péninsule ibérique) s'organiser en « société des individus » (N. Elias) que les interactions du changement culturel et de l'intériorisation psychique étendent les effets du procès d'individualisation au cœur même de l'affectivité et au niveau de la perception que chacun a de lui-même et des autres. La répression des pulsions se renforce tout en se diffusant au-delà de ses premières minorités concernées. Elle provient désormais moins du contrôle social extérieur que d'un contrôle de soi plus rigoureux, issu de la raison personnelle et qui tend à devenir la nouvelle norme auto-déterminante du comportement (*inner-directed*, selon

> l'expression du sociologue David Riesman). L'individualisme se forge et s'approfondit dans ce sentiment d'exister au-dedans et à distance de soi mais aussi en se différenciant d'avantage des autres, en s'affirmant dans une identité singulière (Laurent, 1993, p. 39).

Suivra donc au XIXe siècle une montée fulgurante de l'individualisme bourgeois et libéral qui permettra au capitalisme de se consolider et à la Révolution industrielle de voir le jour. Les craintes du XVIIIe siècle sur le pouvoir arbitraire s'estompent en même temps que le pouvoir séculier prend de plus en plus de place. Mais l'individualisme est l'objet de nouvelles réflexions parce que si est célébré l'individualisme humaniste théorique qui émerge durant les Lumières et qui inspire la solidarité dans les élans révolutionnaires et réformistes du XVIIIe siècle, l'effectivité de nouvelles éthiques individualistes est tout de même troublante.

> Face à cette révolution existencielle des modes de vie courants et à la crise de décomposition/recomposition dans laquelle est engagée la société européenne, tous les penseurs prennent conscience — au-delà des évaluations divergentes qu'ils en font — de l'irruption triomphante du nouveau paradigme individualiste et du conflit qui l'oppose à l'ancien (holiste) dans la représentation de la relation entre l'individu et la société (Ibid., p. 47).

3. 2. L'individualisme entre la modernité et la tradition : l'approche scientifique

C'est dans la montée de la sociologie et des sciences humaines en général que les réponses les plus importantes seront apportées au questionnement du rapport entre l'individu et la société. Déjà, à la fin du XVIIIe siècle Adam Smith apportait des arguments pour l'individualisme qui rompent avec le rationalisme transcendant des contractualistes et de Kant pour situer l'impératif de l'individualisme dans des phénomènes matériels et empiriques. Les

individus ne pouvant dépasser leur nature égoïste, c'est dans la société, économiquement naturalisée, qu'il faut trouver la légitimité de l'individualisme. L'individualisme, dans cette perspective, est donc toujours moral mais la cause est socio-économique et non transcendentale.

L'émergence d'explications scientifiques de la légitimité morale de l'individualisme est peut-être l'effet du climat social du XIXe siècle sur la question de l'individualisme. À cette époque la critique du rationalisme était importante et la quête d'un savoir plus objectif et empirique se faisait sentir. Mais aussi, la nostalgie et les craintes de certains par rapport à un désordre et une anarchie entraînés par de grands bouleversements sociaux (notamment, les révolutions bourgeoises et les résistances ouvrières qui s'en suivent), à un moment où l'individualisme et la démocratie s'incrustrent de plus en plus solidement dans les mœurs, a certainement suscité la nécessité d'explications scientifiques axées sur le progrès pour légitimer un nouveau rapport entre individu et société.

Mais même ces explications « objectives » subiront l'effet des tensions entre le libéralisme et l'individualisme d'un côté et l'ordre et la tradition de l'autre. C'est du moins ce que nous pousse à conclure l'analyse des conflits entre Herbert Spencer et Auguste Comte, tous deux positivistes et organicistes mais dans des contextes bien différents. Ou encore l'équilibre très subtile qu'Émile Durkheim tente de mettre au point entre le pouvoir quasi-absolue masqué de l'État et la liberté et l'autonomie individuelles. Une chose toutefois est commune dans toutes ces rationalisations du nouvel ordre à venir. Durkheim, nous le verrons, saura le remarquer explicitement.

L'individualisme semble être reconnu comme un fait imposant qu'il est nécessaire de légitimer au sein d'un ordre moral nouveau.

L'émergence et la montée de l'individualisme ne s'est donc pas faite sans crainte de voir ses excès éventuels décomposer la société et abolir l'ordre. C'est pourquoi des écrivains européens populaires comme Horace Walpole, après la révolution anglaise, Ann Radcliffe et Matthew « Monk » Lewis tard au XVIIIe siècle et Mary Shelley au XIXe siècle, feront part au public de ces craintes. Leur littérature gothique — un genre littéraire très populaire qui reflétait les peurs de plusieurs sur l'éventualité de la déchéance d'un individualisme sans régulation morale ou d'une société purgée de toutes traditions (chez Walpole et Shelley plus particulièrement) — fera fureur à cette époque et aura même des visées éducatives (de Courville Nicol, 2000).

Cette conjonction problématique entre les thèmes de l'individualisme, du libéralisme, du capitalisme et de la tradition n'est pas non plus étrangère au cas des États-Unis, même si là-bas l'individualisme n'y a jamais trouvé d'opposition sérieuse. Au XIXe siècle, comme Alexis de Tocqueville le rapporte si bien, les émigrants puritains anglais des nouvelles colonies cohabitent dans un environnement individualiste d'entrepreneurs mais en vertu d'une stricte observation des lois religieuses qui les unit. Toutefois, à la différence de l'Europe — et Tocqueville a su le prédire avec perspicacité — l'individualisme américain tardera beaucoup moins à s'implanter de façon incontestablement hégémonique. L'histoire jeune et « individualiste dès ses débuts » des colonies facilitera le passage d'un individualisme moral à un individualisme rationalisé, thèse qui sera l'objet des prochains chapitre.

Il est historiquement probable que le fait du développement des sciences sociales en Europe plutôt qu'en Amérique soit lié aux tensions d'un continent à cheval sur des régimes différents mais présents avec autant d'intensité. Même dans le développement des sciences sociales, les contextes nationaux et les idéologies des penseurs font foi de cette tension, ce qui permet de comprendre que les sciences sociales, et la sociologie tout particulièrement, se préoccupent de l'organisation du rapport entre individu et société. L'organicisme du milieu du XIXe siècle en France et en Angleterre, ceux de Comte et de Spencer donc, auront des portées fort différentes. Alors que Comte écrit en France, où la révolution bourgeoise date de moins d'un siècle et où elle a subit des rebondissements, Spencer écrit dans une Angleterre qui a fait sa révolution il y a déjà près de deux cent ans. Une Angleterre qui restera le géant industriel, capitaliste et bourgeois du monde jusqu'au début du XXe siècle, lorsque les Américains prendront le dessus.

Roberto Miguelez, dans *L'émergence de la sociologie* (1993), fait ressortir très clairement la différence entre les organicismes de Comte et de Spencer. Le premier reflète l'état d'une société qui n'a pas encore su intégrer les tensions entre la bourgeoisie et les masses ouvrières. Mais aussi, supposons-nous, qui n'a pas encore résolu sa rupture avec un régime traditionnel, ce qui expliquerait une insistance sur la dimension de l'ordre et de la force comme deux facteurs nécessaires à l'équilibre social[11]. Le second organicisme quant à lui, est teinté de démocratisme et d'individualisme, ainsi que d'une doctrine de non-

11. Nous connaissons les événements post-révolutionnaires reliés à l'Empire en France qui n'ont pas optimisé les chances de l'institution de la démocratie et par conséquent, du capitalisme libre et de l'individualisme. Quoique la monarchie en Angleterre, jusqu'à ce jour, reste en place, nul ne doute de son importance politique et économique décroissante tout comme de sa collaboration, depuis le XVIIIe siècle, avec le pouvoir séculier et l'entreprise capitaliste et démocratique de l'Angleterre.

interventionnisme étatique qui le rapproche plus de l'économie politique de Smith, comme l'avance Miguelez.

> En d'autres termes, ils (l'individualisme et le démocratisme de Spencer) supposent que, par certains côtés, le « superorganisme social » ne ressemble pas à un organisme biologique. Nous connaissons ses deux différences. D'une part, dans les sociétés, les parties ne sont pas, selon Spencer, dépendants du tout comme c'est le cas dans un organisme biologique mais au contraire, elles sont « libres, discrètes », et la société est à leur service. Puisque les individus constituent les parties d'une société, faire des individus les éléments libres et premiers de la sociologie revient à assumer une position idéologique individualiste. D'autre part, dans les sociétés la conscience ne se trouve pas, selon Spencer, concentrée dans une partie de l'organisme comme dans le cas des organismes biologiques mais, au contraire, elle se trouve dispersée dans tous ses membres, c'est-à-dire dans tous les individus. Il s'agit d'un postulat démocratique puisque tous les individus sont censés être égaux au niveau fondamental de leur conscience (Miguelez, 1993, p. 75).

Nous croyons qu'il est d'autant plus intéressant de faire une lecture de Durkheim et de son individualisme moral à la lumière des déchirements idéologiques qui sévissent en Europe sur le rapport entre individu et société à l'aube de l'hégémonie montante de la bourgeoisie, du capitalisme et de la démocratie. Durkheim voudra s'approprier la méthode scientifique pour tenter de résoudre les conflits qui opposaient les vues théoriques des transcendentaux et des libéraux, tout en affirmant le fait de l'individualisme en rassurant les conservateurs qui y voyaient le pire; soit en ne dépossédant pas l'individu de ses capacités réflexives, autonomes et libres, tout en leur posant des limites naturelles — naturelles parce que sociales.

3. 3. La société et le progrès comme condition morale de l'individualisme

> *Malgré la lettre de notre code moral, les vieilles habitudes survivent plus que nous ne le voudrions. C'est que, pour instituer une morale individualiste, il ne suffit pas de l'affirmer, de la traduire en beaux systèmes, il faut que la société soit arrangée de manière à rendre possible et durable cette constitution. Autrement elle reste à l'état diffus et doctrinaire.*
>
> *Émile Durkheim, Leçons de sociologie, p. 95*

Les *Leçons de sociologie : Physique des mœurs et du droit* (1950) d'Émile Durkheim sont mises au point à un moment où le pouvoir bourgeois, le règne capitaliste et le régime démocratique ne peuvent qu'être reconnus comme la voie à venir des sociétés occidentales modernes. Il en va de même de l'individualisme : « On ne peut pas faire que l'individu ne soit pas devenu ce qu'il est, c'est-à-dire un foyer autonome d'activité, un système imposant de forces personnelles dont l'énergie ne peut pas plus être détruite que celle des forces cosmiques. » (Durkheim, 1950, p. 93) Cela n'échappe pas à l'œil du sociologue français. Il reste qu'il n'est pas satisfait des réponses apportées par ceux qu'il appelle les individualistes — « Spencer et les économistes d'une part, par Kant, Rousseau et les spiritualistes de l'autre » (Ibid., p. 88) — quant à la nature véritable du rapport entre société et individu dans les sociétés modernes.

Pour Durkheim, l'individualisme est un fait inéluctable de l'évolution de la société et constitue donc un fait moral dans les sociétés modernes de la fin du XIXe siècle. Toutefois, les contractualistes et Kant situent le lieu de l'autonomie et de la liberté de l'individu dans les sphères transcendantes de sa raison. En Angleterre, Smith et Spencer élèvent l'individu au niveau d'unité centrale et finale d'analyse pour contribuer à l'idéologie libérale de l'individu

égoïste motivé en dernière analyse par la réalisation de désirs qui ne doivent pas être frustrés par l'intervention étatique. Durkheim condamnera les contractualistes, Kant et les libéraux comme Smith et Spencer[12] pour leur réduction de la « morale civique » à l'individu.

Fût-il transcendentalement rationnel ou intrinsèquement égoïste, l'individu comme facteur déterminant de la morale et de l'organisation sociale est rejeté par Durkheim. Sans partager l'apparent pessimisme de Comte sur l'autonomie individuelle, pour qui la force (et la religion positiviste) est nécessaire afin de maintenir l'ordre, il diminuera quand-même l'importance de l'individu comme facteur déterminant de l'ordre social individualiste moderne pour en investir en grande partie la cause dans la société. Mais l'État sera moins l'organe négatif de la force, comme chez Comte, ou de la surveillance, comme chez les libéraux et les contractualistes, que celui de la conscience collective, positive, partagée par des individus autonomes et libres parce qu'ayant intégré les exigences morales de la solidarité sociale.

L'individualisme moral dans la perspective durkheimienne n'a aucunement sa source dans l'individu mais bien dans la société. Si individualisme il y a, selon lui, c'est en vertu des conditions du progrès posées par la société elle-même et dont l'État se fait le véhicule idéationnel. «Et pourtant, ils [les individus] ne sont pas, comme le veut l'école individualiste utilitaire, ou l'école kantienne, des touts qui se suffisent à eux-mêmes, et que

12. Chez Smith et Spencer il est toutefois important de préciser que la société est quand-même thématisée de façon importante et que l'individu n'est pas la seule unité d'analyse valide. Au contraire, la société est le lieu privilégié de l'activité individuelle et elle a un intérêt productif qui lui est inhérent et qui rejoint — plutôt que de dépasser, comme chez Durkheim — les intérêts individuels.

l'État doit se borner à respecter, puisque c'est par l'État et par lui seul qu'ils existent moralement. » (Ibid. p. 99) Durkheim a voulu théoriquement équilibrer le déchirement de l'individualisme entre la modernité et la tradition mais aussi légitimer, paradoxalement, les notions de liberté et d'autonomie, au sein de limites bien précises posées comme scientifiquement valides; à savoir le progrès et la supériorité de la société sur ses parties.

L'approche comparative en sociologie, dont Durkheim est l'un des pionniers avec les organicistes, l'amène à tirer des conclusions sur la supériorité (évolutive et non morale) des corps sociaux dont la division du travail est à l'image de l'organisme vivant complexe, c'est-à-dire très différencié et très individualisé, point de vue partagé, *grosso modo*, avec Comte et Spencer. Il est dans la nature d'un organisme vivant qu'une solidarité soit entretenue entre ses parties afin de survivre. C'est par une analogie de l'évolution des sociétés avec l'évolution des organismes vivants que Durkheim arrive au constat positiviste selon lequel les sociétés dont le travail social est très peu divisé — tout comme les organismes vivants primitifs — et dont la production est limitée et moins efficace, sont moins évoluées que les sociétés complexes — les sociétés industrielles et modernes avec une grande division du travail social.

Or, ce fait du progrès des sociétés dont la production est plus efficace parce que plus complexe et plus systématisée ne relève en rien de la morale. La morale, chez Durkheim, est toute relative aux besoins de la société. Elle permet à la société de perdurer et ne constitue pas un fait déterminant mais bien une conséquence de la survivance naturelle de la société. Tout comme l'organisme

vivant, la société est un tout supérieur à la somme de ses parties — ces dernières, en bout de ligne, n'ayant pour objet que de contribuer à la santé du tout. Ainsi l'individualisation croissante des sociétés industrielles et modernes — autant dans les fonctions productives que dans les consciences et les éthiques — constitue selon Durkheim un fait moral parce qu'il répond aux exigences du progrès posées par la société elle-même en tant que cette dernière a une vie en soi, qui est objet d'analyse factuelle et scientifique. Dans ces circonstances, « le domaine de la vie vraiment morale ne commence que là où commence le domaine de la vie collective, ou, en d'autres termes, [...] nous ne sommes des êtres moraux que dans la mesure où nous sommes des êtres sociaux. » (Durkheim, 1963, p. 55)

Durkheim pose donc de nouvelles conditions à l'individualisme moral, le situant non pas dans l'individu en soi, mais dans la solidarité et la coopération entre individus en vertu des exigences intrinsèques de la société et du progrès. Qui plus est, cette coopération et cette solidarité, orchestrée par les règles morales et collectives énoncées par l'État est, d'après le sociologue, le moyen le plus eficace de mettre à jour l'individualisme. Pour que la société ne se transforme pas en des sous-groupes (professionnels, familiaux, religieux, etc.) indépendants de la grande société qui empêcheraient dès lors toute forme de liberté individuelle, l'État doit pénétrer partout.

Mais selon Durkheim l'individu n'est pas que le simple outil de la société, ou de l'État. « Les individus peuvent, sans se contredire, se faire les instruments de l'État, puisque c'est à les réaliser que tend l'action de l'État. » (Durkheim, 1950. p. 104) Si l'organe de la pensée de la société est l'État, sa

nature, nous dit Durkheim, n'est pas sacrée ou transcendante comme elle pu l'être dans les sociétés moins évoluées. Au temps de l'individualisation des pensées, des croyances et des sentiments, l'État intéragit avec la masse des gens au moyen de groupes corporatifs, d'associations profesionnelles et de l'éducation morale. L'État peut ainsi connaître les doléances individuelles pour dicter le bien commun et assurer le progrès de la société (et par conséquent la liberté des individus). La démocratie pour Durkheim ne réside donc pas dans le gouvernement de tous sur les affaires de l'État mais bien sur l'étroite collaboration de l'État avec les individus par la voie des associations professionnelles. L'individualisation s'imposant comme fait moral de la société avec le progrès par la division du travail social, c'est à pouvoir optimiser cette condition de l'individualisme moral que l'État, organe délibératif et éclairé, travaille en fonction des intérêts des individus. La relation de l'État avec les individus est en fait pour Durkheim l'exercice d'une conscience *sui generis*, la conscience collective, essentielle et vitale à la survie et à l'évolution de l'organisme social (Ibid., p. 110 - 118). Mais c'est aussi en vertu de leur autonomie que les individus sont capables, selon Durkheim, de concilier leur individualité avec les intérêts du collectif.

Dans *L'éducation morale* (1963), Durkheim postule que l'autonomie n'est pas quelque chose qui se trouve naturellement chez l'être humain. Kant, dit-il, postule que l'autonomie est source de moralité en vertu du caractère intrinsèquement rationnel de l'humain. Durkheim lui oppose le carctère rationnel mais à la fois sensible et insatiable de l'humain qui le conduit, laissé à lui seul, à être irrationnel et immoral (Durkheim, 1963, p. 95). Selon Durkheim, l'autonomie trouve sa source à l'extérieur de l'humain, soit dans la

science et l'éducation. La sociologie par exemple, est à même de révéler les lois morales de la société de par sa capacité scientifique de saisir les faits sociaux. Par son entendement, l'humain peut s'approprier cette compréhension empirique des lois morales et devenir autonome. Ainsi l'autonomie consiste en le respect et l'obéissance aux lois en vertu de la reconnaissance de leur validité. C'est l'adéquation durkheimienne du caractère moral de ce qui est social.

En ces temps démocratiques donc, les individus de plus en plus libres sont aussi de plus en plus éclairés et quoique différents entre eux et libres d'orienter leurs pensées, leurs croyances et leurs sentiments, ils sont en mesure de percevoir, de comprendre, de vouloir et d'accepter ce qui consitue le bien-fondé de la collectivité et de la solidarité, soit-elle réduite, au temps de l'individualisme, à de moins en moins de choses. C'est dans cette mesure que Durkheim postule l'autonomie de l'individu dans sa relation d'obéissance avec l'organe de la pensée sociale qu'est l'État. On peut comprendre que l'argumentation de Durkheim revient à dire que les individus de l'âge démocratique (ce qu'il entend, lui, par démocratie) sont plus autonomes et plus libres que n'importe quels autres en vertu notamment, de l'avancement des sciences. C'est de cette façon que Durkheim tente de concilier une approche conservatrice de l'individualisme qui ne met pas en péril les concepts de liberté et d'autonomie.

> Être autonome, c'est, pour l'homme, comprendre les nécessités auxquelles il doit se plier et les accepter en connaissance de cause. Nous ne pouvons pas faire que les lois de choses soient autrement qu'elles ne sont; mais nous nous en libérons en les pensant, c'est-à-dire en les faisant nôtres par la pensée. C'est là ce qui fait la supériorité morale de la démocratie. Parce qu'elle est le régime de la réflexion, elle permet au citoyen d'accepter les lois de son pays avec plus d'intelligence, partant

> avec moins de passivité. Parce qu'il y a des communications constantes entre eux et l'État, l'État n'est plus pour les individus comme une force extérieure qui leur imprime une pulsion toute mécanique. Grâce aux échanges constants qui se font entre eux et lui, sa vie se rattache à la leur, comme la leur se rattache à la sienne (Durkheim, 1950, p. 124 - 125).

3. 4. Pourquoi un individualisme « moral » ?

Les XVIIIe et XIXe siècles marquent une effervescence au niveau politique en Europe. Les changements sociaux révolutionnaires, le capitalisme et les idées nouvelles sur la nature autonome et libre de l'être humain, issus de l'humanisme et des Lumières et de penseurs comme Kant et les contractualistes, pressent le besoin de légitimer et de délimiter les frontières de la nature d'un rapport entre individu et société. C'est que l'individualisme semble s'insinuer, malgré toutes oppositions, en vertu des forces et des pouvoirs politiques, structurels, institutionnels et technologiques de la société que développe le capitalisme. Ainsi, le nouveau rapport entre individu et société est à plusieurs égards contenu dans le capitalisme, le libéralisme et la démocratie qui s'imposent de plus en plus comme des faits sociaux plutôt que des idées. L'individualisme qui se dessine de plus en plus comme un fait sociologique donc, doit reposer sur une conception subjectivement reconnue de l'autonomie. Les travaux des penseurs politiques de l'époque s'attardent à cette problématique et ils la thématiseront au sein d'une conception morale de la citoyenneté individualiste et d'une scientificisation de la morale.

Alors que les contractualistes et Kant écrivent au moment où la redéfinition de l'être humain en sujet humaniste atteint son point culminant,

c'est au milieu de la « confusion démocratique »[13] que Durkheim propose ses thèses sur l'individualisme moral. Que les motifs de cette entreprise furent plutôt conservateurs, chez Durkheim, plutôt idéalistes, chez Kant, ou plutôt libéraux, chez Spencer, il n'en demeure pas moins qu'il y a là un témoignage de l'effervescence intellectuelle, politique et populaire qui entoure la question de l'individu et par extension, d'une organisation nouvelle de la société. Bien entendu, il est difficile de détacher cette préoccupation du libéralisme et du capitalisme montant. Une approche matérialiste du problème nous apprend que si discussion sur l'individualisme il y eût, c'est en raison des implications du mode de production capitaliste qui ne cesse de prendre de l'ampleur et de faire ressentir ses effets sociaux, dont l'individualisation. Il n'en demeure pas moins toutefois que les citoyens, produits de ces nouvelles idées et de ces nouveaux rapports sociaux, sont les principaux concernés par la légitimation, surtout dans la mesure où la justification de l'ordre nouveau qui se dessine repose sur leur autonomie et leur réflexivité et doit donc provenir « du bas ».

Nous supposons que si la légitimation de l'individualisme dans les sociétés du XVIIIe et XIXe siècles se fait à travers son caractère moral, c'est peut-être en partie parce que les valeurs qu'apportent avec eux le capitalisme, le libéralisme et la démocratie peuvent encore se mesurer aux anciens régimes. L'ordre social de cette époque, vu qu'il ne repose plus dans les mains de la divinité et de la tradition, entre dans une ère de sécularisation puisque c'est la

13. Tocqueville écrit à propos des révolutions démocratiques dans un État aristocratique, que les conflits entre anciennes classes se prolongent. « Ces passions survivent à la victoire, et l'on peut en suivre la trace au milieu de la confusion démocratique qui lui succède (Tocqueville, 1961, t. 2, p. 147). » Cela expliquerait les oppositions à l'individualisme en Europe dans une perspective tocquevillienne. Dans la même perspective on peut expliquer la presque absence d'oppositions à l'individualisme aux États-Unis : « Le grand avantage des Américains est d'être arrivés à la démocratie sans avoir à souffrir de révolutions démocratiques, et d'être nés égaux au lieu de le devenir (Ibid., p. 147) ».

raison, la science ou la société qui viennent prendre le flambeau de la légitimation. Mais le climat révolutionnaire et parfois instable qui persiste en Europe à cette époque favorise selon nous l'appréhension subjective de l'individualisme sur les bases d'une nouvelle conception moralisante du citoyen dans les limites de grands récits légitimateurs. Nous supposons que si l'individualisme est entouré d'un caractère moral fort, c'est parce que l'approche humaniste de l'être humain représente, au sortir du féodalisme, une source d'émancipation, de valeur morale supérieure, de progrès et de justice qui bénéficie d'un effet d'engouement.

Même aux États-Unis où l'individualisme n'a jamais trouvé de grandes oppositions, son effectivité aux XVIIIe et XIXe siècles repose sur des fondements moraux. Tocqueville a su bien démontrer que l'esprit d'entrepreneur des Américains, leurs principes démocratiques et leur égalité tant aimée trouvaient leur ancrage dans les mœurs, et notamment dans la morale puritaine. Il s'agit d'une thèse que reprendra Max Weber de manière plus systématique dans son œuvre*L'éthique protestante et l'esprit du capitalisme.* Mais d'autres thèses de Weber, notamment celles sur la rationalisation, sont à même de venir appuyer nos propositions. Le fait qu'aux XVIIIe et XIXe siècles l'ordre repose sur des critères de moralité nous permet de supposer que la rationalisation de la sécularisation du politique est encore en développement et que cette dernière trouve un état, non pas d'achèvement, mais de développement sans pareil dans la société contemporaine. D'une source de légitimation extramondaine dans le monde moderne — comme le progrès ou la justice universelle — la source de légitimation de l'individualisme contemporain prend des allures intramondaines — comme le quotidien, le soi, la performance.

L'humanisme et la pensée morale qui s'y rattache fait pénétrer le rapport au monde des sujets dans un moment plus rationalisé. Le moi et sa capacité réflexive transforme les enjeux sociaux qui entourent les sujets et ces derniers cherchent à quérir du sens et une autonomie. Ce sens et cette autonomie, dans les deux siècles précédents, s'articulent autour de conceptions morales partagées et entretenues par ce que Jean-François Lyotard appelle , dans *La condition postmoderne* (1979), des grands récits. Le progrès et la justice par exemple, deviennent des enjeux de solidarité au sein d'une conception morale de l'individualisme et de la citoyenneté. En ce sens le citoyen-individu prend une place au sein d'un ensemble dont la cohérence et l'unité repose sur la morale. La société contemporaine, selon nous, représente une rupture d'avec ce modèle en vertu d'une rationalisation du monde. À défaut de pouvoir fournir une explication moins hypothétique des causes socio-historiques de l'individualisme moral, nous pouvons toutefois avancer que l'individualisme, dans le monde contemporain, prend un visage différent — l'objet d'étude de ce qui va suivre est de tenter de comprendre comment se manifeste ce nouvel individualisme du point de vue du sujet pour proposer une compréhension de la nature de l'organisation du collectif.

Chapitre 4 : La rationalisation moderne : foyer du nouvel individualisme

Dans les chapitres précédents, notre attention a été portée sur l'époque des révolutions bourgeoises et démocratiques car, pour pouvoir accéder à sa compréhension, le monde contemporain doit être envisagé dans sa « modernité », comme dans sa spécificité contemporaine. Nous avons d'abord cherché à faire un portrait général de l'individu moderne en scrutant quelques dimensions importantes de la fondation de sa subjectivité, une subjectivité étroitement liée moralement à la dignité, à la vie bonne et au droit moderne. De façon générale, nous avons cherché à caractériser cette subjectivité « du moi » comme un phénomène individualiste qui mettait au centre des préoccupations subjectives et politiques la question de l'autonomie. Le troisième chapitre a été consacré à faire la lumière sur la nature de l'individualisme aux siècles des révolutions bourgeoises jusqu'à la Révolution industrielle. L'individualisme aux XVIIIe et XIXe siècles, surtout en Europe, nous a laissé voir des contours dont la légitimation était soutenue par des bases morales.

Le chapitre présent cherchera à souligner les conditions socio-historiques d'une nouvelle forme idéale-typique de l'individualisme. Il fera aussi l'objet d'une analyse des cadres subjectifs liés à la spécificité de la rationalisation moderne et ayant contribué à la transformation de l'individualisme. Cela nous permettra de mieux saisir, subséquemment, ce que nous proposerons comme étant la désagrégation du caractère moral de l'individualisme. Toutefois, si nous réfléchissons l'époque contemporaine en fonction de l'époque moderne ce n'est pas dans le but de saisir un mouvement de l'histoire qui serait uniforme dans tous les pays ou qui se serait manifesté de la même façon partout. Notre

objectif n'est pas de constater la trame de la « transformation » de l'individualisme mais la présence de forces socio-historiques spécifiques — la rationalisation du monde moderne — ayant contribués à l'émergence d'un individualisme nouveau. Si l'inspiration du troisième chapitre portait plus spécifiquement sur l'Europe, le reste de notre analyse sera surtout inspiré par les États-Unis et l'Amérique du Nord, dans la mesure où le nouvel individualisme y voit le jour.

4. 1. L'émergence d'une nouvelle socialité : le lieu du nouvel individualisme

C'est après la deuxième guerre mondiale que l'Occident va subir des transformations suffisamment importantes pour qu'elles soient symptomatiques d'un débat sur l'appellation de la société dans la littérature sociologique. C'est que ce moment historique marque le début d'une transformation assez importante de la socialité et de la subjectivité pour qu'on postule l'idée de l'émergence d'une formation sociale de nouvelle nature. Or cette société, aussi nouvelle soit-elle, n'est pas révolutionnaire par rapport au monde moderne. C'est-à-dire qu'elle ne se constitue pas, de façon générale, en une coupure radicale d'avec les structures sociales et le discours légitimateur moderne. Les valeurs humanistes des droits naturels, de la dignité et de la vie bonne — le fondement des sociétés de droit — restent, et se consolident même. La naissance des Nations Unies, les indépendances coloniales, la consolidation de l'État providence, la démocratisation de tout l'Occident dans des sociétés de droit ou l'avènement de la société de consommation sont des phénomènes qui perpétuent, à leurs façons, les valeurs humanistes. Ces phénomène perpétuent ces valeurs mais, comme nous le verrons, ils manifestent leur rationalisation.

Or, c'est l'Amérique qui est le premier foyer de cette nouvelle société.

D'ailleurs, jusqu'après la deuxième guerre mondiale, en Europe, le destin de l'individualisme comme facteur apparemment irréversible de la société et à tout le moins sociologiquement difficile à ne pas constater, était loin d'être garantit. Alain Laurent, dans le quatrième chapitre de son *Histoire de l'individualisme* (1993), nous le fait savoir. Qu'elles furent réactionnaires ou progressistes, les oppositions à l'individualisme faisaient légion. Des oppositions qui, selon Laurent, auront une de leurs plus grandes manifestations avec les fascismes allemand et italien qui voulaient subordonner l'individu à une nation « organique ». Pour Laurent, ce n'est que dans les années 1980 — avec les révélations de dissidents du bloc de l'Est entre autres — que l'individualisme abolira toute opposition intellectuelle substantielle en Europe pour s'installer plus confortablement dans les mœurs et cohabiter avec des structures sociales libérales — des structures déjà relativement individualistes donc[14].

Mais alors qu'en Europe les oppositions persistent, en Amérique du Nord, foyer de la société de consommation nouvelle et avec l'Angleterre, des premiers balbutiements de l'entrepreneurship rationnel, la ferveur individualiste inspirée par des individus comme la romancière Ayn Rand, continue de battre son plein, redouble d'efforts même, allant jusqu'à critiquer

14. Les sociétés individualistes peuvent donner lieu à des paradoxes de ce genre en vertu notamment de leur cadre institutionnel rationnel-légal puissant (nous l'aborderons plus loin), ou tout simplement en vertu du moi qui est toujours inévitablement individualiste. Laurent écrit : « [...] on peut professer une idéologie anti-individualiste virulente... tout en menant une existence fort individualiste (non-conformiste); ou décider individuellement de vivre d'une manière communautaire; ou se trouver amené à vivre une existence de type individualiste subie et non-voulue; ou enfin reproduire les modèles les plus conformistes dans une vie (privée) formellement individuelle. » (Laurent, 1993, p. 8)

les mesures sociales-démocrates du *New Deal* et du *Welfare State*. C'est donc en Amérique du Nord que la nouvelle société de l'après-guerre qui se dessine, que nous appelons contemporaine, trouve ses plus grandes assises historiques.

Si les valeurs inaliénables des sociétés de droit se consolident après la deuxième guerre mondiale, on peut l'expliquer en réaction idéologique, peut-être, aux horreurs du nazisme, cristallisées par certains intellectuels comme les apories de la modernité et perçues, dans l'histoire américaine surtout, comme la preuve de la supériorité morale du modèle libéral de la société. Les « horreurs » totalitaires se poursuivront avec le stalinisme et cela va donner lieu à un renforcement des valeurs démocratiques des sociétés de droit, mais du libéralisme économique aussi, contrepartie économique et politique, juste et égalitaire, du communisme totalitaire. Mais à ces hypothétiques causes « idéologiques » du changement social contemporain, on peut faire concorder des causes économiques et culturelles.

D'autres changements significatifs vont survenir : une époque de prospérité économique pour l'Occident ainsi qu'un boom démographique important vont avoir lieu après la seconde guerre mondiale. Cela permettra à un point culminant du changement social d'après-guerre d'être atteint dans les années soixante : ce sont entre autres, la Révolution tranquille qui s'achève au Québec, mai 1968 en Europe, les mouvements sociaux aux États-Unis, dont ceux des Noirs et des Femmes surtout; des phénomènes qui vont même se répercuter dans les sociétés totalitaires avec les mouvements des étudiants tchèques et chinois. Ce sont des phénomènes qui ne sont pas nécessairement orientés directement vers des revendications individualistes mais qui

réclament certainement une plus grande autonomie, qui contestent la légitimité de la tradition et de l'autorité et qui contribueront en bout de ligne à implanter une atmosphère propice au façonnement d'une société nouvelle.

Nous ne supposons pas que les mouvements contestataires ou de résistance sont nouveaux. Au contraire, ils ne le sont pas. Les mouvements des suffragettes ou des prolétaires, par exemple, ont de beaucoup précédé cette époque. Toutefois, les années de l'après-guerre marquent un point tournant dans la mesure où les mouvements sociaux se transforment en acquis sociaux. Ou du moins s'y rapprochent-ils de plus en plus en ayant le droit légitime de contester et en occupant une tribune de plus en plus importante avec les média de masse. Les forces de la tradition, ou ce qu'il en reste, croulent de plus en plus devant les revendications à la liberté et à l'autonomie de plusieurs mouvements identitaires.

4. 2. Les contours d'une nouvelle subjectivité idéale-typique

Il y a une affinité historique entre les événements du milieu du siècle, une transformation de la subjectivité et, comme nous chercherons à le démontrer, une nouvelle façon subjective d'envisager l'organisation de la socialité. Il y aurait aussi plusieurs façons d'expliquer et de tenter de comprendre les changements de la socialité qui mèneront à la société contemporaine. Et le portrait n'est pas homogène; comme les différences entre l'Europe et l'Amérique que nous avons souligné nous l'apprennent. De même, il existe plusieurs lectures possibles de ce changement de socialité, soit des manifestations particulières qui la caractérisent. Un intérêt particulier pourrait

être centré sur l'économique, le politique ou l'institutionnel, qui sont autant de points de vus qui offrent pour le sociologue des avenues de compréhension particulière de la société contemporaine. Notre intérêt se portera au niveau du sujet, sans négliger toutefois le fait que ce dernier ne peut pas être compris si on l'isole de la totalité de la mouvance sociale contemporaine, soit des dimensions que nous avons mentionnées plus haut. Ce que nous tenterons de souligner ici, quant à la mouvance nouvelle de la société qui éclos entre les années 1950 et 1970, et qui se poursuit dès lors, concerne les nouveaux rapports rationalisés entretenus par les individus à l'égard de l'organisation structurelle et institutionnelle de la société.

Nous poursuivons donc l'objectif de nos chapitres précédents, mais dans l'optique de la société contemporaine cette fois. Nous avancerons la thèse selon laquelle la version moralisante de l'individualisme subit une rationalisation dans le contexte contemporain. Nous voulons aussi proposer que l'individualisme contemporain subit les effets d'une rationalisation qui est intrinsèque à l'émergence du sujet humaniste. Cela nous amènera plus loin à proposer que l'individualisme, dans sa façon d'être vécu par les sujets contemporains, est vidé de tout sens de solidarité et de devoir à l'égard de la société, du Tout social.

Il ne s'agira pas de prétendre que le discours moral — ou politique, ou juridique, ou économique — sur les vertus émancipatrices du libéralisme, de la liberté et de l'indépendance individuelle n'est plus, mais plutôt de dire qu'il s'est rationalisé avec des conséquences particulières sur le rapport entre individu et société. En fait, la persistance des catégories morales du droit

naturel moderne, vues au chapitre précédent, reposent maintenant, nous supposons, dans les « fers » de leur rationalisation. À nos yeux toutefois cela ne constitue pas nécessairement une aliénation ou même un désenchantement du monde comme le laissent supposer parfois les thèses wébériennes ou la sociologie critique post-marxienne des années soixante. La société de consommation, comme l'exprime Herbert Marcuse dans *L'homme unidimensionnel* (1968), peut être appréhendée comme une société dont l'acquisition de biens tournent autour de « faux besoins ». Or, il est selon nous possible de regarder la transformation de la société et de ses phénomènes principaux, dont la consommation, comme une nouvelle appropriation subjective de l'autonomie des individus. Une nouvelle autonomie qui se dessinerait sur les bases de l'ébranlement de la légitimation morale de la modernité et des rationalisations qu'elle portait en elle.

Depuis les thèses de Jean-François Lyotard, avancées dans *La condition postmoderne* (1979), le monde intellectuel en général, sans nécessairement postuler l'existence d'une rupture du tissu social spécifiquement moderne par l'anéantissement de ses grands récits légitimateurs, s'accorde du moins sur une crise qui sévit au niveau de ces derniers. Dans la perspective postmoderne, les absolus qu'ont pu représenter la raison ou la société, pour accéder à des vérités universelles sur la justice et la morale, n'auraient plus la capacité de rassembler tout un chacun, sous fond de vérité et dans des idéaux partagés. Lyotard parle en ce sens d'une délégitimation des grands récits spéculatifs et émancipateurs, issus de la science, et de leur capacité à faire concorder les postulats de vérité et de preuve à la justesse normative et morale. Ce phénomène social peut selon nous faire l'objet d'une analyse en fonction d'un accroissement contemporain

de la rationalisation moderne du monde et d'un nouveau positionnement subjectif qui s'en suit, où le moi prend des points de repères inédits dans l'histoire. De façon générale, le moi contemporain est un sujet de la personnalisation et du sens plus que jamais.

Si par ailleurs, des intellectuels d'allégeance moderne comme Jürgen Habermas et Karl-Otto Apel tentent de réhabiliter l'objectif d'un fondement cognitif de la vérité morale, ce n'est pas sans prendre en considération l'ébranlement des repères traditionnels de la modernité et la transformation de la subjectivité des individus. Le fait que ce soit la communication que Habermas et Apel privilégient comme lieu du consensus et de la vérité morale est significatif de l'importance de plus en plus grande du moi dans l'analyse sociologique. Quoique l'on puisse trouver d'après cette école de pensée, une transcendance qui se situerait à l'extérieur des sujets[15] , le véhicule n'en demeure pas moins l'individu et son moi. On peut donc voir chez les modernes comme chez les postmodernes le constat selon lequel les repères qui ont contribué à fonder notre société et notre subjectivité ont perdu leur portée d'antan. Le rapport entre individu et société s'est transformé. Nous sommes d'avis que les grands récits modernes se butent au nouvel individualisme contemporain et à une logique le sous-tendant qui institue la société comme moyen plutôt que comme finalité, ébranlant dès lors la conception de l'individualisme à l'intérieur d'un cadre moral.

Aujourd'hui, l'individu et la société n'ont pas une existence sociologique

15. Habermas (1996, § III) cherche à se détacher de l'aspect transcendantal de la sémiotique telle qu'Apel l'envisage (1987), c'est-à-dire comme fondation ultime de la raison. Mais il postule tout de même que la disposition de sujets à l'intercompréhension orientée en fonction de prétentions à la validité est « quasi transcendantale » (Habermas, 1992, p. 24). Or ce postulat est équivoque.

auréolée d'une idée ou d'un principe méta-fondateur ou légitimateur. Ni même l'individu, en dépit des impressions que certains pourraient avoir, ne constitue ce principe, cette idée ou cette fin en soi. Comme Christopher Lasch l'a fait remarquer dans *The Culture of Narcissism* (1979), le narcissisme contemporain est bien moins une célébration de soi que la manifestation d'un état d'angoisse où l'individu est confronté de plus en plus au problème troublant du sens (ce qui le centre sur lui-même). Ou ne serait-ce son indifférence à l'égard du sens, comme le postule Gilles Lipovetsky ? Il n'en demeure pas moins que le sens — soit-il au centre de la subjectivité moderne ou encore mis en veilleuse — ne fait plus l'objet d'une formulation qui relève d'un moule commun à tous comme celui auquel le citoyen-individu moral des XVIIIe et XIXe siècles était confronté. Dans les perspectives d'une quête de sens personnalisée toujours renforcée ou d'une éthique nihiliste du présent, le fait de l'effritement de l'individualisme moral demeure. Ce n'est pas pour rien non plus que Lipovetsky appelle son recueil d'essais sur l'individualisme contemporain, *L'ére du vide* (1983). Il y adresse d'ailleurs le problème d'une « deuxième révolution individualiste »; comme celle « d'un mode de socialisation et d'individualisation inédit, en rupture avec celui institué depuis les XVIIe et XVIIIe siècles. » (Lipovetsky, 1983, p. 9 - 10)

C'est le Tout, dans ce deuxième temps de l'individualisme, qui se dissipe de l'horizon des acteurs sociaux, non parce qu'il s'éclipse de la socialité — celle-ci reste un tout avec un ordre et des modes de gouvernance qui lui sont propres — mais parce que l'objet du sens moderne s'est déplacé d'un rapport de type moi-communauté ou moi-politique très actif vers un moi dont le rapport au communautaire et au politique s'est régularisé, autonomisé et automatisé. Le

sujet humaniste semble donc dériver. Dans l'optique du monde contemporain, il quitte son rapport originel au collectif pour se rapprocher de plus en plus vers lui-même, laissant tomber du même coup une compréhension de son autonomie qui reposerait sur la collectivité.

L'individualisme moderne a pu s'installer dans les mœurs en vertu de principes collectifs de solidarité pour un projet de société meilleure. Aussi, il a pu s'insinuer assez solidement dans le tissu social dès le XVIIIe siècle parce qu'il contrait l'injustice du pouvoir traditionnel. En Europe, où les oppositions à l'individualisme sont virulentes, surtout de la part des progressistes qui y voient l'outil des bourgeois et des inégalités sociales, c'est parmi les masses populaires que l'individualisme s'installe, pouvant rassembler tout un chacun dans des idéaux de liberté et d'autonomie (Laurent, 1993, p. 56). Mais le lien entre individualisme et collectivité, unis par la morale, va s'évanouir tranquillement surtout à partir de la moitié du XXe siècle en perdant sa substance pour gagner de la consistance. Le type de rationalisation propre à la modernité a eu une incidence de premier plan sur la manifestation d'un nouvel individualisme et d'une subjectivité humaniste transformée. La rationalisation du monde moderne à permis à la société de se transformer. C'est aussi à deux niveaux fondamentaux que ses effets doivent être reconnus; celui des conduites et celui des techniques.

4. 3. La rationalisation des conduites

Deux caractéristiques du processus de rationalisation peuvent être rattachées à la transformation de l'individualisme. La première renvoie à la

forme de la rationalisation spécifique à la modernité en tant qu'elle cherche par la voie d'une rationalité séculière à abolir la domination traditionnelle et l'hétéronomie. La subjectivité moderne et le droit moderne qui en est issu consistent en une critique des formes traditionnelles de la domination qui privent l'individu de son autonomie et de son auto-fondation. Ainsi vise-t-elle une égalité accrue — qui se manifeste dans le monde contemporain en luttes pour la reconnaissance — entre individus qui cherchent à être les agents déterminants de leur propre existence. C'est d'abord dans l'optique de la consolidation de cette rationalité dans le monde contemporain que nous voulons envisager la rationalisation de la subjectivité humaniste et individualiste.

Les restes de la tradition que la société moderne transportait avec elle (comme le patriarcat, le racisme institutionnalisé ou tout autre mode de citoyenneté exclusif) sont dorénavant l'objet de la critique de sujets qui veulent acquérir une plus grande autonomie. Ces acquisitions de reconnaissance — qu'on ne peut envisager de cesser d'exister puisque la société cache toujours des formes de domination[16] — ne se manifestent plus toutefois dans un cadre moral toujours renouvelé et agrandit qui inclurait les nouveaux « membres à part entière de la société ». La notion du citoyen-individu soutenue dans une perspective morale disparaît et est remplacée par un individualisme dont la source de l'autonomie se situe ailleurs. C'est un

16. Max Weber faisait une critique importante du communisme sur la base du fait qu'il reposait sur la fin de la domination de l'homme sur l'homme (Weber, 1995a, p. 256). Pour Weber, la domination est toujours présente. Les critiques contemporaines de la tradition ne signifient en rien que la domination traditionnelle y est disparue et que des attitudes et des phénomènes sociaux subtils ou ouverts ne perpétuent pas des modèles de domination de type traditionnel ou n'en élabore pas de nouveaux. L'idée générale que nous voulons mettre de l'avant dans notre argumentation est que la déconstruction de ces processus de discrimination sont maintenant plus facilement atteingnables en vertu de la subjectivité des individus dans le cadre d'un accroissement de la rationalisation moderne des conduites.

ailleurs dont nous avons déjà effleuré le lieu à la section précédente et qui fera l'objet principal du dernier chapitre.

En attendant, disons que ce premier effet de rationalisation repose sur une deuxième facette de la rationalisation du monde qui vient combler ses besoins matériels. La « démocratisation » du monde repose en fait sur des processus institutionnels et structurels qui lui correspondent et c'est là que nous chercherons à identifier, dans la partie suivante de l'analyse, une deuxième facette de la rationalisation. La rationalisation des exigences matérielles entourant l'accentuation des égalités et la reconnaissance des individus soulève la question de la raison instrumentale et de ses manifestations objectives et subjectives sur l'économique et le politique. On peut donc comprendre ces deux caractéristiques de la rationalisation dans le cadre de ce que Julien Freund appelle un double artifice; l'un[e] externe qui porte sur les objets ou les moyens matériels, l'autre interne qui concerne la conduite de la vie (Freund, 1990, p. 75).

Max Weber n'a pas postulé que la particularité du monde moderne et occidental se situait dans le processus de la rationalisation. Dans l'optique wébérienne, le processus de rationalisation est universel car tous les regroupements humains y ont recours pour organiser un monde dont la manifestation non rationalisée est un « chaos de sensations » (*Ibid.*, p. 73). Toutefois, le monde moderne s'est institué autour d'une modalité de la rationalisation qui lui est exclusive et qui est inédite si on la compare à celle qui prévalait dans les sociétés qui le précèdent. Alors que dans les sociétés pré-modernes, la rationalisation du monde avait des limites institutionnalisées,

comme la religion, la magie ou la tradition (*Ibid.*, p. 74), le monde moderne a repoussé toute forme de limites de par la nature de sa rationalisation séculière. Ainsi est-elle plus en mesure de croître et de s'étendre à tous les phénomènes de la vie d'une part, mais d'autre part de renforcer sa propre cohérence (*Ibid.*, p. 73).

C'est que l'organisation du monde moderne, plutôt que de reposer sur un mysticisme permettant d'apporter des réponses cohérentes à certains phénomènes du monde, comme les hiérarchies et les désastres par exemple, repose sur l'idée d'une rationalité dont l'accessibilité non-médiatisée se fait par l'être humain — par le sujet. C'est ainsi, nous l'avons vu, que le doute, le choix, la remise en cause, le non-sens ou l'angoisse sont des phénomènes qui ont une place privilégiée dans la subjectivité moderne. C'est ce que Weber thématise sous le thème du désenchantement du monde. Dans la mesure où l'individu, avec sa raison, est le principal véhicule de la rationalisation du monde, plutôt que l'autorité mystique traditionnelle, magique ou religieuse, les conditions objectives pour que la vérité se transforme toujours en doute ou en opinion sont présentes. C'est par cette compréhension de la « rationalisation interne » moderne que la transformation du sujet moderne « uniformément moral » en un sujet dont la moralité est de plus en plus individualisée et personnalisée (et dont les référents au collectif n'ont plus de bases morales) peut être comprise. Le monde contemporain est un moment exponentiel de la rationalisation du monde moderne.

Alors que des civilisations ont pu durer plus de mille ans, la modernité, dans la deuxième moitié du XXe siècle et après trois à quatre cent ans

d'existence à peu près, fait déjà l'objet d'un doute quant à sa survivance en tant que « paradigme » d'organisation du social[17]. Il est difficile pour le sociologue de nier le caractère auto-critique de la société moderne qui est issu de son type de rationalisation. Le thème de l'individualisme — et de la subjectivité humaniste qui le sous-tend — n'est pas étranger à ce phénomène et on peut faire la lecture de sa rationalisation à travers son rapport des siècles passés avec la moralité. Ross Poole soutient dans l'introduction de *Morality and Modernity* (1991) que de par sa nature, la modernité, quoiqu'elle ait eu besoin de la morale — envisagée ici, non pas comme jugement et discrimination qualitative quotidienne, mais comme principes transcendants de la société — pour s'instituer, la rend impossible. « Modernity both needs morality, and makes it impossible. » (Poole, 1991, p. ix)

C'est l'individu autonome et libre en tant que membre d'un tout social moral qui prévaut activement dans la conception de l'individualisme aux XVIIIe et XIXe siècles. Mais cette conception de l'individu au sein d'un ensemble moral sanctionnait inévitablement — par extension de ses prémisses sur l'autonomie, la liberté et la raison comme véhicule collectif (mais à la fois individuel) d'accès à la vérité — la volonté individuelle, la réflexion individuelle, bref, le détachement de l'individu d'avec l'ensemble dont il était issu dans l'appréhension originelle de l'individualisme. Ce détachement peut être compris comme une conséquence de la délégitimation des grands récits modernes et du rôle de la science comme fournisseur de vérités morales. Dans

17. C'est une question que nous cherchons à éviter avec le terme « société contemporaine ». Ce débat lourd de conséquences épistémologiques ne fait pas l'objet de cette analyse. Mais par ailleurs, nous croyons qu'il est impossible de comprendre le temps présent de la société sans le rapporter à la modernité « classique ». Cela est admis par les tenants de la modernité comme par ceux de la postmodernité. Par conséquent notre emploi d'auteurs des deux allégeances n'est basée que sur la perspicacité et la pertinence de leurs propos pour les thèses que nous défendons.

la perspective de la rationalisation interne du monde et de sa conjonction avec la chute des grands récits modernes, une conception morale de l'individualisme qui freine le développement de l'auto-détermination subjective est sujet à une virulente critique. Si la conception morale de l'individualisme s'est dissipée c'est parce qu'elle reposait sur une rationalisation du monde qui contenait le germe de son ébranlement.

Quoique Durkheim, par exemple, ait en quelque sorte tenté de faire de l'État le frein de la rationalisation moderne, si l'on peut s'exprimer ainsi, sa conception de l'autonomie et de la liberté individuelles ne concordait pas bien avec l'obéissance « éclairée » à l'organe étatique qu'il proposait. Si le sujet humaniste a une première version de lui-même dans l'individu comme membre et acteur du Tout, il porte en lui le germe de l'individu comme *princeps*, comme nous l'avons souligné au premier chapitre. Dans la logique de la rationalisation du monde moderne et de ses prémisses sur la raison et le jugement individuel, l'individu comme sujet à une morale collective qui a une source extérieure à lui-même est sérieusement ébranlé.

> Modernity has called into play a dominant conception of what it is to have reason to act: this conception has the consequence that the dictates of morality have little purchase on the motivations of those whom they are addressed. Modernity has constructed a conception of knowledge which excludes the possibility of moral knowledge; morality becomes, not a matter of rational belief, but subjective opinion. In a world which is antithectical to faith and dogma, morality can only survive as a matter of personal faith or dogmatic conviction. In neither case can morality retain the authority it needs to play its role in social and individual life (Poole, 1991, p. ix).

La grande faculté critique de la société moderne — issue donc, de la subjectivité humaniste et rationnelle qui fait naître l'individu comme *princeps*

— a joué un rôle important dans la transformation de l'individualisme et dans la déconstruction de la subjectivité moderne originelle. Les mouvements sociaux des années soixante en sont un bel exemple, ainsi que la plus claire des manifestations, dans la mesure où ils prônaient une rupture d'avec les restes de la tradition que la modernité avait traînée avec elle. La chute rapide du marxisme aux États-Unis — idéologisé par les intellectuels au début du siècle et par l'URSS durant près de 45 ans — comme la voie privilégiée de la critique, peut aussi être compris dans la même mouvance que les revendications des Femmes ou des Noirs. C'est-à-dire comme une déconstruction de rapports ou de phénomènes sociaux institutionnalisés qui échappaient à l'accès immédiat de l'individu par sa volonté et sa raison. Dans l'optique de la spécificité de la rationalisation moderne du monde, la soumission impuissante de certains aux Hommes, aux Blancs, ou au Parti, n'avait de sens.

La grande difficulté pour une morale-rationnelle comme vérité universellement valide de se maintenir en vertu du type de rationalisation du monde moderne y est pour beaucoup dans l'avènement de ces mouvements de résistance et de contestation populaires. Mais cela ne constitue qu'un versant de la rationalisation du monde moderne et ne peut suffire à rendre compte d'une subjectivité rationalisée qui se dessine. À la rationalisation des attitudes et des conduites correspond aussi une rationalisation du monde matériel. Alors que Max Weber la reconnaît, dans le monde moderne, sous l'emblème de la bureaucratie, sa prolongation contemporaine est celle de la technocratie, ou de la gestion techno-scientifique du social et du politique. Si le monde contemporain marque une rupture sociologique dans l'organisation du tissu social, c'est parce qu'à la rationalisation interne du monde correspond une

rationalisation externe. Le nouveau visage de la subjectivité ne peut être compris sans les deux dimensions de la rationalisation que nous voulons souligner.

4. 4. La rationalisation des techniques

Un mouvement important de désinstitutionnalisation se met donc en branle au milieu de ce siècle pour secouer la dimension morale de la solidarité de la collectivité par la critique du patriarcat, du racisme, ou de l'autorité. Parallèlement à cet effet de la rationalisation moderne, un autre effet dont la manifestation à certains égards peut sembler contradictoire avec celle du premier, peut être constaté. C'est que l'on peut aussi lire le mouvement de rationalisation et de radicalisation de l'individualisme à travers la question de la raison instrumentale, qui constitue la source hégémonique de la gouvernance des sociétés occidentales. Si un grand mouvement de désinstitutionnalisation de certains rapports sociaux se met en branle à une certaine époque, c'est au même moment que l'institutionalisation de l'individualisme et de son habitat social-historique — le libéralisme économique — se consolide au moyen de la technocratie. Mais il est important, avant d'aller plus loin, d'élucider le rapport entre le monde moderne et contemporain et l'activité de type rationnelle et instrumentale.

Weber a fait remarquer tout au cours de son œuvre que c'est la rationalité instrumentale qui est le type de rationalité engagée dans l'activité la plus importante dans le monde moderne. L'adéquation des moyens disponibles aux fins visées en fonction des conséquences impliquées dans une

activité se distingue d'une activité dont la rationalité est orientée en définitive par une valeur. Dans cette dernière ce sont des valeurs qui dictent les fins à atteindre sans que les moyens disponibles ou les conséquences probables du geste prennent une place de premier plan. Dans le cas de l'activité orientée en finalité, c'est l'aspect du calcul froid et instrumental qui est dominant. On peut tenter d'expliquer la domination moderne et contemporaine de ce type de rationalité à partir de son altérité avec la tradition et de sa tendance à la sécularisation. Elle vise une rationalité accrue des rapports entre individus fondés sur la consistance (bureaucratie/technocratie) et l'efficacité (capitalisme).

Pour ce qui est de la première caractéristique du processus de rationalisation du monde moderne que nous avons voulu mettre en évidence, la question de la rationalité instrumentale est moins au centre de son explication que le rejet de l'activité rationnelle orientée en valeur en fonction d'une rationalisation des conduites. C'est beaucoup moins le fait que la rationalisation du monde moderne incite positivement à adopter une rationalité de type instrumentale, qu'elle incite le rejet de la domination traditionnelle — qui est déjà mis en branle à l'époque des révolutions bourgeoises d'Europe et d'Amérique — qui fait en sorte justement que les attitudes des individus changent à l'égard de l'autorité traditionnelle.

Toutefois, le rejet des récits légitimateurs traditionnels, sous fond d'une subjectivité individualiste et humaniste, s'effectue nécessairement en conjonction avec une transformation de la forme que prend la société dans son ensemble. Par exemple, la Révolution tranquille au Québec n'a pas été qu'un simple changement des attitudes privées des individus, elle a aussi déversé ses

effets sur le fonctionnement global de la société québécoise; en passant par l'économique, jusque dans chacune de ses institutions. C'est que les exigences matérielles d'organisation du monde s'imposent toujours. Le deuxième aspect de la rationalisation moderne que nous tentons d'élucider — celui qui concerne la question des exigences matérielles et techniques de la société — doit être compris à la lumière de la transformation des attitudes à l'égard de la tradition, conjointement au rôle de l'activité rationnelle en finalité dans le processus d'appropriation rationnelle et libre des individus de la gestion du social et de l'économique. Cette appropriation peut être cristallisée dans l'hégémonie d'activités dont la rationalité est instrumentale et qui prennent place sur le terrain de l'activité économique capitaliste, sous fond d'une gestion bureaucratisée de la société.

C'est dans l'optique de l'accroissement de cette rationalisation que le monde contemporain se voit dépourvu d'un référent commun qui transcenderait l'individualité. Aussi le sujet qui tente de s'émanciper de l'emprise d'un individualisme moralisant va-t-il quérir ailleurs que dans le collectif son épanouissement. La consommation, depuis les années cinquante, (ou aujourd'hui la petite entreprise) est à cet effet un des lieux privilégié de la mise à jour de la liberté individuelle et subjective. On verra que la rationalisation du monde moderne est indissociable de l'autonomisation du mode de production capitaliste et de la bureaucratie. Nous supposerons par ailleurs que leur autonomisation entre dans la logique de la transformation contemporaine de la subjectivité. Mais avant de s'attarder à ces préoccupations, il faudrait éclairer la logique de la rationalisation moderne des techniques, surtout dans son rapport étroit au capitalisme et à la bureaucratie.

Dans les analyses de Marx, le fait des révolutions bourgeoises en Europe et des valeurs juridiques, politiques et idéologiques qui y sont associées est déterminé en dernière analyse par les impératifs de l'économie de marché capitaliste. Dans cette perspective, l'appareil institutionnel et idéologique de la société sert la domination capitaliste du monde. Chez Durkheim par ailleurs, c'est dans l'ordre industriel (dont le rapport avec le capitalisme est sociologiquement inévitable) — qui donne naissance à l'individualisme par la division du travail et la division subséquente des consciences — qu'étaient envisagés les fondements de la supériorité évolutive de la société moderne. Ainsi, comme nous l'avons vu, l'individualisme devenait, à cause de l'ordre industriel, un aspect moral de la société. De façon différente et avec des implications politiques séparées par tout un monde, ces deux sociologues n'ont pu éviter de constater, à leur façon, la conjonction entre le capitalisme et les valeurs de l'Europe révolutionnaire. Sans entrer dans le débat du degré d'incidence du libéralisme économique sur la subjectivité humaniste, et vice-versa, leur existence respective ne peut être comprise à l'extérieur d'un cadre commun.

Weber, quant à lui, est loin de réfuter la richesse explicative de la vraisemblable équation entre modernité et capitalisme. Toutefois, à la différence de Marx, il évite de réduire le capitalisme en cause unique de l'émergence du visage socio-juridique qui caractérise les sociétés modernes occidentales. Il propose d'ailleurs, un peu à l'instar de Tocqueville, mais de façon plus systématique, dans *L'éthique protestante et l'esprit du capitalisme,* un type idéal du développement du capitalisme qui repose sur l'influence de la culture, des idées, des valeurs et de la religion, plutôt que l'inverse. Pour

Weber, le changement social véritable se situe à l'intérieur de la subjectivité et il doit relever d'une force charismatique ou extramondaine — il identifie le charisme comme force créative révolutionnaire de l'histoire (Weber, 1995, p. 232). Dans l'optique wébérienne, si le capitalisme est le mode de production privilégié de l'époque moderne, c'est en vertu de son affinité avec la subjectivité que cette époque et ses contingences historiques produit. Autrement dit, il n'y a pas de relations causales déterministes, il n'y a que des probabilités causales d'affinité.

Mais si la domination charismatique est révolutionnaire, les besoins quotidiens (l'économie), la planification et la stabilité (la gestion) finissent par s'imposer sur l'aspect extramondain de la socialité et ce sont des formes rationalisées des valeurs créatives et révolutionnaires qui s'imposent sur le monde. La tradition (comme le patriarcat moderne) ou la bureaucratie (la technocratie, dans sa version contemporaine) jouent le rôle d'une rationalisation intramondaine des valeurs révolutionnaires et substantielles qui s'ancrent dans l'économique et la gestion du social. « The economy, as an permanent organised system of transactions for the purpose of planned provision for the satisfaction of material needs, is the specific home of the structure of the patriarchal structure of domination, and of the bureaucratic structure as it becomes increasingly rationalised to the level of 'enterprise'. » (*Ibid.*, p. 233) Ainsi, dans le monde moderne, l'habitat de la rationalisation est le capitalisme.

Ainsi, la force de la rationalisation matérielle s'impose en vertu de la nécessité de satisfaire efficacement aux besoins quotidiens. Dans la mesure où

l'on peut lire Weber dans l'optique d'un déterminisme idéationnel (le charisme comme force révolutionnaire), les forces de la rationalisation font renverser ce rapport car elles contraignent l'individu à s'adapter aux rationalisations techniques du monde. C'est d'ailleurs à la rationalisation des motivations extramondaines des puritains anglais dans l'accumulation de capital que Weber fait référence quand il postule que si le puritain voulait travailler, nous sommes maintenant forcés de le faire[18]. Aussi le monde moderne est-il le théâtre d'un renforcement de la domination bureaucratique dans la mesure où, la rationalisation des conduites s'accroissant, la rationalisation des techniques rationnelles-légales de gestion du social elle aussi s'accroît. Si par ailleurs la bureaucratie peut sembler révolutionnaire par rapport à la tradition, ses effets sur le changement social n'ont, selon Weber, pas la portée des changements sociaux liés au charisme et se transforment même en contraintes externes.

> As we saw, bureaucratic rationalisation can also be, and often has been, a revolutionary force of the first order in its relation to tradition. But its revolution is carried out by *technical* means, basically 'from the outside' (as is especially true of all economic reorganisation); first it revolutionises things and organisations, and then in consequences, it changes people, in the sense that it alters the conditions to which they must adapt and in some case increases their chances of adapting to the external world by rational determination of means and ends. The power of charisma, by contrast, depends on beliefs in revelation and heroism [...] or judicial wisdom or magical or other favors. Such belief revolutionises men 'from within' and seeks to shape things and organisations in accordance with its revolutionary will (*Ibid.*, p. 231).

18. L'éthique protestante dans son rapport à l'esprit du capitalisme constitue un bel exemple d'une rationalisation intramondaine vers la tradition. Mais il constitue aussi un exemple privilégié d'une rationalisation de la tradition vers la bureaucratie. Toute la place accordée à la divinité dans l'accumulation capitaliste des biens chez les puritains de l'Angleterre, ou même des États-Unis, est disparue du capitalisme et est remplacée par la domination rationnelle-légale qui caractérise nos sociétés. La bureaucratie, comme l'écrit Weber, est une force révolutionnaire dans son rapport à la tradition. En soi toutefois, elle contribue à renforcer certains ordres, sans perpétuer un enchantement, d'où le fait que nous soyons maintenant « forcés de travailler ». Si Weber emploi le verbe « forcer » c'est justement parce que le désenchantement moderne nous le fait apparaître sous ce jour.

Ainsi la bureacratie constitue-t-elle une force révolutionnaire par rapport à la tradition dans la mesure où elle tente de préserver et de consolider une sécularisation des conduites. Mais alors que la tradition est en mesure de préserver un caractère enchantant sur le monde, la bureaucratie, l'entreprise rationnelle et la rationalité instrumentale qui la sous-tend, consistent essentiellement à faire des adaptations des moyens aux fins dans le cadre d'une rationalité instrumentale visant soit la performance (économique) ou la consistance (rationnelle-légale). Dans l'optique wébérienne, il s'agit d'un drame dans la mesure où le pouvoir appartient dit-il à des « spécialistes sans âme ni vision et de voluptueux sans cœur. » (Mongardini, 1975, p. 102) De fait, le monde contemporain plus que jamais, est le lieu de la fin des visions.

Une des grandes différences entre le patriarcat et la bureaucratie comme pourvoyeurs des besoins quotidiens réside dans le type de rationalité qui sous-tend leur légitimité respective. La légitimation moderne ne s'exerce plus en soi à partir d'un ordre supra-individuel, comme dans le cas du patriarcat, qui dicte les modalités de l'organisation de la socialité : même les principes moraux de l'individu-citoyen des XVIIIe et XIXe siècles sont ancrés dans une autonomie subjective et une légitimité qui provient « du bas ». La légitimité du fonctionnement de la société, dans la modernité, devient tributaire de sa rationalisation fondée sur l'individu comme *princeps*. Les individus, leur raison et leurs valeurs humanistes deviennent le cœur du processus de la légitimité de l'ordre capitaliste et bureaucratique.

On peut donc comprendre comment ils adoptent, dans leur vie privée, une culture rationnelle et instrumentale qui correspond aux exigences de la

performance au sein d'une économie de marché et comment ils rendent légitime un système de gestion global de la société qui se veut objectif, parce que dépersonnalisé et rationnel en fonction d'une adéquation des moyens aux fins à atteindre. Si ce modèle persiste dans le cadre de la modernité classique, dans une atmosphère d'enchantement révolutionnaire, de progrès et de justice universelle quasi extramondain, la délégitimation contemporaine des discours modernes dominants change néanmoins ce portrait.

L'univers qui entoure l'individualisme, dès lors, n'est plus moral, mais plutôt fonctionnel et adapté aux besoins quotidiens des individus qui agissent dans un espace social partagé, sans nécessairement qu'il relève de l'idée d'une solidarité sociale. Weber, au début du vingtième siècle, entrevoyait déjà cette désaffection du « dieu » qu'étaient les valeurs humanistes pour aller vers leur autonomisation et leur technicisation.

Dans les vues wébériennes, la bureaucratisation du monde et la consolidation de l'activité économique rationnelle ne sont pas nécessairement le gage de l'émancipation. D'une dimension supra-individuelle et extramondaine, les valeurs humanistes d'égalité, de liberté, de dignité se sont vidées de leur caractère moral pour ancrer leur pertinence dans les structures du quotidien, de l'efficacité et de la performance capitaliste et bureaucratique avec des effets dévastateurs, comme le postule un interprète de Weber :

> Le « désenchantement du monde » apporte de nouvelles conditions d'émancipation à l'homme moderne, mais finit par nier l'humanité elle-même. Elle conduit à la substitution des règles religieuses ou morales par des impératifs abstraits de type rationnel-légal, mais nie à l'homme moderne ces « dieux » sur lesquels se fondaient la coexistence et la durée du lien social (*Ibid.*, p. 97).

Ainsi la rationalité instrumentale est l'attitude à l'égard du monde vouée à croître le plus dans la modernité, autant dans les lieux de la gestion et de l'économique que dans celui de la subjectivité. Comme Philippe Raynaud le rapporte, le constat wébérien de l'autonomisation de la bureaucratie qui accompagne la rationalisation émancipatrice moderne envisage des conséquences désastreuses et paradoxales sur l'autonomie des individus. Elle les enfermerait dans une « cage de fer » et trahirait la rationalisation émancipatrice moderne des conduites :

> Le développement du système économique moderne se traduit donc à la fois par le renversement de l'émancipation en asservissement, par l'intégration des besoins économiques (« souci des biens extérieurs ») dans un système autonome qu'ils [les hommes] ne dominent pas mais dont, au contraire, ils sont dépendants et, enfin, par le risque radical d'une perte de sens et d'une fin de l'autonomie subjective (Raynaud, 1987, p. 185).

Anthony Giddens quant à lui fait remarquer que chez Weber l'institutionalisation de la bureaucratie est irrationnelle dans la mesure où elle fait entrave à certaines des valeurs fondamentales du monde moderne. Elle subordonne l'individu, et Weber, comme le précise le sociologue anglais, n'est pas confiant pour l'avenir de la société moderne :

> Weber [...] perceives a primary irrationality within capitalism. The formal rationality of bureaucracy, while it makes possible the technical implementation of large-scale administrative tasks, substantively contravenes some of the most distinctive values of western civilisations, subordonating individuality and spontaneity. But there is no rational way of overcoming this : this is ' the fate of the times ', to live in a society characterised by ' mechanised petrification ' (Giddens, 1971, p. 217).

Le monde contemporain est le lieu d'une accentuation de la double face de la rationalisation. S'il est le théâtre d'une remise en cause radicale de certaines attitudes traditionnelles (dont Weber n'a pas connu l'envergure) de la

part de certains groupes identitaires qui s'organisent politiquement, le monde contemporain se voit aussi se consolider une domination rationnelle-légale technocratique d'inspiration libérale, caractérisée plus que jamais par une « désubstantialisation » et par la « routinisation » de la pensée morale humaniste. D'un point de vue donné, cela consacre un changement radical dans la nature du politique, si ce n'est sa disparition. Il y a en effet une domination de l'ordre rationnel-légal qui s'étend de l'économique, en passant par la gestion entière de la cité, et Weber a raison de le faire remarquer. À savoir toutefois si cette domination est plus tyrannique que n'importe laquelle, nous en doutons. D'une part, nous ne croyons pas qu'elle consiste en l'anéantissement de l'autonomie subjective. D'autre part, si elle peut s'inscrire dans des stratégies de pouvoir politique de la part de dominants, nous pouvons aussi la lire dans une transformation structurelle de la société correspondant à de nouvelles visées subjectives. Dans cet ordre de pensée, l'expression popularisée de Michel Foucault, « stratégie sans stratège » est tout à fait à propos.

Chapitre 5 : La dissolution de la solidarité et l'autonomie subjective

Nous avons tenté de voir, à la section précédente, comment la désagrégation du caractère moral — que nous n'avons pas pleinement abordé encore — de l'individualisme dans le monde contemporain peut être rendu possible par les modalités particulières de la rationalisation et de la subjectivité et des techniques modernes. Nous avons postulé d'une part que la rationalisation sans limite de l'appréhension du monde moderne compromet les récits explicatifs traditionnels du monde. Nous avons supposé par ailleurs qu'une rationalité de type instrumentale, c'est-à-dire qui adapte les moyens disponibles et les conséquences probables d'une action aux fins visées, était devenu le moyen privilégié pour les individus de contrer les effets de la domination traditionnelle afin d'ancrer leur volonté et leur autonomie dans le quotidien. Nous avons aussi mis de l'avant la thèse selon laquelle le nouvel ancrage moderne des individus dans le quotidien, en remplacement de la forme traditionnelle de production/consommation, est essentiellement circonscrit dans la bureaucratie et l'économie de marché.

La bureaucratie d'une part, parce qu'elle représente une révolution « plus » rationnelle et « plus » juste du modèle de gestion traditionnel de la société qui correspond mieux aux visées d'auto-fondation des individus modernes. Elle implique une foule de recours universaux fondés sur des critères objectifs, rationnels-légaux, raison pour laquelle elle est révolutionnaire par rapport à la tradition. L'économie de marché d'autre part, car elle constitue le lieu privilégié de la production des biens de la modernité, mais parce qu'elle constitue aussi un espace d'activité économique dont la légitimité est le libre-agent et la volonté individuelle. Comment toutefois cela

équivaut-il à la dissolution de l'individualisme avec une dimension morale et solidaire en vue d'un projet social partagé, tel qu'il existait aux XVIIIe et XIXe siècles ?

Le politique et l'économique sont aujourd'hui plus que jamais techniques. On assiste à une autonomisation du marché et à une automatisation des décisions politiques (technicisées et « expertisées ») en fonction des impératifs de la performance économique. Dans *Les conséquences de la modernité* (1994), Anthony Giddens affirme que « l'économie est complètement distincte, ou « isolée » des autres arènes sociales, et particulièrement des institutions politiques. » (Giddens, 1994, p. 62) Cette isolation signifie que l'économique est à l'abris du politique et du social et que par conséquent, « les relations économiques ont une emprise considérable sur les autres institutions. » (*Ibid.*, p. 62) Ainsi « l'autonomie de l'État est conditionnée, et non déterminée, par sa dépendance à l'accumulation du capital, sur laquelle son contrôle est loin d'être complet. » (*Ibid.*, p. 63) Le politique épouse plus que jamais les formes des impératifs économiques dans un cadre de décisions de plus en plus expertes et de moins en moins « politiques ».

Le politique et la gestion du monde contemporain reposent sur ce que Giddens appelle une dé-localisation des processus ou des systèmes. Pour le sujet contemporain cela se manifeste par une certaine foi (issue d'un peu de savoir induit) en le bon fonctionnement de l'ensemble social technicisé. Cette confiance en les systèmes experts (comme la justice) ou en les gages symboliques (comme l'argent) prend aussi forme dans l'appréhension de

l'autre, c'est l'inattention polie. La confiance révèle donc le visage anonyme des interactions sociales qui reposent sous fond d'une confiance partagée en le respect d'autrui ou en la confiance en la stabilité du système. De façon générale, elle relève aussi d'une routinisation du fonctionnement de la société et d'une acceptation tacite de circonstances qui laissent peu de choix (Ibid., p. 96). Aussi ces systèmes abstraits qui caractérisent le monde contemporain sont-ils plus que jamais difficilement accessibles à la fuite du sujet, « [...] personne ne peut complètement échapper aux systèmes abstraits des institutions modernes. » (*Ibid.*, p. 90) De façon générale, la société, comme le lien social, sont technicisés et la solidarité est contenue dans des formes procédurales et désubstancialisées. « La technique postmoderne, produit de la rationalité moderne, est devenue le pivot de l'être-ensemble, c'est-à-dire le type mécanique de solidarité dont le pivot est constitué par l'organisation. La solidarité technique multiplie les contacts entre l'organisation et les organisés, mais aussi entre les organisés. » (Jacques Zylberberg IN Boisvert, 1995, p. 31)

Le visage de la domination rationnelle-légale contemporaine prend *grosso modo* cette allure. La solidarité est remplacée par un lien abstrait et technique entre les individus et le lieu de l'épanouissement subjectif n'est plus relié à la sphère publique. Le sujet contemporain est d'une part détaché du fonctionnement global de la société et il est par ailleurs conditionné et « forcé »[19] par cette domination. Est-ce que cela signifie que le sujet moderne n'est que le produit d'une domination rationnelle-légale dont la portée lui échappe définitivement, comme Weber le laisse supposer ? Peut-il être envisagé comme un élément ayant conditionné cette nouvelle socialité,

19. Par « forcé », nous faisons référence au sens wébérien de cet emploi du terme dans le postulat selon lequel nous sommes forcés de travailler.

comme ayant conditionné son cantonnement dans un individualisme rationalisé, un agent de la consommation et un chercheur de sens au quotidien ? Par ailleurs le sujet est-il relégué à une confrontation perpétuelle à l'aliénation subjective ou est-il au contraire le sujet d'une autonomie nouvelle ?

5. 1. La rationalisation de la société contemporaine en technique

Des phénomènes importants sont au centre de l'explication de notre postulat de la rationalisation de l'individualisme. Un point de départ fertile est celui de la science et de la technique dans leur rapport au politique et à l'économique. Dans sa version technicisée, la science, sous son emblème dorénavant performative (Lyotard, 1979), est graduellement devenue le véhicule de la gestion de la société en remplacement de la tradition — et des traditions qui persistaient dans la modernité « classique ». Dans un monde où la rationalité stratégique est le moyen inévitable de l'ancrage des sujets dans le monde des besoins quotidiens (la nécessité de gagner sa vie, notamment), la techno-science devient la forme hégémonique de la planification du social, ce dernier étant dorénavant envisagé comme système dont le but est la fonctionnalité ou la performativité. Parmi les causes vraisemblables de ce phénomène, qui ne peut selon nous se résumer à des stratégies de domination, se trouve le cantonnement des quêtes de sens des individus dans leur quotidien et dans un monde économique qui leur est immédiatement accessible. Autrement dit, nous supposons que si la domination rationnelle-légale se consolide au point d'atteindre une dimension strictement technique dans le monde contemporain, ce ne peut être sans que des visées subjectives

nouvelles émergent. Il y a une affinité historique de causalité — plutôt qu'une déterminité — qui peut expliquer l'émergence de la domination technocratique et de la subjectivité contemporaine.

L'individu contemporain laisse le destin de la solidarité et de la collectivité entre les mains d'institutions et de spécialités. Aussi, la moralité de l'ensemble social contemporain est supplanté par sa fonctionnalité comme finalité — surtout économique. Par ailleurs, l'individualisme, tout en restant chargé d'un bagage culturel et moral issu de l'humanisme — en façonnant des sujets qui valorisent l'autonomie, la liberté et la dignité et qui se frayent une voie dans le monde selon ces valeurs —, prend une forme de plus en plus objective et de moins en moins subjective. Les catégories morales modernes se posent moins en valeurs politiques qu'en faits sociaux institutionnalisés. Le cantonnement de la subjectivité dans l'activité économique et le quotidien pour quérir du sens a donné lieu à une organisation de la société qui est axée sur son automatisation et sa stabilité institutionnalisée, ce qui explique en partie des phénomènes comme l'apathie de l'électorat. Dans cette optique, les structures institutionnelles — qui instituent des conditions objectives nouvelles — de la société contemporaine, de façon générale, rendent l'individualité et l'individualisme des questions dont la portée morale est symbolique.

Certaines thèses de Jürgen Habermas, élaborées dans les années 1960 à la suite des travaux de Herbert Marcuse dans *L'homme unidimensionnel,* constituent une source de premier plan pour l'explication de la légitimation techno-scientifique du cadre institutionnel du monde contemporain.

Toutefois, alors que les deux semblent aller dans une direction laissant supposer la mise en œuvre d'une stratégie concertée en vue de la domination des masses par les puissants de ce monde, nous croyons plutôt que le processus de rationalisation de la société vers une légitimation techno-scientifique de l'ordre est le fait de la conjonction des effets objectifs de la rationalisation des techniques et des effets subjectifs de la rationalisation des conduites. C'est beaucoup moins, en notre sens, la quête de certains de pouvoir absolu, que la quête subjective d'ancrage dans la stabilité et le sens — qui, dans sa forme moderne, trouve son foyer dans le quotidien et l'activité économique libérale — qui est à même de fournir l'explication la moins naïve et la moins réductionniste du cantonnement du politique, des valeurs morales modernes et de l'individualisme dans la fonctionnalité du système sous fond d'une légitimation techno-scientifique. C'est donc aussi beaucoup moins une « illusion » ou de la « programmation », comme le propose Habermas, que le fait de l'incarnation subjective des acteurs sociaux dans une forme rationalisée et « détraditionalisée » du social.

Dans *La technique et la science comme « idéologie »*, Habermas s'intéresse à la question de la légitimation dans les sociétés capitalistes avancées. Il postule que la rationalité instrumentale qui sous-tend l'activité socio-économique moderne permet graduellement, en visant la stabilité et la fonctionnalité, à la technicisation de la science, dans les champs de la politique, de l'économique et du social, de devenir la plaque tournante de la légitimité de l'ordre social. Habermas met de l'avant un argument semblable au nôtre en ce qui concerne l'investissement des sujets dans le quotidien — dans l'activité économique de marché plus spécifiquement — en vue d'obtenir de la stabilité.

À l'origine, la légitimité de l'activité économique de marché, selon Habermas, repose sur le fait que c'est « du bas » (des acteurs sociaux) qu'elle provient avant tout. C'est parce qu'il correspond aux catégories morales de la subjectivité individualiste et humaniste que le capitalisme peut exister. Le fait que des libres-agents en interaction soient au centre des transactions, des rapports de production et des échanges sur la scène du marché remplace la légitimité de la domination politique légitimée par la tradition et la rend, dès lors, impossible.

Mais alors que la domination traditionnelle était à proprement parler politique — parce qu'elle légitimait l'ordre social, les classes sociales, les rapports de production, bref, le cadre institutionnel de la société, strictement en fonction de la sphère du politique —, la domination rationnelle-légale de la société moderne est ancrée ailleurs que dans la stricte politique. Si par exemple, la rationalisation du charisme se transforme en tradition patriarcale, la légitimité du cadre institutionnel social reste politique. D'un autre côté, la rationalisation de la subjectivité moderne en une attitude orientée vers l'activité économique rationnelle à des conséquences tout à fait différentes. Nous l'avions fait remarquer plus haut; la question de l'économique est au cœur de la question politique.

> Ce n'est qu'avec le mode de production capitaliste que la légitimation du cadre institutionnel peut être directement lié au système du travail social. C'est seulement alors que le statut de la propriété, de *rapport politique* qu'il était devient un *rapport de production*, car il trouve sa légitimation dans la rationalité du marché, dans l'idéologie de la société de l'échange et non plus dans un statut de domination légitime en soi. C'est bien plutôt le système de domination qui peut être justifié en invoquant la légitimité des rapports de production : et tel est le véritable sens de la rationalité du droit naturel de Locke à Kant. Le cadre institutionnel de la société n'est que médiatement politique; immédiatement il est économique (l'État constitutionnel bourgeois comme « superstructure ») (Habermas, 1973, p. 31).

C'est donc un système de domination rationnel-légal, sous fond d'un déterminisme de l'activité économique de marché, comme lieu de la mise à jour de l'autonomie, de la volonté et de la liberté des acteurs sociaux qui s'impose comme cadre institutionnel de la société moderne. Immédiatement économique, le cadre institutionnel bourgeois n'est que médiatement politique en tant qu'il cherche à ajuster sa légitimité avec les exigences du libre-marché.

Mais le capitalisme moderne repose encore sur une dimension politique et morale afin de légitimer son existence et de contrer les résistances, comme nous l'avons fait remarquer au chapitre troisième. Le calcul rationnel de la finalité fonctionnelle du système s'imposera graduellement pour remplacer la forme politico-morale de la légitimité, ce fût l'objet du quatrième chapitre d'y déceler les principaux mécanismes. C'est la science, au service de la technique, qui s'imposera éventuellement comme moyen le plus efficace de maintenir en place ce système et de le garder fonctionnel. Toutefois, ce n'est pas avant que le système économique capitaliste se soit presque écroulé, ni sans instituer des mesures sociales-démocrates que son autonomisation en tant que système de production/consommation presque naturel, apolitique et amoral, soit engendré.

Selon Habermas, le rôle légitimateur de la science et de la technique dans la domination rationnelle-légale et capitaliste de la société contemporaine naissante se consolide après les expériences des crises économiques. C'est donc vers les années de l'après-guerre que se solidifie l'aspect fonctionnel, amoral et apolitique de l'hégémonie rationnelle-légale. Dans le but de mettre le système à l'abri des défaillances d'une activité économique non-régularisée et non-

régulée, la science et la technique deviennent les outils de la transformation de la société en un lieu fonctionnel d'activité, mais aussi la source de légitimation des mesures politiques, économiques et sociales du cadre institutionnel qui prend une forme de plus en plus rationnelle. Ces nouvelles mesures, ce « *new deal* », ce seront surtout l'interventionnisme de l'État dans des programmes sociaux et dans l'activité économique.

Or, il s'agit en quelque sorte d'un retour du politique, si on le compare au libéralisme pur, écrit Habermas, dans la mesure où on ne peut plus entièrement faire reposer les cadres institutionnels sur l'économique et ses lois internes. Mais c'est une politique négative, de « guidage » technique et instrumental, plutôt qu'une politique positive, morale et « solidariste » qui est mise en place, écrit-il. C'est la science, au service de la technique, qui prend les commandes de la gouvernance et qui représente la nouvelle légitimité du cadre institutionnel. Dans ces circonstances, le contenu de l'ordre de la vie bonne et des finalités morales de la politique est substitué par des finalités d'ordre techniques et fonctionnelles reliées surtout à la stabilité et au maintient du système économique[20]. Le politique, dans ces circonstances, devient de plus en plus l'affaire de spécialistes qui font plus souvent qu'à leur tour le pont entre leur domaine d'expertise et la santé économique de l'État.

Ainsi, parmi les implications de cette domination techno-scientifique de la gestion socio-économique se trouve ce que Habermas appelle, dans un autre texte de la même époque, la « scientificisation de la politique ». La scientificisation de la politique contribue à la démoralisation de

20. Ce qui revient à dire que le politique se dénaturalise, si on se fie à la proposition de Habermas selon laquelle le politique, en dernière instance, n'aurait pas de fondements rationnels mais serait plutôt ancré dans le domaine des valeurs (Habermas, 1973b, p. 99).

l'individualisme et à l'évanouissement de la politique en tant que sphère de débat sur les valeurs et les visions de solidarité que la société comme ensemble devrait adopter. D'abord, parce qu'elle réduit de plus en plus en plus la sphère du politique à des questions d'ordre technique pour une finalité qui est la performativité du système (économique surtout), ce qui réduit la valeur révolutionnaire de l'individualisme libéral pour l'ancrer dans la stabilité et le quotidien. Mais aussi, parce que la scientificisation du politique effectue un clivage. Il y a un récif entre le politique dorénavant soumis au règne de spécialités (plutôt que des spécialistes) qui gèrent l'ensemble de la société et les individus qui performent leur autonomie autour du sens dans la vie quotidienne et qui font des projets à des échelles de plus en plus réduites, c'est-à-dire de plus en plus personnelles et de moins en moins sociales.

Mais il ne faut pas réduire cette démoralisation et dépolitisation de la population à une stratégie bourgeoise, comme Habermas et Marcuse semblent vouloir le laisser entendre. À la perspective de l'opposition entre une classe de dominants, aidée d'une technocratie, mettant à leur service des masses aliénées, nous préférons celle d'une dynamique globale d'une société qui s'organise autour d'exigences subjectives et dont les conséquences varient, selon le point de vue duquel on les aborde. C'est en ce sens que nous empruntions plus haut l'expression foucaldienne de « stratégie sans stratège ».

5. 2. La domination technocratique et l'autonomie subjective

Ainsi l'hyper-rationalisation contemporaine et ses conséquences sur le politique comme gestion du social technocratisée ne sont peut-être pas si

étrangères à la façon pour le sujet contemporain, qui émerge au XXe siècle, d'appréhender son environnement social. On peut lire le désistement généralisé du politique dans le monde contemporain à travers la lunette d'une aliénation dues aux forces de la rationalisation et de la domination technocratique rendant l'accès au politique impossible. Ou encore peut-on lire ce désistement à travers la lunette de la subjectivité. La première possibilité nous apprend que le sujet est objectivement impuissant et la deuxième nous propose un sujet dont les enjeux existentiels ne se jouent pas au niveau du politique et qui s'approprie un système objectivant, voire qui participe à sa création. Nous supposons que dans le monde contemporain, si le sujet revendique plus son droit à l'égalité et à l'équité, il ne cherche pas à joindre une grande communauté de citoyens, mais plutôt à accéder à son quotidien en toute liberté. Nous supposons aussi, par ailleurs, que si les conditions objectives de la domination contemporaine peuvent être explicatives de cette nouvelle attitude — en tant qu'elle circonscrirait le sujet dans des lieux bien précis d'activités —, il n'en demeure pas moins qu'une appropriation subjective de ce terrain social par les acteurs sociaux — vers des formes nouvelles d'autonomie — ne permet pas d'exclure de l'analyse de la société contemporaine le rôle créateur des visées subjectives contemporaines.

À travers le caractère automatique et routinier du politique, l'autonomisation de l'économique, la spécialisation et la technicisation des prises de décisions et des débats, le caractère auparavant moral de la solidarité fait place à son institution dans un cadre formel et fonctionnel qui érige la société en système qui vise la consistance et la fonctionnalité plutôt que l'entretient d'un discours sur le progrès ou la justice universelle. Or, cela n'a

rien d'étranger à une autonomie subjective dont le lieu se dessine dorénavant ailleurs que dans le collectif. Michel Maffesoli fait remarquer l'opposition diamétrale entre le monde moderne pré-contemporain (qu'il appelle le bourgeoisisme) et le monde contemporain en ce qui à trait aux orientations éthiques et l'attitude à l'égard du politique :

> Il est intéressant de noter qu'en général cette rétention quant à l'investissement public, va de pair avec une « dépense » dans l'ordre existentiel (jouissance, hédonisme, *carpe diem*, corps, soleil). Alors que dans le bourgeoisisme c'est le contraire que l'on peut observer : frilosité, économie de (et dans) l'existence, et dépense énergétique dans l'ordre public (économie, service public, grandes idéologies motivantes...) qui lui est triomphant (Maffesoli, 1988, p. 76).

Toutefois, si pour Maffesoli cette rétention de l'investissement dans la sphère publique se manifeste essentiellement par la fin de l'individualisme vers une dépense dans l'ordre existentiel hédoniste, mais tribaliste — car l'individualisme serait le propre du politico-moral moderne : l'individu que l'on cherche à protéger de la communauté[21] (*Ibid.*, p. 100) — nous croyons qu'il inaugure au contraire un individualisme dont la forme est nouvelle par rapport à la modernité « classique ». Tribalisme il y a, certes — de *la* morale, c'est de plus en plus *les* morales qui remplissent le paysage éthique de la société —, mais l'individualité est, du point de vue objectif (institutionnel) incontournable, et sa version doctrinaire libérale[22] — morale donc — reste au fondement de la société contemporaine, aussi rationalisée soit-elle. Autrement dit, le tribalisme contemporain n'a rien à voir avec la fin de l'individualisme, il

21. Nous ne sommes pas en désaccord avec cette proposition. L'individualisme moral a en effet tout à voir avec la protection de l'individualité et de sa propriété des dangers de la communauté, surtout chez les contractualistes. Toutefois, nous ne croyons pas que l'on puisse réduire l'existence de l'individualisme à cette stricte manifestation. Cela revient à nier le caractère de plus en plus sociologique (objectif) de l'individualisme contemporain — dont l'existence ne peut être détachée de l'individualisme doctrinaire — que Anthony Giddens, entre autres, contribue à éclairer.
22. Nous avions fait allusion au premier chapitre à la difficulté de détacher complètement l'individualisme sociologique de l'individualisme doctrinaire.

s'agit plutôt d'une forme nouvelle de manifestation de l'individualisme dans un monde rationalisé, où l'individu plus que jamais, devient le *princeps*.

Si les individus s'organisent en tribus, ce n'est pas pour le bien-être de la tribu, mais pour le bien-être individuel : on fait et défait les tribus et les groupes comme bon nous semble (Maffesoli le fait remarquer d'ailleurs). Qui plus est, on choisit d'en faire partie — exclusivité socio-historique moderne qui se renforce définitivement dans le monde contemporain. « La réflexivité de la vie sociale moderne, c'est l'examen et la révision constantes des pratiques sociales, à la lumière des informations nouvelles concernant ces pratiques mêmes, ce qui altère ainsi constitutivement leur caractère. Toutes les formes de vie sociales sont partiellement constituées par la connaissance qu'en ont ses acteurs. » (Giddens, 1994, p. 45) Cette dernière phrase est particulièrement vraie du monde contemporain où les référents moraux et éthiques s'individualisent de plus en plus. Qui plus est, le moi moderne et une certaine dose de réflexivité sont rendues inaliénables de la subjectivité, car constitutives de la création de tout acteur social contemporain (moderne). Qu'on le veuille ou non, paradoxalement !

Giddens traite de façon éclairée, dans ce passage, du statut de la tradition (que l'on peut étendre au tribalisme) dans la société contemporaine en vertu d'une individualité institutionnellement mais aussi subjectivement incontournable.

> Approuver une pratique parce qu'elle est traditionnelle ne suffit pas; on peut justifier la tradition, mais uniquement à la lumière d'un savoir qui n'est pas lui-même authentifié par la tradition. Si l'on ajoute à cela l'inertie résultant de l'habitude, cela veut dire que même dans la plus modernisée des sociétés modernes, la tradition continue de jouer un

> rôle. Mais ce rôle est généralement bien moins signifiant que ne le supposent les auteurs s'intéressant à l'intégration de la tradition à la modernité dans le monde contemporain. Car la tradition légitimée n'est qu'une parodie de tradition, et ne tire son identité que de la réflexivité moderne (*Ibid.*, p. 44-45).

Cette réflexivité moderne prend une place de plus en plus personnalisée dans le monde contemporain. Elle ne s'inscrit pas dans une évaluation morale (de dignité et de vie bonne) de soi par rapport au tout social ou à une communauté virtuelle de citoyens-individus semblables. Elle trouve plutôt son foyer dans l'individualité et la personnalité. L'aspect substantiel et moral qui a motivé les révolutions bourgeoises et démocratiques modernes s'est rationalisé, s'est circonscrit dans le cadre de principes procéduraux et fonctionnels. À la limite pourrions-nous dire que c'est la première catégorie morale de la subjectivité humaniste — le respect d'autrui — qui se circonscrit dans la procèdure (l'innatention polie, les cours de justices), rendant dès lors possible l'appréhension subjective de la dignité et de la vie bonne dans des paramètres de plus en plus personnalisés et dans un climat de sécurité.

Pour Weber, l'effet technique de cette transformation (qui n'a jamais atteint de son temps son envergure actuelle) est une menace directe à la possibilité de l'autonomie subjective par la confrontation à une absence totale de sens et un effet d'aliénation et de subordination à l'appareil bureaucratique — gestionnaire des procédures d'une société complètement intellectualisée, désenchantée et désenchantante. Si ce point de vue est vraisemblable, le théâtre du monde contemporain peut dévoiler un portrait qui ne soit pas aussi fataliste du point de vue de la subjectivité, comme nous l'avons déjà suggéré. Qui plus est, on peut lire la rationalisation croissante de la société

contemporaine à travers les visées subjectives des acteurs sociaux, à travers un enchantement dont la modalité ne doit plus nécessairement reposer sur le collectif et la solidarité. Dans la perspective du collectif, la société est plus que désenchantée, on peut même douter de son existence comme tout. Dans la perspective du sujet toutefois, la société contemporaine rationalisée et technocratique apparaît sous un nouveau jour[23].

Le fait que ce soit la technocratie qui organise, qui régit et qui gouverne ne signifie pas qu'elle aliène. Elle a au contraire une existence légitime dans l'œil du sujet contemporain justement parce qu'elle est conditionnelle à son existence retirée de la sphère publique et ancrée dans le quotidien et l'intimité. Cette légitimité ne peut donc pas être réduite à l'objet d'une programmation démagogique des masses dans la mesure où elle correspond aux moyens nouveaux d'atteindre l'autonomie subjective. De ses moyens, la consommation est certainement parmi les plus importants. « Dans cette perspective, la consommation vise d'abord et avant tout le bien-être des individus; elle se veut la voie idéale pour atteindre le plaisir. Elle engendre [et est le produit aussi[24]] donc un individualisme pur, débarrassé des anciennes valeurs métaphysiques. » (Boisvert, 1995, p. 30)

C'est un peu dans la même optique de Michel de Certeau que nous considérons l'appréhension de la consommation; à la fois comme « produit objectif », mais aussi comme « moyen subjectif ». Sans tenter de faire une

23. Il est intéressant de constater que chez Max Weber ou chez Charles Taylor (dans *Grandeur et misère de la modernité*(1992)) la hantise de la dissipation du collectif et de la solidarité les pousse à considérer une société dont l'enchantement ne reposant pas sur le collectif, offre peu de perspectives pour les individus. L'analyse sociologique que nous mettons de l'avant dans cette analyse tente de réfléchir à ce problème de façon différente, en tentant le moins possible de la teinter de politique.

24. Notre ajout.

analyse atomiste de l'individu, il propose une appréhension de la société contemporaine, et de la société de consommation plus précisément, qui puisse tenir compte de l'appropriation subjective des individus dominés — « ce qui ne veut pas dire passif ou dociles » (de Certeau, 1990, p. xxxvi) — par la rationalisation technologique et par la consommation. Dans cette perspective, la domination capitaliste et rationnelle-légale reste une domination mais ne constitue pas nécessairement la fin de l'autonomie.

> À une production rationalisée, expansionniste autant que centralisée, bruyante et spectaculaire, correspond une *autre* production, qualifiée de « consommation » : celle-ci est rusée, elle est dispersée, mais elle s'insinue partout, silencieuse et quasi invisible, puisqu'elle ne se signale pas en produits propres mais en *manières d'employer* les produits imposés par un ordre économique dominant (*Ibid.*, p. xxxvii).

Dans la continuation de la logique humaniste et révolutionnaire des Lumières, la prise en charge de la dimension du quotidien revient à l'individu. Si la consommation est l'un des lieux privilégiés de cette prise en charge, « l'esprit du capitalisme » comme phénomène généralisé de l'appréhension des besoins quotidiens est aussi important à souligner. L'individualisme nouveau, émancipé, s'est réalisé dans l'arène de l'économie de marché et c'est l'esprit d'entreprise et la raison instrumentale qui sont nécessaires pour accomplir cette tâche sur un terrain de la sorte; l'activité rationnelle en finalité est typiquement celle du capitalisme.

D'un autre côté, la gestion technocratique de la société est la contrepartie idéale de la gestion traditionnelle du monde et elle repose elle aussi sur une appréhension instrumentale du monde en vue de pouvoir d'abord traiter tout un chacun de façon juste, équitable et égalitaire, mais aussi de fournir de la stabilité. Aussi, si les deux constituent les pôles de la figure de la domination

du monde contemporain, à savoir une domination rationnelle-légale et capitaliste, cela n'élimine pas la possibilité de réfléchir à ce phénomène dans la perspective des lieux nouveaux d'autonomie subjective, ni même de créations sociales-historiques de la part du sujet. En sillonnant un peu dans le raisonnement tocquevillien sur l'avènement « prophétique » de l'État tutélaire et de son « despotisme doux », nous voulons proposer la thèse selon laquelle cette technicisation politique est ancrée dans la culture individualiste rationalisée.

Tocqueville faisait remarquer dans *De la démocratie en Amérique,* l'éventualité de deux phénomènes qui pouvaient advenir dans les sociétés capitalistes et démocratiques. D'abord, la disparition de l'ordre et ensuite, l'avènement d'un État tutélaire et la fin des grandes révolutions. Dans la même optique, nous avons proposé que la rationalisation du monde moderne consiste en un double mouvement : l'un de critique de la tradition par la rationalisation des conduites, et l'autre consistant dans l'ancrage des sujets dans le quotidien et la stabilité, surtout par la voie de la rationalité instrumentale — qui, dans le monde moderne s'avère être la forme dominante et croissante de la rationalisation des techniques. Tocqueville avait su se rendre compte, à sa façon, de ces deux pôles relativement aux conséquences de l'égalité, qui, selon lui, était poussée vers deux tendances dont « [...] l'une mène directement les hommes à l'indépendance et peut les pousser tout à coup jusqu'à l'anarchie, [et] l'autre les conduit, par un chemin plus long, plus secret, mais plus sûr, vers la servitude. » (Tocqueville, 1961, t. 2, p. 396,)

Si l'égalité moderne et l'individualisme, issus de l'humanisme, peuvent

entraîner d'une part une abolition de l'ordre dans son ensemble par une révision critique du tout, il y a par ailleurs un danger plus insidieux et moins prévisible qui guetterait l'individu moderne et qui le mettrait dans un état de servitude; l'État tutélaire. Tocqueville qualifie l'État tutélaire d'état de servitude. Mais dans l'approche tocquevillienne, il semble que ce soit moins l'État qui prend les individus sous tutelle, que les individus qui, de par leur culture et leur subjectivité, instituent les conditions favorables à la mise en place d'un État tutélaire.

> Parce que les hommes de la démocratie paraissent toujours émus, incertains, haletants, prêts à changer de volonté et de place, on se figure qu'ils vont abolir tout à coups leurs lois, adopter de nouvelles croyances et prendre de nouvelles mœurs. On ne songe point que, si l'égalité porte les hommes aux changements, elle leur suggère des intérêts et des goûts qui ont besoin de la stabilité pour se satisfaire; elle les pousse, et, en même temps, elle les arrête, elle les aiguillonne et les attache à la terre; elle enflamme leurs désirs et limite leurs forces (*Ibid.*, t. 2, p. 360).

Tocqueville voit clair lorsqu'il postule que la souveraineté individuelle est le principe qui est au centre de l'organisation sociale et de la morale américaine : il parle du patriotisme des Américains comme d'une extension de leur individualisme. Mais cette dimension souveraine de la subjectivité américaine est indissociable de la question de l'entreprenariat. « Les passions qui agitent le plus profondément les Américains sont des passions commerciales et non des passions politiques, ou plutôt ils transportent dans le politique des habitudes du négoce. » (*Ibid.*, t. 1, p. 423) Nous avons tenté de l'insinuer plus haut; le type idéal du sujet de la rationalisation du monde moderne incite le développement de l'autonomie de l'individu dans les sphères de l'économie de marché. Et quoiqu'il persiste une dimension morale

de l'individualisme aux États-Unis à cette époque[25], les conditions de sa rationalisation sont beaucoup plus avancées qu'ailleurs à cause de la providence qui a fait naître une nation avec des individus « égaux *a priori* » sans qu'il y ait d'opposition farouche à l'individualisme.

Dans cette perspective, la politique américaine n'est déjà plus une fin en soi dans l'esprit des gens (si on la compare à ce qui se passe dans la France révolutionnaire du XIXe siècle), elle est un moyen, voire même un « mal nécessaire » et elle prend déjà l'apparence d'une organisation technique pouvant assurer la fonctionnalité du système et prévenir les débordements. L'individu américain « obéit à la société, parce que l'union avec ses semblables lui paraît utile et qu'il sait que cette union ne peut exister sans un pouvoir régulateur. » (*Ibid.*, t. 1, p. 118) Une allusion au pouvoir régulateur qui se rapproche des systèmes de confiance dont parle Giddens : objets de résignation mais aussi de confiance.

Dans cette optique, l'État tutélaire tocquevillien qui prend la destiné du quotidien et la stabilité socio-politique en charge, et qui s'apparente, par ses effets, à la scientificisation de la politique et à la gestion technicisée et instrumentale de la société, ne peut être réduit à une stratégie démagogique visant à asservir les individus. L'État tutélaire s'inscrit dans la mouvance de la rationalisation propre au monde moderne et aux visées subjectives des

25. En fonction du puritanisme protestant, dont les mœurs et les valeurs sont partagées par la plupart des citoyens et qui constitue une des grande bases de l'ordre social américain. Mais aussi en fonction de la doctrine de l'intérêt bien entendu, valeur morale unanimement admise chez les Américains, selon Tocqueville, qui permet de faire concorder les intérêts individuels avec ceux de la collectivité (chapitre VIII, tome 2). Dans cette perspective, la stricte fonctionnalité du système a une valeur morale. À l'époque contemporaine, quoique les finalités soient les mêmes, soit la fonctionnalité du système, nous supposons que sa dimension morale se soit dissipée parce qu'elle est définitivement vidée de toute doctrine, contrairement à l'intérêt bien entendu qui reste substantiel, doctrinaire et moral.

individus, celle de l'ancrage du sens de leur vie dans la stabilité du quotidien, dans le respect de leur intégrité et de leur dignité. Or, c'est un quotidien auréolé de la gestion rationnelle-légale de la société dans son ensemble, et dont les pôles d'activité sont situés entre l'activité économique, dont la consommation et l'entreprenariat surtout, et une multitudes d'autres phénomènes dont les répercussions sont à des échelles réduites et personnalisées.

Aussi, les éthiques et les esthétiques individuelles se multiplient et cohabitent sans que le visage global de la société ne change. Les relations à autrui se limitent à une reconnaissance formelle de sa différence et sont contenues dans le paradigme de la performativité du système. Ce n'est plus le politique qui prend en charge la destinée et l'intégration de groupes revendicateurs, mais une instance rationnelle de gestion des litiges et de la moralité : la justice. Autrui fait l'objet de nos communications, de nos relations et de nos jugements mais peut tout aussi bien faire l'objet de l'indifférence. Les relations contemporaines sont laissées au bon vouloir des participants dans la mesure où il leur appartient de déterminer le lieu de leur épanouissement. La confrontation de points de vus moraux entre des groupes de citoyens aux allégeances contradictoires n'est même plus l'objet des délibérations qui ont lieu dans les instances politique. Ce sont des instances bureaucratiques, les cours de justice, qui prennent en charge la résolution de ces conflits. Les injustices du système ne sont plus dénoncées en fonction d'une justice universelle mais en fonction de contraintes locales et les doléances visent, en dernière analyse, l'intégration où l'accroissement de l'efficacité du système pour laisser libre cours aux visées subjectives individualisées. Nonobstant des influences multiples du politique sur les nominations des

magistrats ou sur l'orientation des décisions, les citoyens occidentaux contemporains s'en remettent à ces instances fonctionnelles pour faire valoir leurs différends.

Certes, l'insurrection contre l'injustice existe toujours, mais elle prend moins la forme d'une prise en charge de la collectivité pour le bien-être de tout un chacun dans l'optique d'une vision idéale de la société que de luttes pour la reconnaissance localisées et personnalisées qui visent l'intégration. Par ailleurs, si certains individus supportent des causes ou des luttes sociales précises au nom de la justice, on peut supposer que l'espace entre ce support abstrait et une intervention concrète peut être très éloigné; on s'en remet généralement en la confiance aux recours légaux et à la rationalité du système.

Excursus. La question des droits « des homosexuels » est très intéressante à l'effet de son rapport avec les deux caractéristiques de la rationalisation — celle des conduites, critique et interne, et celle des techniques, fonctionnelle et externe — et de la remarque que nous venons de faire sur les visées personnelles de l'organisation identitaire et/ou en groupe. À savoir, deux attitudes parallèles du sujet à l'égard du monde : la critique des formes traditionnelles de légitimité en tant que contestation et résistance; et la consolidation d'une attitude instrumentale à l'égard du monde en général qui a comme fondement l'individualisme.

La lutte de ce groupe pour la reconnaissance est toujours en cours à l'aube du XXIe siècle et ce n'est que tout récemment, au Canada, que leur ont été accordés des droits civils similaires à ceux des hétérosexuels en ce qui concerne

le transfert des pensions aux conjoints. Cette décision de la plus haute instance judiciaire canadienne est d'une importance sans pareille dans les efforts de certains individus à être reconnus comme des citoyens à part entière et constitue un signe de l'effondrement de certaines attitudes traditionnelles et discriminatoires qui sont institutionnalisées à leur égard. Toutefois, on ne peut pas dire des homosexuels qu'ils constituent un groupe révolutionnaire, et encore moins anti-individualiste. Ils vont à contre-courant de certaines attitudes traditionnelles mais leurs visées[26] sont au niveau de l'intégration de la société telle qu'elle est, dans sa fonctionnalité rationnelle-légale et capitaliste. Dans cette perspective, ce mouvement correspond aux deux manifestations de la rationalisation de l'individualisme. Sous fond d'arguments sur la liberté, la dignité et le droit, ce groupe fait une critique virulente de la tradition mais il cherche aussi à ce que l'individualité et l'autonomie de ses membres puisse s'exercer sur le terrain fonctionnel de la société bureaucratique et capitaliste contemporaine. Or, de façon générale, les résistances à l'égard de leurs doléances sont circonscrites dans des poches de groupes de droite et de façon générale font l'objet, sinon de la sympathie, de l'indifférence[27].

26. Nous ne supposons surtout pas que les visées « des homosexuels » soient toutes les mêmes. Ni encore que la catégorisation du groupe « homosexuels » constitue une catégorie en soi — ce qui constitue un autre ordre d'analyse. L'image médiatisée du groupe de pression caractérisé comme « les homosexuels » reflète toutefois l'image de ce que nous mettons de l'avant : un contre-mouvement de la tradition conformiste quant à l'ordre rationnel-légal contemporain.

27. Stockwell Day, le chef de la droite canadienne — qui prend des allures de plus en plus Républicaines (au sens américain) et de moins en moins conservatrices (au sens canadien) — malgré ses oppositions personnelles morales à l'homosexualité — il est d'allégeance chrétienne traditionnelle — a postulé qu'il n'entâmerait pas de lutte politique contre les droits qu'auraient acquis les homosexuels. Cela révèle selon nous le fait que la politique se joue de moins en moins sur une dimension morale. Par ailleurs, cela révèle que la rationalisation de la morale humaniste d'intégration, de justice, d'égalité et d'équité est à toute fin pratique difficile à déloger et pourrait avoir des conséquences désastreuses sur le plan politique d'un candidat. On pourrait supposer que la dimension morale de la justice sociale contemporaine ne réapparaîtrait que si elle était compromise pour l'ensemble de la population. En attendant, son caractère moral est latent et n'a de visage que dans son appréhension fonctionnelle et procédurale.

Ainsi, l'individualisme, dans sa forme contemporaine, est immédiatement influencée par les catégories morales de la subjectivité humaniste, mais sa manifestation concrète et les bases sur lesquelles elle repose ne sont pas morales. Les catégories morales des Lumières se sont rationalisées et ont été intégrées à une logique de la performativité et de la fonctionnalité. La moralité humaniste s'est sédimentée et s'est transformée en contenant de moralités et, à l'occasions des litiges, en gestionnaire des moralités. Ce passage de *L'ère du vide* de Gilles Lipovetsky illustre bien le virage local et personnel que l'individualisme contemporain à pris en fonction de ce que nous avons présenté comme la dynamique de la rationalisation de la subjectivité humaniste.

> L'individualisme disciplinaire et militant, héroïque et moralisant a été relayé par un individualisme *à la carte*, hédoniste et psychologique, faisant de l'accomplissement intime la fin principale des existences. Déroute des grands projets de société, érosion des identités sociales et des normes coercitives, culte de libre disposition de soi-même dans la famille, la religion, la sexualité, le sport, la mode, l'adhésion politique et syndicale, la seconde révolution individualiste est celle qui a concrétisé, dans la vie quotidienne, l'idéal libéral de l'autogouvernement de soi, là où autrefois les valeurs et institutions travaillaient à en conjurer le déploiement. La coupure historique néo-individualiste signifiait au plus profond la généralisation et l'explosion des aspirations à l'autonomie subjective dans toutes les couches sociales, dans les diverses catégories d'âge et de sexe. S'est ainsi imposée la figure d'un individualisme post-moderne dégagé de l'emprise des idéaux collectifs autant que du rigorisme éducatif, familialiste, sexuel (Lipovetsky, 1993, p. 316).

L'individualisme ne subsiste donc plus comme une valeur morale partagée dans des visée collectives de progrès et de solidarité. Par les effets de la rationalisation de la quête subjective d'accès à l'autonomie et à la liberté, il se manifeste dans le monde contemporain comme condition objective d'une société dont les fondements sont contenus dans les limites d'un espace

stratégiquement préservé en système fonctionnel. Mais si l'individualisme n'est plus un choix, en fonction des impératifs institutionnels des sociétés et des États occidentaux, il n'en demeure pas moins que cette imposition n'a pas le visage de la tyrannie. Au contraire, c'est dans la mouvance des quêtes subjectives d'autonomie, de liberté et d'accès au sens et à la dignité — étroitement liées à l'existence réflexive du moi — que la domination rationnelle-légale et capitaliste de la société peut s'instituer et que la valeur morale de l'individualisme peut estomper légitimement. Dans cette perspective, le politique tel qu'il était traditionnellement appréhendé par les acteurs sociaux, comme par les intellectuels, fait l'objet de grandes transformations. La société elle aussi, puisque la nature de ce qui la définit et la constitue, le lien social, prend dans le monde contemporain une forme inédite dans l'histoire. En fait, la nature du lien social contemporain semble n'exister justement, que dans une forme.

Conclusion

Pour être menée à terme cette analyse devait silloner à travers des thèmes variés. Nombreuses sont les contributions, dans la littérature sociologique, aux questions de l'individualisme, de la subjectivité, de la rationalisation ou de la société contemporaine. Or, les développements de cette analyse ne prétendent surtout pas les avoir directement ou indirectement toutes adressées. Dans cette optique, la portée de cette analyse est limitée et laisse plusieurs portes entrouvertes pour des précisions et des discussions ultérieures. Elle propose toutefois une lecture différente des thèmes qu'elle aborde en vertu de leur agencement dans une méthode particulière à la lumière de points d'intérêts spécifiques; à savoir l'appréhension par le sujet du caractère formel et procédural du lien social contemporain par rapport au lien moral l'ayant précédé.

Notre lecture de la subjectivité, surtout constituée à partir d'une discussion des thèses de Charles Taylor, fut à la base de nos développements. Le sujet moderne, par l'existence de son moi et par sa quête de sens et de dignité individualisée dans le but d'atteindre l'autonomie subjective, est le lieu ultime de la détermination des jugements qu'il porte sur le monde. La thèse d'inspiration taylorienne sur la conjonction de la morale et de l'éthique dans le cadre phénoménologique des activités sociales a été le point d'ancrage du contraste que nous avons cherché à soulever entre la société moderne des XVIIIe et XIXe siècles et son héritière contemporaine. Aussi, ce chapitre nous a permis d'avancer des thèses sur l'autonomie et l'aliénation qui situaient leur lieu d'existence dans la subjectivité en vertu de ce que nous avons appelé l'incarnation subjective. De ce point de vue, nous pouvions mieux être en

mesure de comprendre le rapport entre l'individu et la société à partir de la subjectivité. Selon nous, le jugement moral contemporain n'a plus de visée (subjective) et encore moins d'effet (en vertu des conditions objectives), dans un espace qui comprend la collectivité, le Tout social.

Alors que dans les grandes tendances sociales historiques de l'organisation du collectif antérieures au monde contemporain, le rapport à l'autre est circonscrit dans une série de jugements substantiels (sur le bon citoyen, sur la race, sur le sexe, sur la famille), dans la version contemporaine du lien social, ce rapport n'est presque plus imprégné de jugements. S'il l'est, sa portée n'a toutefois pas de répercussions énormes au niveau du collectif, dans la mesure où le rapport à l'autre est circonscrit dans des procédures rationalisées qui bénéficient d'une autonomie propre. Le jugement moral dont la visée est collective, s'il en est, dans le monde contemporain, se frappe à l'écueil que constituent les systèmes experts et la gestion technocratique de la socialité.

C'est dans la mesure où le rapport à l'autre n'est plus immédiatement[28] circonscrit dans un rapport de jugement substantiel, mais plutôt formel et procédural (dans une logique technicisée autonomisée) que nous avons postulé que l'individualisme contemporain n'est plus contenu dans une conception morale de l'individu-citoyen. Or, nous avons cherché à appréhender cette problématique à partir d'une subjectivité idéale-typique dont nous avons inscrit le développement au sein du processus de rationalisation de la modernité. Nous cherchions à cet effet à nous interroger sur la nature subjective de la

28. Nous écrivons « immédiatement » à la lumière d'une remarque que nous avions précédemment faite dans une autre note. Nous y mettions de l'avant l'hypothèse selon laquelle le jugement moral du monde contemporain, dans le rapport à l'autre, était latent.

disparition du caractère moral de l'individualisme.

Au premier regard, la domination technocratique contemporaine peut sembler se limiter à une cage de fer. Elle se révèle en effet en condition objective incontournable du système qui accapare l'espace d'intervention subjectif de contraintes externes. La technocratie, les systèmes experts délocalisés et le caractère individualisant des institutions contemporaines émettent sur les sujets des cadres objectifs presque incontournables. Il est difficile de nier ce fait sociologique. Or, une lecture de ce phénomène structurel et institutionnel qui explique sa provenance dans les limites du pouvoir politique et capitaliste — par la programmation des masses ou la démagogie des discours — n'était pas selon nous en mesure de saisir un portrait juste de la subjectivité contemporaine et de la manifestation de son individualisme. Nous pensions plutôt que la rationalisation contemporaine des techniques s'était mise en branle en conjonction avec une rationalisation de la subjectivité concordante.

Ce qui apparaissait au yeux de Max Weber comme une aliénation en vertu du désenchantement issu de l'autonomisation des techniques de gestion rationnelles de la socialité, pouvait apparaître, sous un jour nouveau, comme un glissement des lieux d'acquisition d'autonomie subjective. Or, nous avons proposé que cette nouvelle forme d'autonomie se réalisait dans les cadres structurels de la domination technocratique contemporaine. Dans nos vues, cette domination pouvait être comprise non seulement comme la cause d'une subjectivité nouvelle qui se dessine — qui dans ces strictes limites prend des allures d'aliénation objective — mais aussi comme l'effet de cette dernière —

nous permettant de la réfléchir plus facilement en fonction de l'autonomie subjective. La domination de la société technocratique, dans nos vues, prend moins le visage de la tyrannie que celui de l'affinité sociale-historique entre des visées subjectives transformées et l'hyper-rationalisation technocratique de la domination rationnelle-légale moderne.

Ainsi, la disparition de la légitimation du lien social sur la base de jugements moraux intersubjectivement partagés peut être compris comme l'effet d'une rationalisation de la subjectivité. Le sujet contemporain, plus hédoniste et très axé sur son quotidien, s'approprie les effets de domination de la consommation, de l'entreprenariat et des sytèmes experts en vue de se réaliser dans un espace dont les autres ne doivent pas nécessairement être partie prenante. Qui plus est, comme nous avons tenté de le démontrer dans notre chapitre sur la rationalisation, de même que dans nos interprétations des thèses de Tocqueville entre autres, on ne peut exclure le fait que cette appropriation subjective des effets de la domination contemporaine ne soit aussi une création subjective de conditions sociales favorables à une réalisation subjective différente.

Dans l'optique de la mise de côté graduelle des grands récits supra-individuels de la légitimation du lien social au moment des révolutions culturelles du milieu du XXe siècle, le lieu de l'enchantement contemporain passe du collectif au moi, ou à la communauté tribale[29]. Des trois catégories morales du sujet humaniste, le respect d'autrui est dès lors contenu dans des procédures, alors que la vie bonne et la dignité n'ont de référents que dans les

29. Et encore faut-il noter, comme nous l'avons fait, le caractère très individualiste de l'association tribale contemporaine.

limites de ce que le sujet reconnaît lui-même comme étant valide. Moralité post-conventionnelle ? Le terme est équivoque dans la mesure où, justement, l'idée du dépassement de la convention fait référence à une organisation politique qui se dessine non pas dans les sphères d'une conception morale du monde, mais d'une conception rationnelle et instrumentale ayant sa logique propre et qui est indépendante de tous jugements qualitatifs subjectifs.

Le monde moderne pré-contemporain est le théâtre d'une délocalisation de la morale en tant qu'elle régit les rapports intersubjectifs des individus à travers des jugements sur la valeur de l'autre, sa dignité, la conduite de sa vie, etc. La morale, de par le rôle qu'elle joue à ce moment de l'histoire, peut donc être comprise comme ayant une valeur universelle. On peut dans cette optique la comprendre séparément de l'éthique dans la mesure où elle la régit sous des termes semblables et subjctivement incarnés, pour tout un chacun. L'éthique moderne pré-contemporaine est celle qu'Émile Durkheim reconnaissait dans les associations professionnelles et les divers sous-groupes, toujours subordonnés à la conscience collective. Mais elle n'est ainsi que parce que conscience collective il y a.

Dans cette perspective, la dichotomie habermassienne de la morale et de l'éthique tient. Or, elle ne tient que dans un contexte social-historique précis. Dans une perspective ontologique cela ne révèle en rien une vérité sur le statut de la morale comme transcendance de l'éthique, mais plutôt une subjectivité différente, en l'occurence rationalisée. La morale contemporaine est localisée, hyper-localisée, et elle se multiplie en moralités dans la mesure où ses référents ne sont plus à l'égard d'autrui et du Tout social mais d'individualités. La

personnalisation de la morale implique une explosion des identités qui se manifestent par leur désir de reconnaissance. Or cette quête de reconnaissance n'est pas évaluée par le système sur des fondements moraux mais à travers une logique procédurale autonomisée. Toute approche moralisante et universalisante, dans le monde contemporain, est engloutie dans et par les systèmes experts et les procédures rationnelles-légales désubstancialisées et délocalisées.

Cette analyse s'est limitée à vouloir faire un constat sociologique. Elle a cherché volontairement à ne pas s'aventurer au niveau politique. Le point de vue politique peut poser problème dans l'approche sociologique. C'est d'ailleurs une préoccupation qui revient toujours comme hantise de l'analyse sociologique; celui que Max Weber avait soulevé entre le jugement de fait et le jugement de valeur. Mais c'est peut-être au niveau du politique que la société contemporaine peut donner lieu à une situation angoissante. Les conjectures sur le futur sont très délicates. Néanmoins, il semble que le monde contemporain entre graduellement et plus que jamais dans une ère de stabilité où, comme Tocqueville avait su le remarquer, les grandes révolutions risquent de se faire rares.

La nature du lien social contemporain n'a d'ailleurs rien de révolutionnaire par rapport à celle du monde moderne. En laissant de plus en plus de place à l'intégration et à l'individualité, le mode de production n'en reste pas moins capitaliste et la science, quoiqu'elle ne joue plus le rôle de la légitimation morale du Tout social, est quand même plus que jamais, dans sa version technologique et technocratique, au centre du politique. Le visage de la

société contemporaine est en fait celui de la rationalisation dans la mesure où l'on peut comprendre et expliquer vraisemblablement ses différences par rapport aux époques qui le précèdent en fonction d'un raffinement inédit, entre autres, des techniques et des conduites. Le premier par la technocratie et le second à travers la venue d'un individu qui existe de plus en plus comme *princeps*. Or, c'est en vertu de la stabilité, de la performance et de la consistance dans un cadre procédural et rationnel-légal mise au monde par cette socialité que les paramètres de l'interrogation politique sur le monde deviennent ambigus et angoissantes. On dirait que ni les structures sociales ni la subjectivité contemporaine ne sont en mesure d'être compatibles avec une réorientation[30] proprement politique du monde; son orientation étant déjà contenue dans une logique qui ne lui appartient pas. Dans ces circonstances, il semble que parmi les candidats les plus sérieux comme causes du changement social, la catastrophe écologique ou nucléaire font bonne figure.

30. Vers où ? La question reste ouverte et appartient justement au domaine du politique.

Bibliographie et références

Apel, Karl-Otto, *La sémiotique transcendantale et les paradigmes de la Prima philosophia* IN « Revue de métaphysique et de Morale », 92, 2, 1987, p. 147 - 163.

Birnbaum, Pierre et Leca, Jean, *Présentation* IN « Sur l'individualisme », Birnbaum, Pierre et Leca, Jean (dir.), Paris, Presses de la Fondation nationale des sciences politiques, 1986, p. 11 - 24.

Boisvert, Yves, *Le postmodernisme*, Canada, Boréal, 1995.

de Certeau, Michel, *L'invention du quotidien : 1. arts de faire*, Paris, Gallimard, 1990.

de Courville Nicol, Valérie,

— *La production de « l'homme » moderne : le passage de la peur à l'intérieur*, Manuscrit, 2000.

— *Le soupçon gothique*, Thèse de Doctorat, Ottawa, Université Carleton, 1999.

de Tocqueville, Alexis, *De la démocratie en Amérique*, tomes 1 et 2, Paris, Gallimard, 1961.

Durkheim, Émile,

— *L'éducation morale*, Paris, Presses universitaires de France, 1963.

— *Le çons de sociologie : Physiqe des mœurs et du droit*, Paris, Presses universitaires de France, 1950.

Freund, Julien, *Études sur Max Weber*, Genève-Paris, Librairie Droz, 1990.

Giddens, Anthony,

— *Les conséquences de la modernité*, Paris, L'Harmattan, 1994.

— *Capitalism and Modern Social Theory*, Cambridge, Cambridge University Press, 1971.

Habermas, Jürgen,

— *Notes programmatique pour fonder en raison une éthique de la discussion* IN « Morale et communication », Paris, Éditions du Cerf, 1996, p. 63 - 130.

— *De l'Éthique de la discussion*, Paris, Les Éditions du Cerf, 1992.

— *La technique et la science comme « idéologie »* IN « La technique et la science comme « idéologie » », Paris, Gallimard, 1973, p. 1 - 74.

— *La scientificisation de la politique et opinion publique* IN « La technique et la science comme « idéologie » », Paris, Gallimard, 1973a, p. 97 - 132.

Lasch, Christopher, *The Culture of Narcissism*, New-York, Warner Books Edition, 1979.

Laurent, Alain, *Histoire de l'individualisme*, Paris, Presses universitaires de France, 1993.

Lipovetsky, Gilles, *L'ère du vide : Essais sur l'individualisme contemporain*, Paris, Gallimard, 1983 (1993 pour la postface).

Lyotard, Jean-François, *La condition postmoderne*, Paris, Les Éditions de Minuit, 1979.

Maffesoli, Michel, *Le temps des tribus : Le déclin de l'individualisme dans les sociétés de masse*, France, Librairie des Méridiens, 1988.

Marcuse, Herbert, *L'homme unidimensionnel*, Paris, Les Éditions de Minuit, 1968.

Miguelez, Roberto, *L'émergence de la sociologie*, Ottawa, Presses de l'Université d'Ottawa, 1993.

Mongardini, Carlo, *Max Weber et les contradictions de la modernité* IN « Max Weber et le destin des sociétés modernes », Pierre million (dir.), Grenoble, rehcerches sur la philosophie du language, 1975, p. 89 - 103.

Poole, Ross, *Morality and Modernity*, London-New-York, Routledge, 1991.

Raynaud, Philippe, *Max Weber et les dilemmes de la raison moderne*, Paris, Presses universitaires de France, 1987.

Renaut, Alain, *L'avènement de l'individu comme dissolution du sujet* IN « Individualismes et individualités », Côté, Jean-François (dir.), Sillery, Les éditions du Septentrion, 1995, p. 27 - 56.

Schnapper, Dominique, *La relation à l'autre : Au cœur de la pensée sociologique*, Paris, Gallimard, 1998.

Taylor, Charles,

— *Les sources du moi : La formation de l'identité moderne*, Canada, Boréal, 1998.

— *Multiculturalisme : différence et démocratie*, Paris, Aubier, 1994.

— *Grandeur et misère de la modernité*, Canada, Bellarmin, 1992.

Villey, Michel, *Préface : La doctrine du droit dans l'histoire de la science juridique* IN Emmanuel Kant, « Métaphysique des mœurs », Première partie : Doctrine du droit, Paris, Vrin, 1971, p. 7 - 26.

Weber, Max,

— *The Nature of Charismatic Domination* IN « Weber : Selections in Translations », W. G. Runciman (dir.), Cambridge, Cambridge University Press, 1995, p. 226 - 250.

— *Socialism* IN « Weber : Selections in Translations », W. G. Runciman (dir.), Cambridge, Cambridge University Press, 1995a, p. 251 - 262.

MQ

66048

UMI

MICROFILMED 2002

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

PERMISSION DE REPRODUIRE ET DE DISTRIBUER LA THÈSE

PERMISSION TO REPRODUCE AND DISTRIBUTE THE THESIS

NOM DE L'AUTEUR / NAME OF AUTHOR:	HASSAN, Husam R.
ADRESSE POSTALE / MAILING ADDRESS:	1700D Lamoureux Dr. Orléans, Ontario K1E 2N2
GRADE / DEGREE:	**ANNÉE D'OBTENTION / YEAR GRANTED**
M.A.SC. ELECTRICAL ENGINEERING	2001
TITRE DE LA THÈSE / TITLE OF THESIS: OPERATIONAL (IN-FIELD) CALIBRATION METHODOLOGY OF VIDEO QUALITY EVALUATION OF MPEG2 COMPRESSED DATA FOR BROADCAST.	

L'auteur permet, par la présente, la consultation et le prêt de cette thèse en conformité avec les règlements établis par le bibliothécaire en chef de l'Université d'Ottawa. L'auteur autorise aussi l'Université d'Ottawa, ses sumccesseurs et cessionnaires, à reproduire cet exemplaire par photograhie ou photocopie pour fins de prêt ou de vente au prix coûtant aux bibliothèques ou aux chercheurs qui en feront la demande.

Les droits de publication par tout autre moyen et pour vente au public demeureront la propriété de l'auteur de la thèse sous réserve des règlements de l'Université d'Ottawa en matière de publication de thèses.

The author hereby permits the consultation and the lending of this thesis pursuant to the regulations established by the Chief Librarian of the University of Ottawa. The author also authorizes the University of Ottawa, its successors and assignees, to make reproductions of this copy by photographic means or by photocopying and to lend or sell such reproductions at cost to libraries and to scholars requesting them.

The right to publish the thesis by other means and to sell it to the public is reserved to the author, subject to the regulations of the University of Ottawa governing the publication of theses.

N.B. LE MASCULIN COMPREND ÉGALEMENT LE FÉMININ

30/Nov./2001

DATE

Husam Hassan

(AUTEUR) SIGNATURE (AUTHOR)



Université d'Ottawa · University of Ottawa



Université d'Ottawa • University of Ottawa

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

HASSAN, Husam R.

AUTEUR DE LA THÈSE - AUTHOR OF THESIS

M.A.Sc. (Electrical Engineering)

GRADE - DEGREE

School of Information Technology and Engineering

FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

Operational (In-Field) Calibration Methodology of Video Quality Evaluation of MPEG2 Compressed Data for Broadcast

Emil Petriu

DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

V. Groza

R. Hafez

J.-M. De Koninck, Ph.D.

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

SIGNATURE

DEAN OF THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

Operational (In-Field) Calibration Methodology of Video Quality Evaluation of MPEG2 Compressed Data for Broadcast

By

Husam R. Hassan

A thesis submitted to the

Faculty of Graduate And Postdoctoral Studies

In partial fulfillment of the requirements for the degree of

Master of Applied Science

In

Electrical and Computer Engineering

Supervised by:

Prof. Emil Petriu

Ottawa-Carleton Institute for Electrical and Computer Engineering

School of Information Technology and Engineering

Faculty of Engineering

University of Ottawa

September 2001

© Husam R. Hassan, Ottawa, Canada

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-66048-6

Canada

Contents

		Page
i-	List of Figures	v
ii-	List of Tables	vii
iii-	Abstract	viii
iv-	Acknowledgement	ix
1-	Introduction	1
	1.1 The TV Network Problem	2
	1.2 Thesis Organization and Contribution	3
2-	Background on Video Compression & MPEG-2	5
	2.1 Entropy Coding	5
	2.2 Spatial Coding	9
	2.3 Discrete Cosine Transform (DCT)	9
	2.4 Quantization (Weighting)	11
	2.5 Scanning	12
	2.6 I, P & B Frames	12
	2.7 Profiles & Levels	13
	2.8 Macroblocks	14
	2.9 Slices	15
	2.10 MPEG-2 Transport Stream	16
3-	Digital TV and Video Standards	20
	3.1 History of TV	20
	3.2 Advanced Television Standards Committee (ATSC)	22
	3.3 Digital Video Broadcast (DVB)	23
	3.3.1 Coded Orthogonal Frequency Division Multiplexing	27
	3.3.2 Uncoded OFDM	28

3.4 Digital Video Tape Standard 30

4- Basics of Video Quality Measurement 35

4.1 The Human Visual System (HVS) 35

4.2 Video Artifacts 44

4.2.1 Analog Video Artifacts 44

4.2.2 Digital Testing Concepts 46

4.2.3 Digital Signal Artifacts 47

4.3 The ITU-R.BT-500 standard and Subjective Quality 52

4.4 Objective Quality Measurement 57

4.4.1 Picture Comparison Method 57

4.4.2 Feature Extraction method 58

4.4.3 Single Ended testing Method 58

4.5 Quality Metrics 59

4.5.1 Metrics Based on the Human Visual System 60

4.5.2 Metrics Based on Visual Artifacts 62

5- Algorithms of Video Quality 63

5.1 Double Stimulus Systems (Full and Reduced Reference FR/RR) 63

5.1.1 JND algorithm, Sarnoff, USA 63

5.1.2 The NTIA model, USA 68

5.1.3 Image Segmentation model, CPqD, Brazil 71

5.1.4 DVQ Model, NASA, USA 74

5.1.5 The PDM Metric, EPFL, Switzerland 76

5.1.6 Criticality Model, NHK/Mitsubishi, Japan 77

5.1.7 KDD/Pixelmetrix Model, Japan 80

5.1.8 The KPN/Swisscom CT model, Netherlands 82

5.1.9 Tapestries Model, UK 83

5.2 Single Stimulus Systems (No Reference NR) 84

5.2.1 IFN/R&S Algorithm 84

6 Calibration of Objective to Subjective values 88

6.1 The VQEG work 88

6.1.1 Objective Model Evaluation Criteria 89

6.2 Evaluation Metrics 90

6.2.1 Metrics relating to Prediction Accuracy 90

6.2.2 Metrics relating to Prediction Monotonicity 92

6.2.3 Metrics relating to Prediction Consistency 93

6.3 The VQEG test and conclusion 94

6.4 The Suggested In-Field Calibration Standard 95

6.4.1 Subjective Measurement of quality for video 97

6.4.2 Objective Testing 99

6.5 Practical Experiment 101

6.5.1 Subjective Evaluation 101

6.6 The test setup 103

6.7 The test results and analysis 105

7 Conclusion and Future Work 112

8 References 115

9 Acronyms and Glossary 122

10 Appendix A 133

i- List of Figures

Figure 2.1	Elements of Video Information	7
Figure 2.2	Scene Complexity	8
Figure 2.3	The MPEG basis function and a DCT	10
Figure 2.4	Scanning Techniques	11
Figure 2.5	I,P and B frames in a group of Pictures Structure	13
Figure 2.6	MPEG-2 Levels and Profiles	14
Figure 2.7	The MPEG-2 188 byte packet	17
Figure 2.8	The MPEG-2 transport stream	19
Figure 3.1	Multimedia Interfaces in the Home using DVB	25
Figure 3.2	The DVB transport stream	25
Figure 3.3	DVB Implementation Worldwide	30
Figure 4.1	Anatomy of the Human Eye	35
Figure 4.2	The Rods and Cones cells	36
Figure 4.3	The Human Visual System	37
Figure 4.4	The Spectrum of Electromagnetic waves including Visible light	38
Figure 4.5	Relative Constant Sensitivity of the HVS with spatial frequency	40
Figure 4.6	Variation of the HVS contrast Sensitivity to temporal Frequency	41
Figure 4.7	Threshold Elevation as a function of local contrast	42
Figure 4.8	Luminance sensitivity as a function of Grey Value	43
Figure 4.9	Analog Systems Artifacts	45
Figure 4.10	Bit Rate Reduction Impairments according to CCIR1089	50
Figure 4.11	Original Picture (Susie)	47

Figure 4.12	Original Indian before compression, with multiple errors	48
Figure 4.13	Edge noise effect	48
Figure 4.14	Blocking oe tiling effect to the DCT.	49
Figure 4.15	Aliasing Effect	50
Figure 4.16	Blurring Effect	51
Figure 4.17	The slider mechanism evaluation device	56
Figure 4.18	Picture Comparison or Full Reference System	57
Figure 4.19	Feature Extraction or reduced Reference System	58
Figure 4.20	Single Stimulus or No Reference System	59
Figure 5.1	Block diagram of JND model	64
Figure 5.2	JND model Architecture	66
Figure 5.3a	Reference Image	67
Figure 5.3b	Processed Image	67
Figure 5.3c	The difference map or JND map	68
Figure 5.4	The extracting feature approach for quality measurement	69
Figure 5.5	Sample spatial-temporal region (S-T)	70
Figure 5.6	Objective Parameters Computation	71
Figure 5.7a	Part of the Mobile and Calendar	73
Figure 5.7b	Result of segmentation	73
Figure 5.8	Overview of the DVQ Algorithm processing steps	74
Figure 5.9	Block diagram of Perceptual Distortion metric	76
Figure 5.10	Configuration of criticality measurement	77
Figure 5.11	Procedure to derive picture quality distribution characteristics	78
Figure 5.12	The 3 layered model for quality evaluation (KDD)	79
Figure 5.13a	The KDD model	79

Figure 5.13b	The 3 layer model of noise weighting	80
Figure 5.14	Sample calculation process for the amplitude difference of adjacent pixels	84
Figure 5.15	Average pixel amplitude differences	85
Figure 5.16	Averaged pixel amplitude difference	86
Figure 6.1	Relation between objective and subjective Measurement	89
Figure 6.2	Model with greater accuracy	90
Figure 6.3	Model with lower accuracy	90
Figure 6.4	Model with more Monotonicity	91
Figure 6.5	Model with less Monotonicity	92
Figure 6.6	Model with large outlying errors	93
Figure 6.7	Model with consistent errors	93
Figure 6.8	Single ended system's evaluation,	98
Figure 6.9	Double Ended system's evaluation	99
Figure 6.10	Test Setup for the Subjective/objective calibration	103

ii- List of Tables:

Table 3.1	Popular ATSC picture formats	23
Table 4.1	Quality and Impairment scales	54
Table 4.2	Comparison Scale	55
Table 6.1	525/60 Hz format sequences	101
Table 6.2	Test Conditions (HRCs)	101

iii- Abstract:

This Thesis discusses the different techniques used in subjective and objective measurement for MPEG-2 compressed video for broadcast according to the ITU-R BT-500 recommendation for both world standards ATSC and DVB.

The technique for subjective/objective quality correlation described in ITU recommendations and VQEG work [VQEG00] for Double Stimulus Continuous Quality Scale (DSCQS) & Single Stimulus Continuous Quality Evaluation (SSCQE) is used to generalize an application for In-Field operational calibration standard of subjective / objective digital video evaluation for TV network and broadcast stations.

A practical experiment for validating the correlation between the subjective evaluations of some MPEG2 coded streams vs. the objective measurement obtained using the IFN / Rohde & Schwarz single stimulus blockiness measurement algorithm is realized, demonstrating the correlation to other measurements and the ease and applicability to perform this in the field.

iv- Acknowledgement

I would like to express my deep thanks and gratitude to Prof. Dr. Emil M. Petriu for his guidance and support throughout the full course of my Masters studies and Thesis.

I also would like to express my deep thanks and appreciation to the team of the Digital TV and Video research group at the Communication research Center – Industry Canada, Ottawa, Canada: Philip Corriveau (currently at Intel Corp.), Ron Reneaud & Andre Vincent for their help and support during my work, and allowing me to use their facility for the practical part of my study.

My special thanks to the whole team of division 7 of Rohde & Schwarz in Germany, USA and Canada for their continuous help and support during my work, especially:

1. Thomas Bichlmaier, Germany
2. Harald Ibl, Germany
3. Peter Foulger, Canada
4. Alexander Woerner, USA

There are a number of people who have helped me with papers, support or encouragement including the VQEG members. I especially want to thank:

5. Scott Daly, Sharp, USA
6. Charles Fenimore, NIST, USA
7. Brian Flowers, VQEG
8. Prof. Mohamed Ghanbary, Univ. of Essex, UK
9. Margaret Pinson, NTIA, USA
10. Stefan Winkler, EPFL, Switzerland
11. Stefan Wolf, NTIA, USA

My biggest thanks goes to my parents and family for their continuous help, support and encouragement throughout my whole life, and my daughter Nadine, who have sacrificed many weekends without going to the movies or having fun!

Chapter 1

Introduction

"The basic purpose of television systems is to extend the sense of vision and hearing beyond their natural limits" [Bru99]

The first television systems were mechanical, and then they became electronic. The first color broadcast was an NTSC in the USA in 1953. SECAM followed in France in 1957, then PAL in Germany in 1961. The next revolution was in HDTV, which had some momentum going for it, but didn't pick up worldwide until the introduction of digital television.

The first success of digital TV was in post-production applications, where the high cost of digitizing the video stream was compensated by the capability of limitless layering of effects while preserving the same quality. This is one of the huge drawbacks of analog TV. The use of digital TV could only be extended if the bandwidth of data could be reduced. As the digitized video stream of a standard definition picture is over 200 MBits/sec.(ITU-601 standard), and the new digital HDTV standard (ITU-292) is about 1.5 Gbits/sec., it's extremely difficult and costly to manage this bandwidth in storage and transmission. The cost of transmission is directly proportional to the bit rate transmitted.

Thus compression emerged as a practical solution to this problem. "Compression is the science of reducing the amount of data used to convey information" [Sym98]

A typical bit rate for standard definition signal, which is currently in our digital Satellites and digital cable systems, is around 2-8 Mbits/sec., and around 20 Mbits/sec. for HDTV. This is almost a 100:1 compression ratio, and this puts a big question mark on the quality of signal compared to the original one.

Compression is not new to Television. Interlace is a simple form of compression giving a 2:1 reduction in bandwidth. The use of color difference signals C_B, C_R instead of RGB is another

form of compression. "The human eye is more sensitive to luminance than Chrominance"[Tek98].

When color TV was introduced, the composite color systems like PAL, SECAM and NTSC were a form of compression, because the color was superimposed on the monochrome signal, and used the same bandwidth.

"With the variety of video compression methods in use and being developed, especially in the encoding and decoding, there is a strong requirement for picture quality evaluation, which is independent of the compression algorithm and it's related artifacts" [Tek98].

1.1 The problem of TV Networks

Currently there are a few systems for objective video quality evaluations based on some of the algorithms mentioned in this thesis. Subjective video quality evaluation has been the dominant method of quality measurement at TV stations, TV Networks and distribution companies. The later are sometimes cable companies, satellite companies or even phone utility companies. They would have a group of professional quality evaluators referred to sometimes as "Golden Eyes" to see the programming material and sense any problems in quality. However with the increased number of TV channels to monitor, and the need for TV networks to increase that number on the available transport streams and links they have, it has been a problem to maintain quality to a certain level. In fact there has been no standards to follow in order to decide quality levels acceptable, as is the case of analog TV.

At this time, we see many satellite companies worldwide advertising hundreds of channels in their networks, and it's not difficult to calculate how many programs are put in each transport stream, and deduce that some programs are left with 1 Mbits/sec for example, which is not enough for some programs like sports.

The other issue for broadcasters or other companies involved in the distribution and transmission of digital TV signal is the newly introduced objective measurement systems. There has been a few systems introduced, mainly double-ended systems, which would test the

degradation in transmission. They are used also by manufacturers of encoders and MPEG2 codecs for design purposes. In broadcast, they are used in an out of service mode, before the actual transmission or after, and also in cases of evaluation of MPEG2 CODECs.

The first single stimulus technique system was introduced by IFN/R&S, Germany, in 1998, and a few other systems have been developed since then.

The main issue for broadcasters, is what do the readings they see really mean, and how do they correlate to the subjective viewer evaluation? How do the values correlate as well to other objective systems they may already have. And the big question after that how do they use all that to create a standard of quality for MPEG2 compressed video that is homogeneous all over the network.

1.2 Thesis Organization and Contribution

Chapter 2 gives a general overview of the MPEG2 compression standard, giving a glimpse on the DCT transform, the construction of the stream with I,P and B frames, and goes up to the blocks, macroblocks and slices, and then the whole structure of MPEG transport stream with the different tables, flags and bit structure of the MPEG2 packets. It demonstrates along the way, the factors used in compressions that may affect quality.

Chapter 3 gives a brief history of TV, and the standards adopted in the World today for digital TV broadcast. Added to that, there is a description of some of the main digital tape standards used in TV stations and Networks worldwide, describing the level of compression used in them. This is another cause of quality degradation in the broadcast cycle.

Chapter 4 describes the basics of video quality. This is the core of the whole subject. Describing the human eye concept of vision, luminance and chrominance sensitivity in the human vision system (HVS) and perceived concept of quality. Then we describe the different video artifacts known in video streams. Then describe the ITU-500 recommendation for classifying the methods for subjective quality evaluation, followed by corresponding objective

quality evaluation techniques adopted. Then describing types of video quality metrics for objective systems' performance.

Chapter 5 gives an over view of the World standard algorithms developed for objective quality evaluation including 9 double stimulus systems, and one single stimulus system. These were presented at the VQEG evaluation round done in 1998. There has been some more standards developed afterwards, and may be presented at future work for VQEG or other international bodies.

Chapter 6 discusses the VQEG work for the subjective/objective quality evaluation, and the practical test done for subjective /objective evaluation.

Chapter 7 presents the conclusion of this thesis, and the need and applicability of this new suggested standard in the Broadcast Industry. It also highlights current and future work areas and potential areas.

The contribution of this Thesis is:

Generalization of the ITU recommendations and VQEG work to develop a simple application for In-field calibration procedure for TV networks to validate the video quality subjective/objective correlation for an objective video quality monitoring system based on their individualized set of Network characterization. The correlation is realized using the Pearson linear regression technique.

Chapter 2

Background on Video Compression and MPEG2:

"The quality of presentation that can be derived by decoding the compressed multimedia signal is the most important consideration in the choice of the compression Algorithm. The goal is to provide near transparent coding for the class of multimedia signals that are typically used in a particular service" [Has97]

There are a number of compression techniques and methods used, depending on the application. They include JPEG, MPEG-1, MPEG-2, MPEG-4, MPEG-7, H261, H263, fractals and wavelets among others. It's beyond the scope of this study to get into the details of each one of those.

Since the broadcast industry has adopted MPEG2 as the standard of choice in both systems chosen worldwide (ATSC and DVB), we will focus on MPEG2, discussing the techniques involved, and the potential areas of possible signal degradation affecting quality.

2.1- Entropy Coding:

In any program material, there are 2 types of components; The real information of the signal, which is new and not repeated, and the redundant information, which can be predicted from previous information. The true information is called "Entropy". And ideal encoder would take only the entropy of the signal and encode it, leaving all the redundant information and thus saving bandwidth.

Redundancy can be spatial (Related to space or the values of one picture or frame), like a large area of pixels having almost the same value – a blue sky for example. Redundancy can also be temporal (Related to time), where the successive frames are almost identical. For example a car moving along the same background, or scenery. This stream can be represented by the

information from previous frames, plus the difference signal between frames and the motion vector of the car.

Therefore a talking announcer has much redundancy, whereas the water surface in a lake or tree leaves blowing in the wind has a large amount of information. In fact streams with water surfaces are commonly used to stress test encoders in real life applications, because of the complexity involved in compressing such streams.

Fig. 2.1 shows the entropy versus redundant information in a sample program material. There are many techniques used in Entropy coding. One of the most common ones is "Hoffman Coding". This technique uses the probability of existence of a certain symbol. For example, symbols that occur often, will use small codes, and the code becomes larger for symbols that occur less frequently.

The amount of information transferred in a symbol A that occurs with probability p is:

$$I_A = Log_2\left[\frac{1}{P_A}\right] \qquad [2.1]$$

Where I_A is the number of bits required to express the amount of information conveyed in symbol A, and P_A is the probability of symbol A occurring.

The entropy of code sequence is the average amount of information contained in each symbol of the sequence:

$$H = \sum P(s) * Log_2\left[\frac{1}{P_S}\right] \qquad [2.2]$$

Where H is the entropy of the coding representation, and s ranges through all symbols in the alphabet of symbols. Ref. to [Wes97] as an example, the entropy of the symbols in a sample of 6 symbols shown below is:

H = (1/2) X 1 + (1/4) X 2 + (1/16) X 4 + (1/16) X 4 + (1/16) X 4 + (1/16) X 4 = 2

If we use a fixed word length of 3, we could represent the 6 symbols in the table, but using Hoffman, we get an average code length of 2.

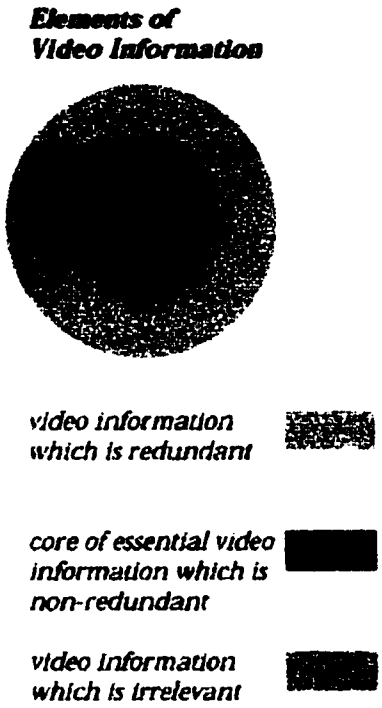


Fig.2.1 Elements of Video Information [Pan00]

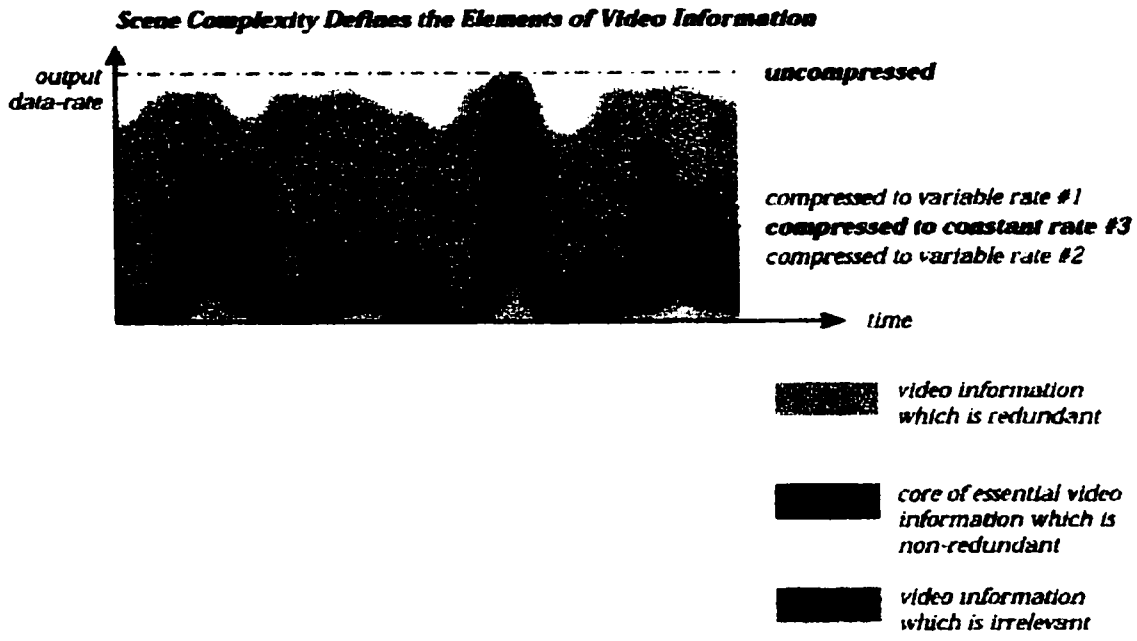


Fig 2.2 Scene Complexity [Pan00]

Intra coding (intra=within) is a technique that exploits spatial redundancy, or redundancies within one picture or frame. Inter coding (inter=between) is a technique exploits temporal redundancy.

Intra coding can be used alone as in JPEG, or along with inter-coding as in MPEG. Intra-coding depends on 2 things in typical images. First, not all spatial frequencies are present, and also, the higher the spatial frequency, the lower the amplitude.

Cutting out high frequency components from images significantly cuts the components, and this is one of the ways of compression used. The coefficients created in transforms like wavelets, FFT or DCT are much less, where most of them will be zero.

Inter-coding deals with similarities between successive images (Temporal). The picture of an object moving on the same background needs sending motion vectors to compensate for the motion between each frame and the next. Then using predicted information from other frames, this frame would be recreated at the decoder side.

The vector transmission requires less data than sending the picture difference data.

2.2- Spatial Coding:

In spatial coding, we transform the waveform into another domain, changing the frequency components in a picture into coefficients.

Fourier transform is one of the most common transforms, where we multiply the input signal by a "Basis Function" sharing different multiples of the frequency and integrating the product. When the input signal contains that frequency, there will be a coefficient from the integral giving the amplitude of that frequency. When the frequency doesn't exist, the integral will be zero at that frequency.

The Fourier transform requires coefficients for both sine and cosine components of each frequency. However, in the cosine transforms, the input waveform is time-mirrored with itself before multiplying it with the basis function. This mirroring cancels the sine coefficients and doubles all of the cosine components.

2.3- Discrete Cosine Transform (DCT)

The discrete cosine transform is the sampled or discrete version of the cosine transform and is used in 2-dimension form in MPEG.

In video display systems, we deal with pixels in a frame, and the next level of entities is called blocks. The block in MPEG case is an 8X8 pixel array. There is another term, called Macro block, which is a 16X16 pixels. A block of 8X8 pixels is transformed into a block of 8X8 coefficients via the DCT transform. The DCT equation is:

$$F(u,v) = 0.25\, C(u)\, C(v) \sum_{x=0}^{x=7} \sum_{y=0}^{y=7} f(x,y) \frac{\cos(2x+1)u\pi}{16} \frac{\cos(2y+1)v\pi}{16} \quad [2.3]$$

Where $C(u)=C(v)=\frac{1}{\sqrt{2}}$ for (u=v=0) ; C(u)=C(v) = 1 (Otherwise)

In some cases, the coefficients give longer values than the pixel length. However, the next steps of weighting or quantization and coding use those values to create a much-compressed value. The inverse DCT on the decoder side will create back the original signal. There is of course some loss induced in this process, because quantizing these coefficients into discrete values, the decoder on the other side cannot determine the original values, but rather deals with the quantized discrete values created in this process.

Fig.2.3 shows the DCT for an 8X8 block. For an 8X8 DCT block, the top left coefficient is the average brightness or DC component of the whole block. Then there are 63 AC coefficients for the

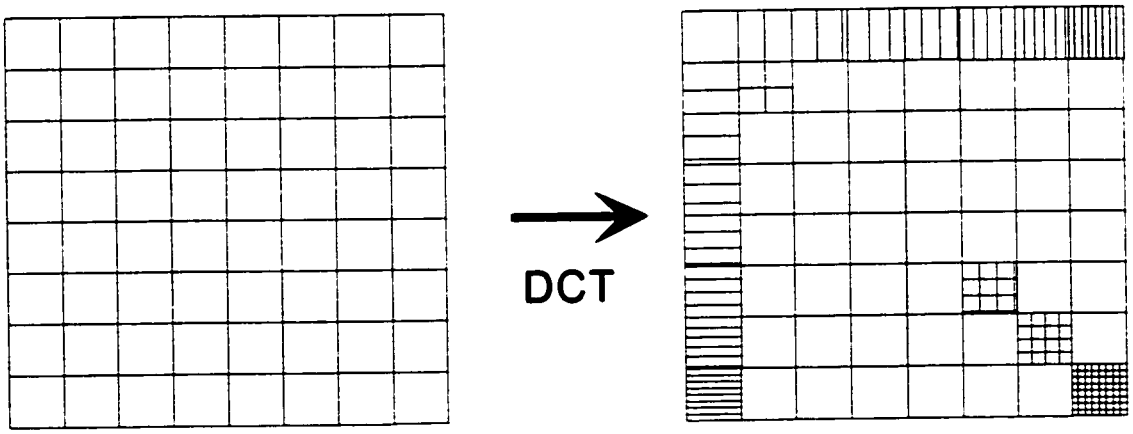


Fig 2.3 The MPEG2 Basis Function and a DCT

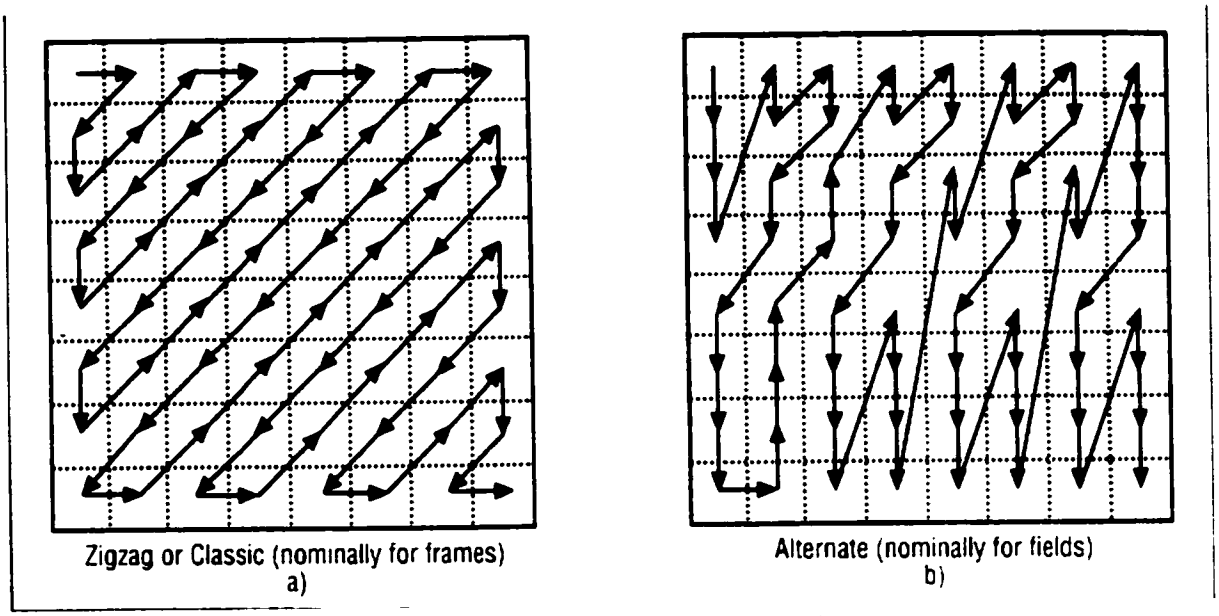


Fig. 2.4 Scanning Techniques [Tek98]

8X8 values. Going from left to right, the spatial function increases, and top to bottom the vertical frequency increases. For color difference signals C_b and C_r, there are separate 8X8 arrays other than the luminance Y array and transformed separately.

2.4- Quantization (Weighting):

"Human perception of noise in pictures is not uniform but is a function of the spatial frequency. More noise can be tolerated at high spatial frequencies" [Tek98]. Thus giving accuracy to low frequency coefficients and less accuracy to high frequency coefficients will not affect the human perception of the content. This is done in the weighting process. In the same token, the human eye is more sensitive to luminance or brightness than Chrominance (Color), and giving less accuracy to the color coefficients will also not affect the human perception as well.

In this case, low frequencies are divided by small numbers to create quantizing levels, whereas high frequencies are divided into high numbers making fewer levels. In the

inverse process of re-quantization on the decoding side, low frequency coefficients will suffer low noise added to the signal, whereas high frequency coefficients will suffer from more noise. High frequency components in a picture are like sharp building edges for example. And those could be seen blurred sometimes in MPEG compressed streams.

2.5 Scanning:

In regular programs, the top left pixels are the ones that have the significant values, and the rest of the block consists mostly of zeros. Sending those values first, then putting a string for the rest of pixels with one value which is zero, will give even more compression of data sent. This technique is called Run-length coding. The scanning methods used in MPEG accomplish this task, by scanning the pixels towards the top left corner of the block first, and moving in sort of a bottom right direction. Fig 2.4 shows 2 of the most common scanning techniques used (Zigzag and Alternate scanning).

2.6 I, P and B frames:

There are 3 types of pictures or frames in an MPEG stream. "I" is the "Intra-coded" picture that is completely encoded and when decoded doesn't need any further information.

The "P" picture is the "Forward Predicted" picture, is predicted from earlier pictures, that could be "I" or "P" frame. "P" picture data comprise of vectors that describing where in the previous picture, the macro blocks should be taken from, and transform coefficients that describe the correction or difference data that must be added to the macro blocks. "B" pictures are "Bi-directionally Predicted" from earlier or later "I" or "P" pictures. B-picture data consist of vectors describing where in earlier or later pictures data should

be taken from, and transform coefficients that provide the correction. Because of the feature of bi-directional in "B" frames, it only requires one quarter the data of an "I" frame, and half the data of a "P" frame.

There is a term called GOP or "Group of pictures" in an MPEG stream. It's an "I" frame followed by "B" and "P" frames, and the GOP ends before the next "I" frame.

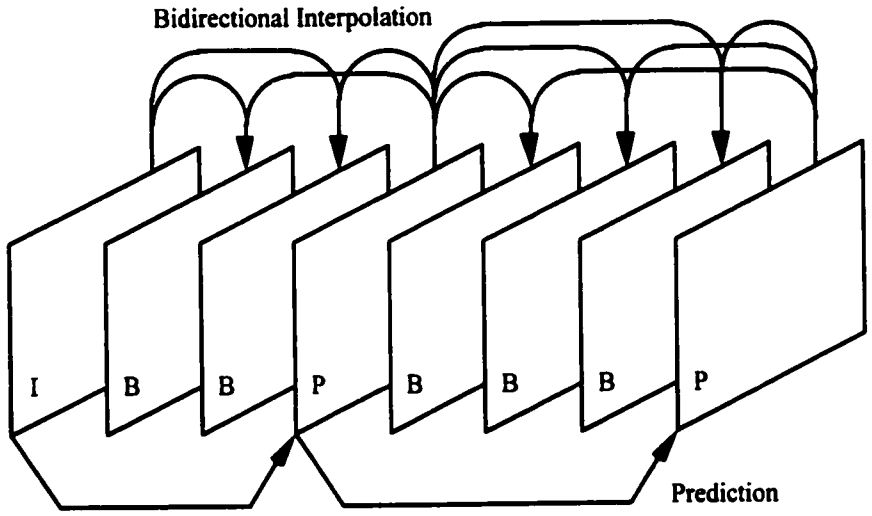


Fig 2.5 I, P and B frames in a Group of Pictures Structure (GOP)

2.7 Profiles and Levels:

MPEG can be used for a number of applications and with different performance. That's why MPEG2 is divided into different profiles, each profile has different levels. For example the profile 4:2:2 indicates 4 samples of luminance to 2 samples of both color difference components C_b and C_r. This profile is used mostly for off Air Broadcast. 4:2:0 uses less one color difference signal sampled, and has more compression though. This application is popular in cable and satellite applications.

Fig. 2.6 shows the different levels and profiles. The simple profile doesn't have B frames, thus needs simpler hardware and has less delay in decoding. The main profile is designed for large portion of applications, especially at main level.

		Profile							
					Spatial		High		
Level	Description	Simple	Main	SNR	Enhancement Layer	Base Layer	Enhancement Layer	Base Layer	4:2:2
	Frame Type	I & P	I, P & B	I, P & B	I, P & B	I, P & B	I, P & B	I, P & B	I, P & B
	Chroma Sampling	4:2:0	4:2:0	4:2:0	4:2:0	4:02:00	4:2:0, 4:2:2	4:2:0, 4:2:2	4:2:0, 4:2:2
High	Samples/line		1920H				1920H	960H	1920H
	Lines/frame		1152V				1152V	576V	1152V
	Frames/sec		60 fr/s				60 fr/s	30 fr/s	60 fr/s
	Max Bit rate		80 Mbps				100 Mbps	100 Mbps	
High 1440	Samples/line		1440H		1440H	720H	1440H	720H	
	Lines/frame		1152V		1152V	576V	1152V	576V	
	Frames/sec		60 fr/s		60 fr/s	30 fr/s	60 fr/s	30 fr/s	
	Max Bit rate		60 Mbps		60 Mbps	60 Mbps	80 Mbps	80 Mbps	
Main	Samples/line	720H	720H	720H			720H	352H	720H
	Lines/frame	576V	576V	576V			576V	288V	512/608V
	Frames/sec	30 fr/s	30 fr/s	30 fr/s			30 fr/s	30 fr/s	30 fr/s
	Max Bit rate	15 Mbps	15 Mbps	15 Mbps			20 Mbps	20 Mbps	50 Mbps
Low	Samples/line		352H	352H					
	Lines/frame		288V	288V					
	Frames/sec		30 fr/s	30 fr/s					
	Max Bit rate		4 Mbps	4 Mbps					

Fig 2.6. MPEG-2 Levels and Profiles [Sym98, Tek98, Rob98]

2.8 Macroblocks

A macroblock is a group of DCT blocks, which correspond to the information content of a window of 16X16 pixels in the original picture [Rob98]. It contains the basic number of luminance and chrominance blocks in the sampling structure used. So, in a 4:2:2 profile

signal, a macroblock contains 4 (8X8) blocks of luminance, 2 blocks of Cb color difference signal, and 2 blocks of Cr color difference signal.

[bha97] states: "When the coded bit stream has no scalable extensions, then the blocks within the macroblock are 8x8 pixels. With scalable extensions, macroblocks may contain scaled blocks with lower resolution like 1x1, 2x2 or 4x4".

The macroblocks are used also for temporal redundancy prediction. When comparing one 16X16 pixel region from one frame to the next in the following time slot, the temporal redundancy processor calculates the difference between the 2 frames. If the 2 frames have a high degree of temporal redundancy, then the difference frame would have a large number of pixels with values near zero.

The 16x16 pixel size is chosen because of the compromise between providing efficient temporal redundancy and requiring moderate computational requirements. The macroblock header contains information about the type (Y or Cb or Cr) and the corresponding motion compensation vectors.

2.9 Slices

The slice is a string of consecutive Macroblocks of arbitrary length running from left to the right of the picture. The slice could be as small as one block or as large as the whole picture. The first and last macroblocks in a slice must be transmitted, but ones in between could be skipped. A slice cannot exceed the right edge of the picture, and slices must not overlap.

The slice header contains information about its' position within the picture, and its' quantizer scaling factor, in case an error occurs and re-transmission is necessary.

2.10 MPEG Transport Stream:

The MPEG2 transport stream normally contains several elementary streams. The elementary stream is the output of the MPEG2 encoder for a single audio or video stream. The content of the transport stream is described in the program specific information table (PSI). Each transport stream contains a Program Associated Table (PAT), as well as one or more Program Map Table (PMT). The PAT is contained in the transport stream packet with the PID 0X000. It refers all the programs contained in the transport stream indicating the program number and the corresponding PID for the Program Map Table (PMT).

The elementary streams whether video, audio or data that belong to the individual programs are described in a PMT.

The elementary stream can be packetized in a packetized elementary stream (PES). The packets have a fixed length of 188 bytes. Each packet is synched with a byte (47Hex). The packets can also have the length of 204 bytes in DVB and 208 in ATSC. The different 16 or 20 bytes are used for Forward error correction before transforming the signal from base-band MPEG into a modulated RF signal. The forward error correction used in broadcast applications is (Reed-Solomon forward error correction).

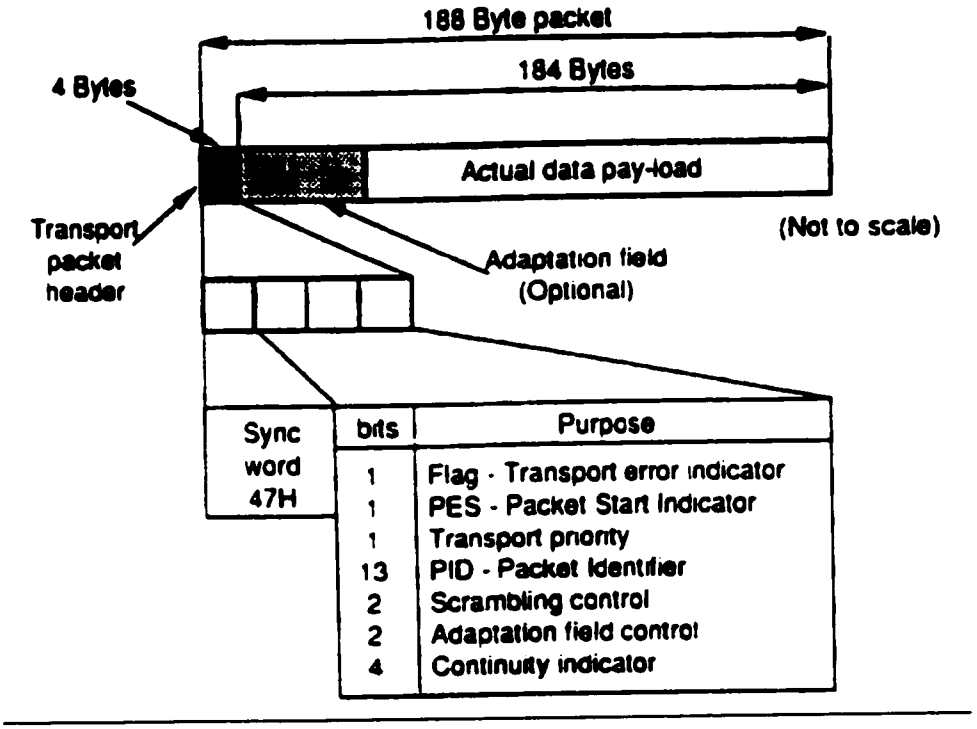


Fig 2.7 The MPEG-2 188 Byte Packet [Rob98]

For program synchronization, there is a common clock reference called "Program Clock reference (PCR)". This signal is carried every 40 ms of time at the most to synchronize the whole program. The individual elementary streams contain time stamps such as Decoding Time Stamps (DTS), and Presentation Time Stamps (PTS).

For service, there are tables like Program Specific Information (PSI), Service Information (SI). PSI contains partly PAT, PMT, CAT (Conditional Access Table) and NIT (Network Information Table). The CAT gives information for encrypted programs, whereas NIT contains data given by the network operator for tuning the receivers (e.g. orbit positions or transponder numbers).

BAT (Bouquet Association Table) contains information about programs from a supplier irrespective of the propagation paths of the programs. SDT (Service Description Table) describes the programs offered.

EIT (Event Information Table) supplies the database for an electronic TV guide with info about the type of program and age classification of the viewer.

RST (Running Status Table) comprises status information about the individual programs and especially serves for controlling video recorders.

TDT (Time and Data Tables) provides information about the date and current time. TOT (Time Offset Table) provides information about the local time offset in addition to date and time.

SI (Stuffing Table) has no relevant content. It's generated when invalid tables are overwritten during transmission.

The MPEG transport stream is an extremely complex structure, with linked and interlinked tables and identifiers to separate elementary streams, vectors, coefficients and quantization tables.

Failures could be in within the encoder, decoder or due to the transmission medium. Modulation errors, RF effects in noise, multipath fading among others could contribute to the reception of a defective stream without the right information to decode it. Fig 2.8 shows a typical transport stream and packets.

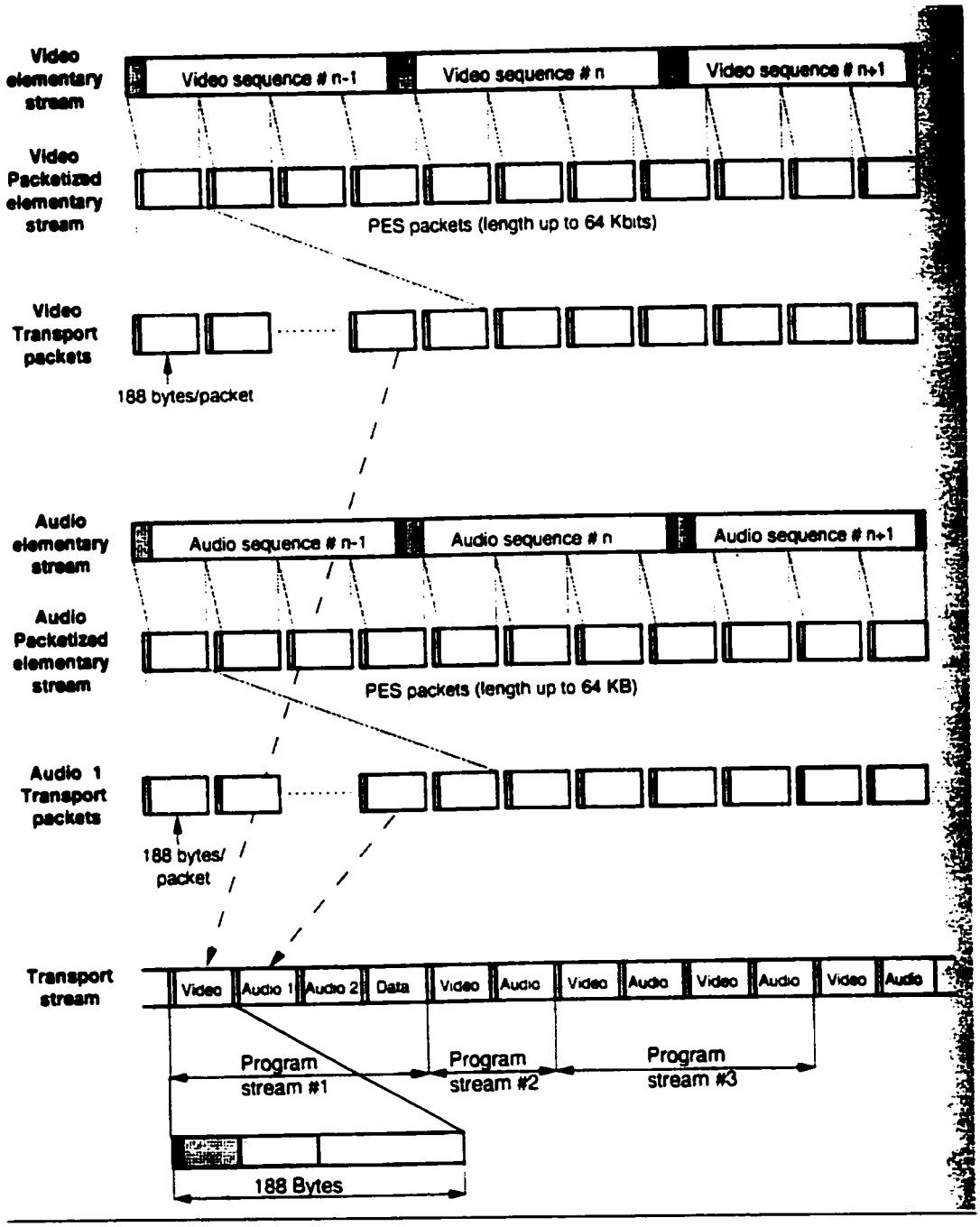


Fig 2.8 The MPEG-2 Transport Stream [Rob98]

Chapter 3

Digital TV and Video Standards

3.1 History of TV

The development of TV started with the German Paul Gottlieb Nipkow invented his mechanical television system in 1884 [Bru99]. The Americans and Europeans worked on that until electronic TV was created. There were different line systems introduced like 405 line system in Great Britain in 1935 developed by EMI under Sir Isaac Shoenberg, and adopted by the BBC in 1836. A 441 interlaced lines, 50Hz introduced in Germany in 1937. Then later in France emerged an 819 line interlaced system. The USA followed in 1941 with the 525 lines 60 Hz system, and then Japan followed in 1953 with a similar system.

As for color television, there was a German patent for such a system dating back to 1904 [Bru99]. Baird demonstrated the first color mechanical TV system in 1928 with a Nipkow disk with 3 spirals, one for each primary color (Red, Green & Blue). The first color TV broadcast started in the USA in 1951 using "Abortive frame sequential system". Then in 1953 the National Television systems Committee (NTSC) developed their system, and it started broadcasting in 1954 in the USA followed by Japan in 1960.

The NTSC system was sensitive to some distortions during transmission causing color hues. Therefore its' acronym sometimes means "Never The Same Color". In 1957, Henri de France developed the "Systeme Electronique Coloeur avec Memoire" (SECAM), which tackled the hue error problem. Later in 1961 W. Bruch in Germany developed the "Phase Alternating Line" system (PAL).

The term High Definition has been an old one in the industry. Baird called his mechanical 30-line system a High Definition one. Nowadays, we refer to HDTV as 1000 line systems. There is other work on higher resolution systems like ultra HDTV in Japan with 3000 lines.

The real work on HDTV started in the sixties in Japan when Dr. Takashi Fuijo from NHK started development on a high quality system that is compatible with 35 mm quality and CD quality audio. In the second half of 1970s the first broadcast started with an 1125 lines 60 Hz system. NHK also developed a Multiple Sub-Nyquist Sampling Encoding (MUSE) for satellite service in 1984 with 1125 lines 60 Hz system.

In 1985, the European "Comite' Consultatif International des Radiocommunications CCIR Interim working party (IWP) made a proposal for a 1125 line 60 Hz system similar to the Japanese system. This was not successful. This is when the USA and Europe went into different paths in the HDTV development.

In Europe they developed a system called High Definition Multiplexed Analog component (HDMAC) for satellite transmission using a 1250 lines and 50 Hz system.

In 1987 in the USA, the FCC asked the industry for proposals for a HDTV terrestrial system and received 21 proposals. However, only 4 of those were compatible with the NTSC system. In 1992 the 4 parties involved:

1- General Instruments (later bought by Motorola),

2- AT&T / Zenith

3- David Sarnoff Research Center/ Philips/ Thomson

4- MIT (Massachusetts Institute of Technology)

Formed what is known as the Grand Alliance, with the mandate of developing one HDTV standard for terrestrial TV in the USA. They chose MPEG2 for source coding, Dolby AC-3 for audio coding. The GA was later transformed into the Advanced Television Standards Committee ATSC. The HDTV system they developed in called ATSC for terrestrial broadcast.

In 1993, the European Launching Group with 84 European broadcasters, telecommunications organizations, manufacturers and regulatory authorities signed a memorandum forming the Digital Video Broadcast project (DVB). The DVB system adopted MPEG-2 as well for source

coding for both audio and video. There are DVB standards now for terrestrial, cable as well as satellite broadcast.

Thus there are 2 main standards in HDTV terrestrial systems worldwide:

1- ATSC

2- DVB-T

3.2 Advanced Television Standards Committee (ATSC):

The Advanced Television Systems Committee (ATSC) was formed in the USA in 1982 by the member organizations of:

- The Joint Committee on Inter Society Coordination (JCIC)
- The Electronics Industries Association (EIA)
- The Institute of Electrical and Electronics Engineers (IEEE)
- The National Association of Broadcasters (NAB)
- The National Cable Television Association (NCTA)
- The Society of Motion Picture and Television Engineers (SMPTE)

The ATSC currently has over 200 members. The ATSC standards include standard definition TV, high definition TV, data broadcasting, multi-channel surround sound audio and satellite direct to home broadcasting.

The Federal Communications Commission (FCC) in the USA adopted the ATSC digital TV standard (document A/53) on Dec. 24th/1996. Canada adopted that on Nov. 8th, 1997. South Korea on Nov. 21st, 1997, Taiwan on May 8th, 1998 and Argentina on Oct. 22nd, 1998.

The ATSC standard document A53 was issued in Sept./95 and later revised in April/2001 [www.atsc.org]. It defines the structure of the ATSC transport stream, which is 19.39 Mbits/sec, and based on MPEG-2 source coding standard as specified in ISO/IEC 13818-1 standard.

For terrestrial transmission, the modulation technique used is 8VSB (Vestigial side band with 8 discrete values). The standard also allows for 16VSB, although nothing much has happened in

that direction. For picture formats, the ATSC standard allows for 36 picture formats (including 18 NTSC-friendly formats) . Probably the most 2 popular formats adopted by broadcasters in the USA are the 720 line progressive and 1080 line interlaced formats.

ATSC Picture Formats				
Aspect Ratio	Active H-Pixels	Active Lines	Scanning Mode	Frame Rate
4:3	640	480	Progressive	60(59.94), 30(29.97), 24(23.98)
			Interlaced	30(29.97)
4:3	704	480	Progressive	60(59.94), 30(29.97), 24(23.98)
			Interlaced	30(29.97)
16:9	704	480	Progressive	60(59.94), 30(29.97), 24(23.98)
			Interlaced	30(29.97)
16:9	1280	720	Progressive	60(59.94), 30(29.97), 24(23.98)
16:9	1920	1080	Progressive	30(29.97), 24(23.98)
			Interlaced	30(29.97)

Table 3.1 Popular ATSC picture formats

The FCC in the USA has issued a ruling in 1996 that all broadcasting stations in the USA have to broadcast at least one ATSC station by the year 2006, or they may lose the spectrum they already have. At the moment there are more than 200 stations in the USA broadcasting ATSC terrestrial signal. [www.atsc.org], and are increasing every day. The total number of off the air stations is about 1500 stations in the USA.

3.3 Digital Video Broadcast (DVB):

The Digital Video Broadcasting Project (DVB) is a consortium of over 300 broadcasters, manufacturers, network operators and regulatory bodies worldwide, committed to designing a global standard for the delivery of digital television. Numerous broadcast services using DVB standards are now operational in Europe, North and South America, Africa, Asia, and Australia.

The DVB family of standards include:

DVB-S A satellite system that can be used with any transponder, current or planned (Uses QPSK modulation).

DVB-C	A matching cable system to suit the characteristics of all cable networks (Uses QAM-64 modulation).
DVB-T	A digital terrestrial system (Using COFDM modulation).
DVB-MC/S	A microwave multipoint video distribution systems.
DVB-SI	A service information system, enabling the user to navigate through the DVB environment;
DVB-CA	A common scrambling or Conditional Access system.
DVB-CI	A common interface for conditional access and other uses.

DVB-T which uses the multi-carrier Coded Orthogonal Frequency Division Multiplexing (COFDM) modulation technique, is capable of delivering crystal clear picture to television connected to portable, set-top antennas in hostile reception environments such as city apartments, or even to receivers on the move. In trials in Germany since 1997, DVB-T has been tested in slow moving city trams and on the autobahn at speeds in excess of 275 km/hr.

The vision of DVB is to bring many types of digital video broadcast services to various devices in the home of the future.

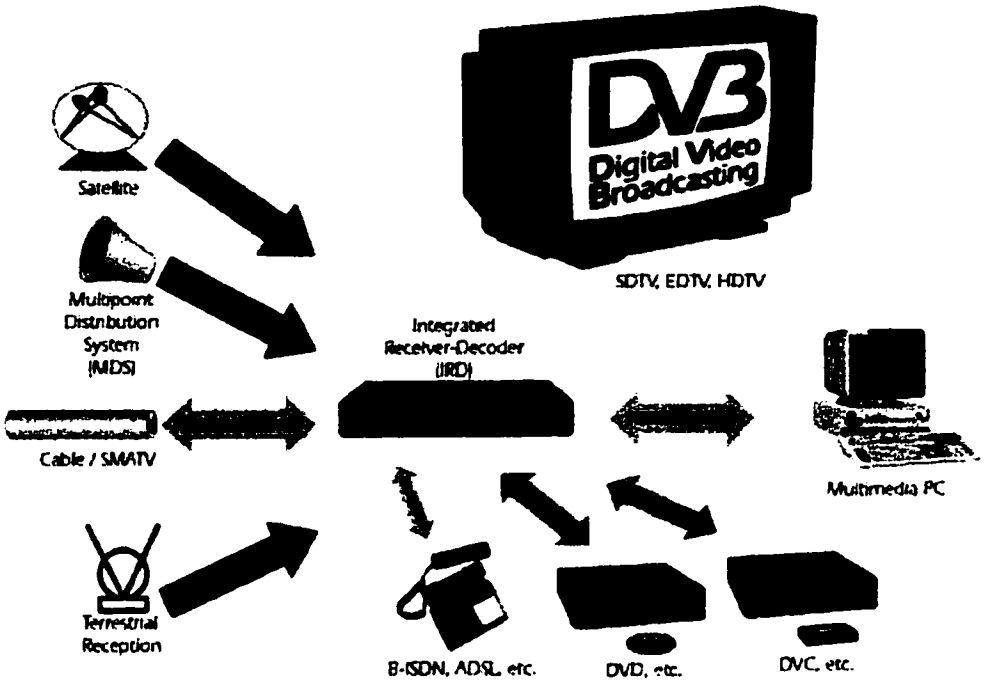


Fig 3.1 Multimedia Interfaces in the Home using DVB

DVB has defined the delivery media (for satellite, cable, terrestrial, etc.), and the interfaces (to ISDN, PSTN, ATM, PDH, SDH... etc.), which enable a cluster of interconnected devices in the home, all receiving and processing digital broadcast services. It all comes together on the set top box (STB) also known as the integrated receiver-decoder (IRD). MPEG-2 is used as a common sound and vision-coding system. In fig. 3.2, four MPEG-2 "data containers" are shown:

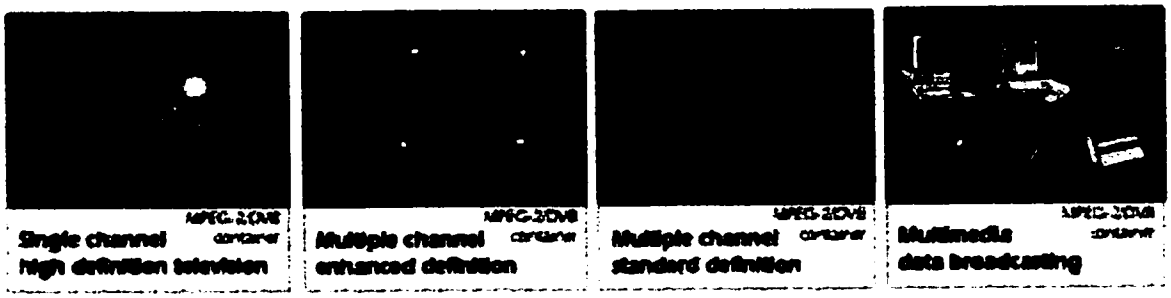


Fig 3.2 The DVB transport Stream

The DVB approach provides great flexibility in terms of transmitted digital information, owing to its data "container" concept.

DVB simply delivers to the receiver "containers" with compressed image, sound or data. No restrictions exist as to the kind of information, which can be stored in these containers. The DVB Service Information acts like a header to the MPEG-container, ensuring that the receiver knows what it needs to decode.

DVB selected the "MPEG-2" approach developed by the Moving Picture Experts Group (MPEG) for the compression of image and sound data prior to transmission. MPEG-2/DVB compliant systems are capable of delivering anything from multiple-channel Standard Definition (SDTV) or Enhanced Definition (EDTV) to single channel HDTV in the same transport stream. Or can also deliver a SDTV signal along with other audio or data services including interactive services. Delivering this service via many mediums, this will open the door for the Convergence era everybody is talking about.

The sound coding system specified for all DVB systems is the MPEG Layer II audio standard (MUSICAM). MPEG Layer II is a digital compression system, which takes advantage of the fact that a sound element will have a masking effect on other nearby lower-level sounds (or on noise). This is used to facilitate the coding of the audio with low data rates. This system can achieve a sound quality, which is very close to Compact Disc. The system can be used for mono, stereo, or multilingual sound, and as well for surround sound.

A DVB "Blue Book" called "Implementation Guidelines for the use of MPEG-2 Systems, Video and Audio in Satellite and Cable Broadcasting Applications in Europe" has been provided which describes in detail the subset of MPEG-2 elements to be used by DVB.

The baseline SDTV decoder is based on MPEG-2 Main Profile at Main Level (MP @ ML). The HDTV baseline decoder uses MPEG-2 Main Profile at High Level (MP @ HL), ensuring backwards compatibility with existing DVB/MPEG-2 bit streams.

While the DVB MPEG-2 Data Container gives a flexible range of service options, the Implementation Guidelines recommend HDTV broadcasters use the so-called Common Image Format (CIF) proposed by the ITU and DAVIC (i.e. 1080 lines by 1920 pixels). They also detail

the implementation requirements necessary for Integrated Receiver-Decoders (IRDs) to be used.

When HDTV broadcasts are directed towards populations of DVB-compliant IRDs of which some are not HDTV-enabled, the DVB Implementation Guidelines recommend multiplexing SDTV program streams into the MPEG-2 bit stream alongside the HDTV programs, in an approach known as Simul-casting.

3.3.1 Coded Orthogonal Frequency Division Multiplexing (COFDM)

COFDM is an un-orthodox, rather sophisticated modulation scheme that has been considered on-and-off for terrestrial broadcasting for over 6 years. It was originally considered as the modulation approach for one of the proposed digital ATV systems but was abandoned for another method because of the time it would have taken to design and build a world-class version for testing. Later it became of personal interest because of the opportunity it offered to support distributed transmission.

Development of COFDM has been largely in Europe, with some work coming later in Japan. In Europe, it was very close to adoption by the Digital Video Broadcasting (DVB) group as the standard method for terrestrial digital transmission. A late-in-the-game effort was mounted by some North American Broadcasters to gain consideration of COFDM as a replacement for the Grand Alliance VSB scheme, but it was too little, too late to succeed.

Coded Orthogonal Frequency Division Multiplexing (COFDM) has been specified for digital broadcasting systems for both audio -- Digital Audio Broadcasting (DAB) and terrestrial television -- Digital Video Broadcasting (DVB-T).

COFDM is particularly well matched to these applications, since it is very tolerant of the effects of multi-path (provided a suitable guard interval is used). It is not limited to 'natural' multi-path as it can also be used in Single-Frequency Networks (SFNs) in which all transmitters radiate the same signal on the same frequency.

A receiver may thus receive signals from several transmitters, normally with different delays and thus forming a kind of 'unnatural' additional multi-path. Provided the range of delays of the multi-path (natural or 'unnatural') does not exceed the designed tolerance of the system (slightly greater than the guard interval) all the received-signal components contribute usefully.
Multi-path can be viewed in the frequency domain as a frequency selective channel response.
Another frequency-dependent effect for which COFDM offers real benefit is the presence of isolated narrow-band interfering signals within the signal bandwidth. Note that conventional analogue television signals (NTSC/PAL/ SECAM) essentially behave like narrow-band interferers to COFDM.
COFDM copes with both these frequency-dependent effects as a result of the use of forward error coding. However, rather more is involved than simply adding coding -- the 'C' -- to an un-coded OFDM system. The coding and decoding is integrated in a way, which is specially tailored to frequency-dependent channels and brings much better performance.

3.3.2 UNCODED OFDM

OFDM spreads the data to be transmitted over a large number of carriers -- typically more than a thousand (1705 carriers in case of 2K systems, and 6814 carriers for 8K systems. The data rate to be conveyed by each of these carriers is correspondingly reduced. It follows that the symbol length is in turn extended. These modulation symbols on each of the carriers are arranged to occur simultaneously.
The carriers have common frequency spacing. This is the inverse of the duration, called the active symbol period, over which the receiver will examine the signal, performing the equivalent of an 'integrate-and-dump' demodulation. This choice of carrier spacing ensures orthogonality (the 'O' of OFDM) of the carriers -- the demodulator for one carrier does not 'see' the modulation of the others, so there is no cross talk between carriers, even though there is no explicit filtering and their spectra overlap.

A further refinement adds the concept of a guard interval. Each modulation symbol is transmitted for a total symbol period, which is longer than the active symbol period by a period called the guard interval. This means that the receiver will experience neither inter-symbol nor inter-carrier interference provided that any echoes present in the signal have a delay, which does not exceed the guard interval.

The addition of the guard interval reduces the data capacity by an amount dependent on its length. The concept of a guard interval could in principle be applied to a single-carrier system, but the loss of data capacity would normally be prohibitive. With OFDM it is possible to protect against echoes with prolonged delay, simply by choosing a sufficient number of carriers that the guard interval need not form too great a fraction of the active symbol period. Both DAB and DVB-T have a guard interval which is no greater than 1/4 of the active symbol period, but can protect against echo delays of the order of 200 μs (depending on the mode chosen).

Fortunately the apparently very complex processes of modulating (and demodulating) thousands of carriers simultaneously are equivalent to Discrete Fourier Transform operations, for which efficient Fast Fourier Transform (FFT) algorithms exist. Thus integrated circuit implementations of OFDM demodulators are feasible for affordable mass-produced receivers.

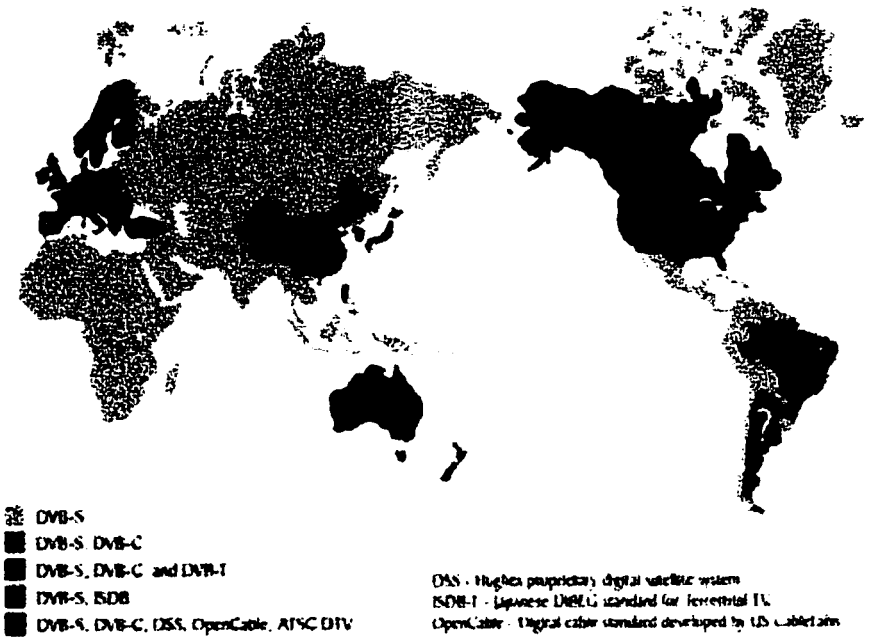


Fig 3.3 DVB implementation Worldwide

As seen from this map, the DVB standard is dominant in the World, except in North America and a few other countries for terrestrial transmission. In Japan they adopted a variation of the DVB called ISDB.

ISDB is the Japanese Digital Broadcasting Experts Group (DiBEG's) variant of the DVB-T system. It differs in one key respect only, the use of an intermediate (software driven) data segmentation system, whereby services such as radio, SDTV, HDTV and Mobile TV can be flexibly allocated pieces of the overall service bandwidth. The Japanese ISDB system is anticipated to launch in Japan no earlier than 2003.

At the moment DVB-T broadcast has already started in the UK, Sweden, Spain and Australia. Many European countries and New Zealand will follow very soon.

3.4 Digital Video Tape Standards:

There are a number of digital tape standards in the market nowadays. And of course there is the concept of DDR or digital disk recorders. These recorders record on a hard drive medium, and

can be without compression. However, many TV stations use one or more of the other tape standards for their materials. The main digital tape standards are:

D1 (Introduced 1986)

A format for component digital video tape recording working to the ITU-R 601, 4:2:2 standard using 8-bit sampling. There is no compression in this format, and thus it's one of the best quality systems, however it's expensive. The tape is 19 mm wide and allows up to 94 minutes to be recorded on a cassette. Being a component recording system it is ideal for studio or post-production work with its high chrominance bandwidth allowing excellent chroma keying. Also multiple generations are possible with very little degradation and D1 equipment can integrate without transcoding to most digital effects systems, telecines, graphics devices, disk recorders, etc. Being component there are no color framing requirements. Despite the advantages, D1 equipment is not extensively used in general areas of TV production, at least partly due to its high cost. (Often used incorrectly to indicate component digital video.)

D2 (Introduced 1988)

The VTR standard for digital composite (coded) NTSC or PAL signals that uses data conforming to SMPTE 244M. It uses 19 mm tape and records up to 208 minutes on a single cassette. Neither cassettes nor recording formats are compatible with D1. D2 has often been used as a direct replacement for 1-inch analog VTRs. Although offering good stunt modes and multiple generations with low losses, being a coded system means coded characteristics are present. The user must be aware of cross-color, transcoding footprints, low chrominance bandwidths and color framing sequences. Employing an 8-bit format to sample the whole coded signal results in reduced amplitude resolution making D2 more susceptible to contouring artifacts. (Often used incorrectly to indicate composite digital video.)

D3 (1990)

A composite digital video recording format that uses data conforming to SMPTE 244M. Uses 1/2-inch tape cassettes for recording digitized composite (coded) PAL or NTSC signals sampled at 8 bits. Cassettes are available for 50 to 245 minutes. Since this uses a composite signal the characteristics are generally as for D2 except that the 1/2-inch cassette size has allowed a full family of VTR equipment to be realized in one format, including a camcorder.

D4

A format designation never utilized due to the fact that the number four is considered unlucky (being synonymous with death in some Asian languages).

D5 (1994)

A VTR format using the same cassette as D3 but recording component signals conforming to the ITU-R BT.601-2 (CCIR 601) recommendations at 10-bit resolution. With internal decoding D5 VTRs can play back D3 tapes and provide component outputs. Being a non-compressed component digital video recorder means D5 enjoys all the performance benefits of D1, making it suitable for high-end postproduction as well as more general studio use. Besides servicing the current 625 and 525 line TV standards the format also has provision for HDTV recording by use of about 4:1 compression (HD D5).

D6 (1996)

A digital tape format which uses a 19mm helical-scan cassette tape to record uncompressed high definition television material at 1.88 GBps (1.2 Gbps). D6 is currently the only high definition recording format defined by a recognized standard. D6 accepts both the European 1250/50 interlaced format and the Japanese 260M version of the 1125/60 interlaced format, which uses 1035 active lines. It does not accept the ITU format of 1080 active lines. ANSI/SMPTE 277M and 278M are D6 standards.

D7 (1996)

DVCPRO. Panasonic's development of native DV component format, which records an 18-micron (18x10-6m, eighteen thousandths of a millimeter) track on 6.35 mm (0.25-inch) metal particle tape. DVCPRO uses native DCT-based DV compression at 5:1 from a 4:1:1 8-bit sampled source. It uses 10 tracks per frame for 525/60 sources and 12 tracks per frame for 625/50 sources, both use 4:1:1 sampling. Tape speed is 33.813mm/s. It includes two 16-bit digital audio channels sampled at 48 kHz and an analog cue track. Both Linear (LTC) and Vertical Interval Time Code (VITC) are supported. There is a 4:2:2 (DVCPRO50) and progressive scan 4:2:0 (DVCPRO P) version of the format, as well as a high definition version (DVCPROHD).

D8

There is no D8. The Television Recording and Reproduction Technology Committee of SMPTE decided to skip D8 because of the possibility of confusion with similarly named digital audio or data recorders (DA-88).

D9 (Formerly Digital-S) (Introduced 1996)

A 1/2-inch digital tape format developed by JVC, which uses a high-density metal particle tape running at 57.8mm/s to record a video data rate of 50 Mbps. The tape can be shuttled and search up to 32x speed. Video sampled at 4:2:2 is compressed at 3.3:1 using DCT-based intra-frame compression (DV). Two or four audio channels are recorded at 16-bit, 48 kHz sampling; each is individually editable. The format also includes two cue tracks. Some machines can play back analog S-VHS. D9 HD is the high definition version recording at 100 Mbps.

D9 HD

A high definition digital component format based on D9. Records on 1/2-inch tape with 100 Mbps video.

DV

This digital VCR format is a cooperation between Hitachi, JVC, Sony, Matsushita, Mitsubishi, Philips, Sanyo, Sharp, Thomson and Toshiba. It uses 6.35 mm (0.25-inch) wide tape in a range of products to record 525/60 or 625/50 video for the consumer (DV) and professional markets (Panasonic's DVCPRO, Sony's DVCAM and Digital-8). All models use digital intra-field DCT-based "DV" compression (about 5:1) to record 8-bit component digital video based on 13.5 MHz luminance sampling. The consumer versions, DVCAM, and Digital-8 sample video at 4:1:1 (525/60) or 4:2:0 (625/50) video (DVCPRO is 4:1:1 in both 525/60 and 625/25) and provide two 16-bit/48 or 44.1 kHz, or four 12-bit/32 kHz audio channels onto a 4 hour 30 minutes standard cassette or smaller 1 hour "mini" cassette. The video recording rate is 25 Mbps.

Digital Betacam (1993)

Introduced by Sony. It has a compression of 2:1.

Betacam SX (1993)

Introduced by Sony. It has a compression of 10:1.

Chapter 4

Basics of Video Quality

4.1 The Human Visual System (HVS)

The eye perceives images with a very unique perception. Fig. 4.1 shows the eye structure.

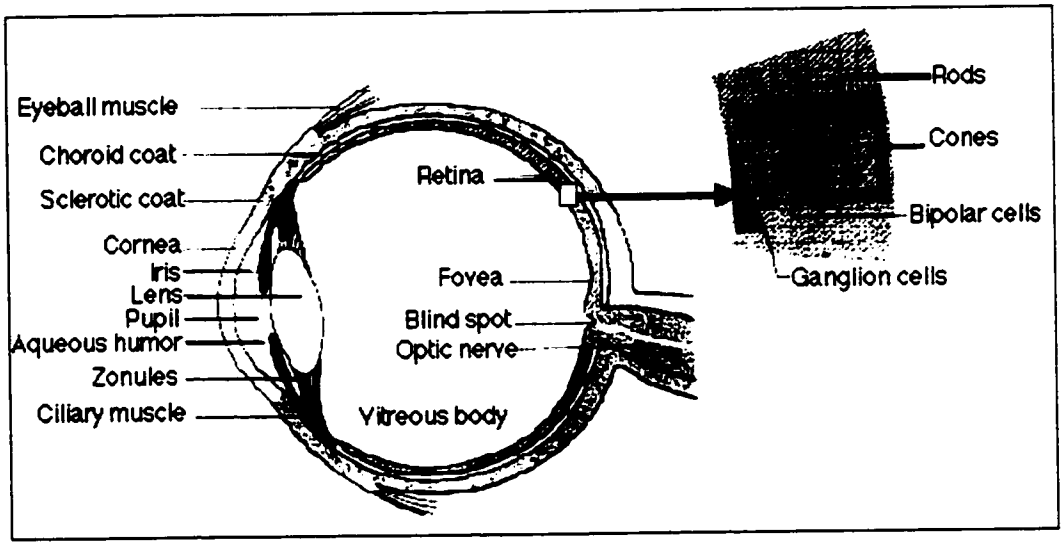


Fig 4.1 Anatomy of the Human Eye

The light bounces off different objects, and then is refracted by the cornea toward the pupil. This is the opening of the iris where the light enters the eye. The light is then refracted by the lens onto the back of the eyeball, forming an image on the retina. The retina consists of receptors sensitive to light called photoreceptors, and those are connected by nerve cells. The photoreceptors contain chemical pigments, which absorb light and create a neural response. "The light absorbed by photoreceptors initiates a chemical reaction that bleaches the pigment, which reduces the sensitivity to light in proportion to the amount of bleached pigment" [Has97]. The amount of bleached pigment corresponds to the amount of light.

There are 2 types of photoreceptors:

1- Rods, which are responsible for low light vision, like night (scotopic) colorless vision.

2- Cones are responsible for chrominance (color) and detail under normal light conditions. There are 3 types of cones that are individually sensitive to Red, green and blue wavelength lights, and the combination of those create all the other different colors.

The light of various wavelengths causes the sensation of colors, or also called hue. In the retina, there are roughly between 110-130 million rods, and about 6 and 7 million cones [Rob98]. That's one of the reasons the human eye is more sensitive to light or luminance than chrominance or color. Fig 4.2 shows the rods and cones cell structure. [http://www.ultranet.com/~jkimball/BiologyPages/V/Vision.html]

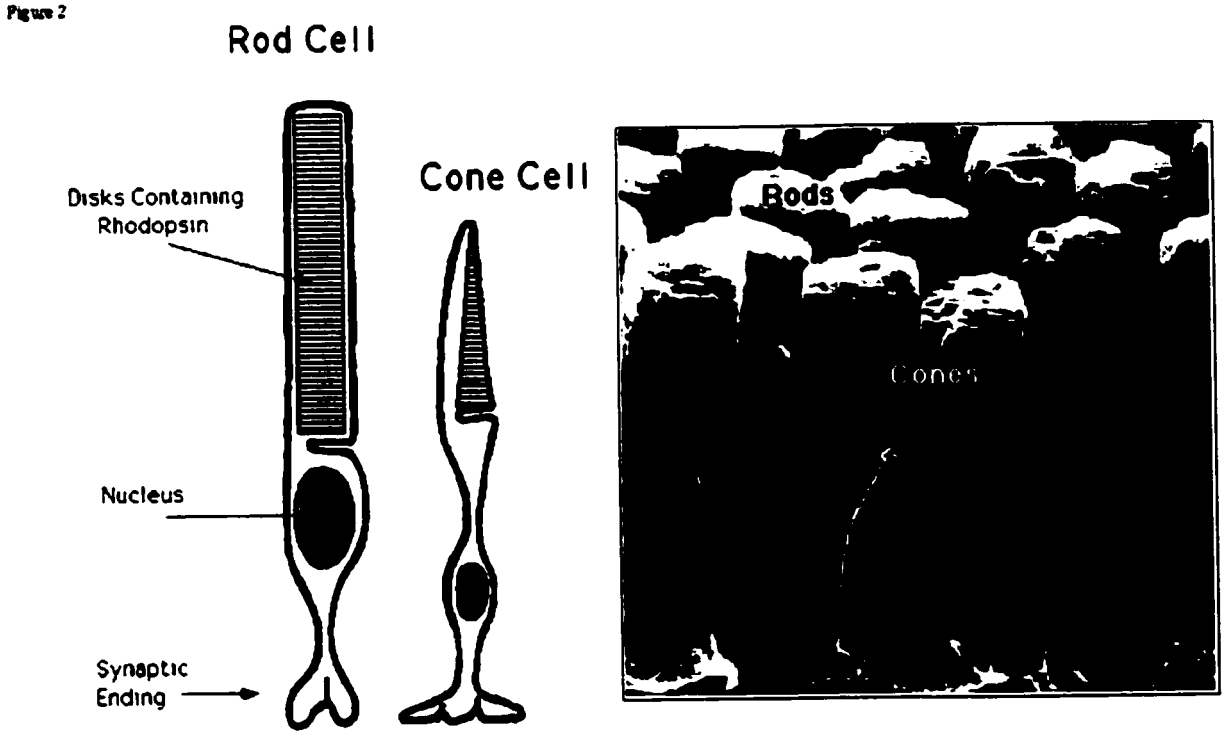


Fig. 4.2 a The Rod and Cone cells

Fig. 4.2.b. Rods and Cones

The information from the retina is transformed through the optical nerve (which has about 800,000 fibers) to the brain. Beyond the retina, the visual information takes place before reaching the brain in places called "The lateral geniculate", and "visual cortex" [Rob98]. A "ganglion cell" feeds each fiber in the optical nerve. Each ganglion cell is connected to hundreds of rod cells, and tens of cone cells. Fig. 4.3 shows the eye-brain connection.

The cones are in the middle of the retina in an area called (Fovea). At high light intensity, the cones have a high colorless visual acuity, and degraded color acuity. When light intensity reduces, the perception is shifted more towards areas with rods.

In the fovea area, every cone has a direct connection to a dedicated ganglion cell, and as well as sharing other ganglion cells with groups of cones and rods. This accounts for the high acuity of vision in the center of the visual field. This acuity degrades as the light intensity decreases [Rob98]. The eye perceives color as the various wavelengths or hues. The RGB wavelengths are: R: 700nm, G: 546.1 nm, B: 435.8 nm. Fig 4.4 shows the visible light wavelength.

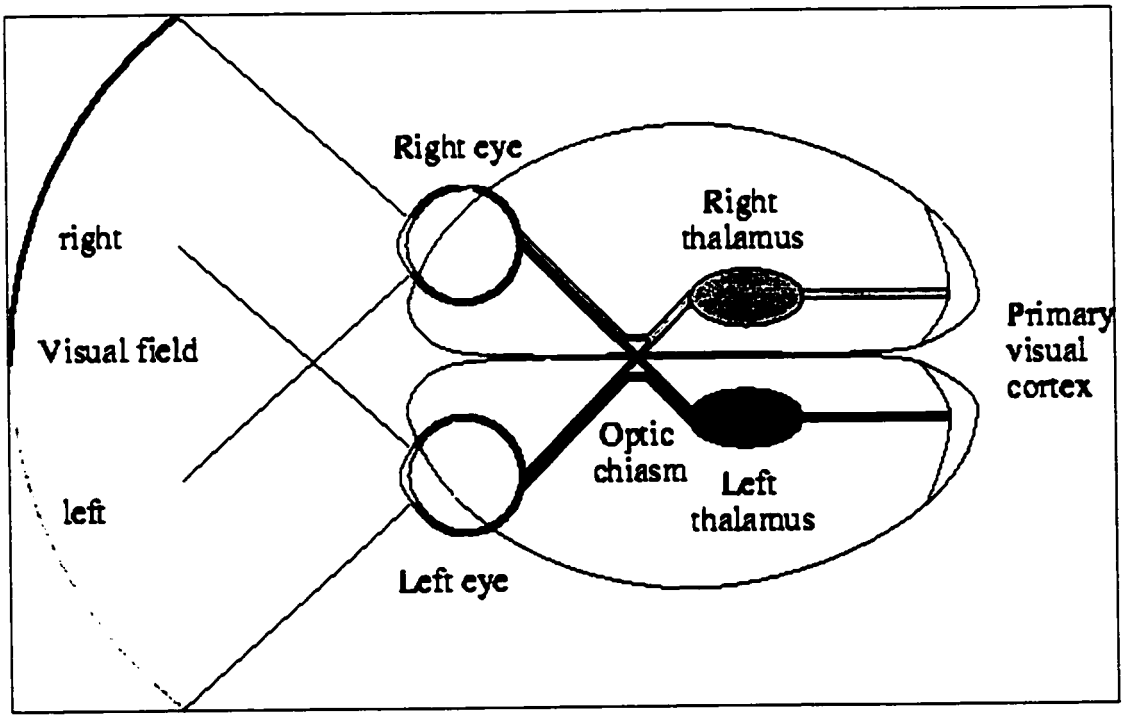


Fig. 4.3 The Human Visual System (HVS)

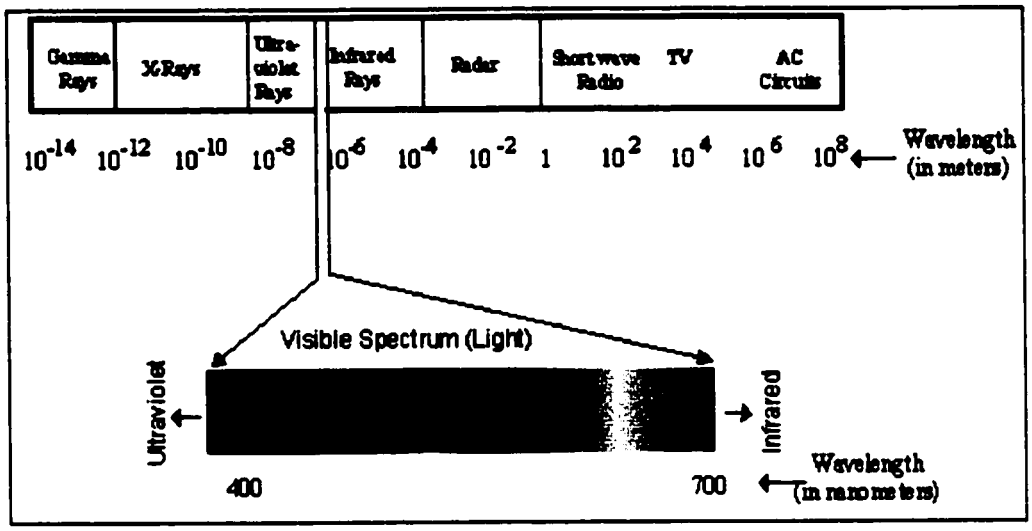


Fig 4.4 The Spectrum of Electromagnetic waves including Visible light

The human eye and the brain's visual cortex make up the Human Vision System (HVS), which can operate over a wide range of light intensities, detect color differences and understand picture contrast as a function of spatial frequency and light intensity.

Since the picture width and viewer distance influence the visibility of pictures, then the image frequency contents is depending on the viewer position. Therefore, it's important for specifying viewing systems to state the viewing distance, and resolution of display system (No. Of lines). The viewing distance for normal TV displays is 6H (where H is the picture height) [Rob98, Has97, Meer97]. This causes the full visual resolution of the details of the picture. However, for high definition systems, this distance is reduced to 3H only, and still maintaining the same level of spatial frequency eye discrimination.

The term "Visual Acuity" means the ability of the human eye to perceive and resolve details. It's measured by "the angle subtended by the smallest visible detail in an object" [Rob98]. In TV systems it's customary to calculate a one-minute of arc as the human acuity. To explain details in a picture, this is called resolution. Television resolution is equal to the number of alternating black and white horizontal lines that can be resolved vertically over the height of the picture.

Assuming normal viewing condition (viewing distance 6H, the vision acuity calculation formula is:

$$N = \frac{1}{\alpha n} \qquad [4.1]$$

Where N = Total number of elements to be resolved in the vertical direction.

Alfa α = Minimum resolvable angle of the eye (in radians)

$N = \frac{D}{H}$ (Viewing distance divided over picture height) [4.2]

If we assume α= 1 minute of arc or 2.91×10^{-4} radians, and n=6, then N becomes approx. 572 lines. This is the origin of why the 2 systems in the world picked numbers of 525 and 625 lines. Another characteristic of the HVS is that it doesn't necessarily perceive equally the digitized samples of a picture [Rob98], thus, the errors in this area do not affect the judgment of perceived quality by a viewer. That's the reason that removing some sample values could be done without affecting the quality of picture from a human perception.

The acuity of the human visual system depends on the following factors (Rob98):

1- The Luminance of the background. The visual acuity increases with luminance, to a maximum level of 100 foot/Lambert or 340 cd/m2 (candles per square meters).

2- The contrast of the picture luminance and chrominance. The image contrast is defined as the difference between the maximum and minimum light intensity to the sum of those two intensities.

$$Contrast = \frac{I_{max.} - I_{min}}{I_{max.} + I_{min.}} \qquad [4.3]$$

Thus picture details are only visible when there is a difference between them and the background. Fig. 4.5 shows the eye contrast sensitivity to spatial frequency in cycles /degree. As well, the eye contrast sensitivity varies with the temporal frequency of the picture (Fig 4.6).

The characteristics of the HVS in relation to spatial and temporal redundancies are summarized as follows [Rob98]; For Spatial Redundancy:

1- The eye is more sensitive to lower spatial frequencies, and less sensitive to higher frequencies. Actually, when the picture is electrically or optically transformed from one stage to another, the different spatial frequency components suffer from different responses. The amplitude of the high frequency components is reduced compared to the low frequency ones.

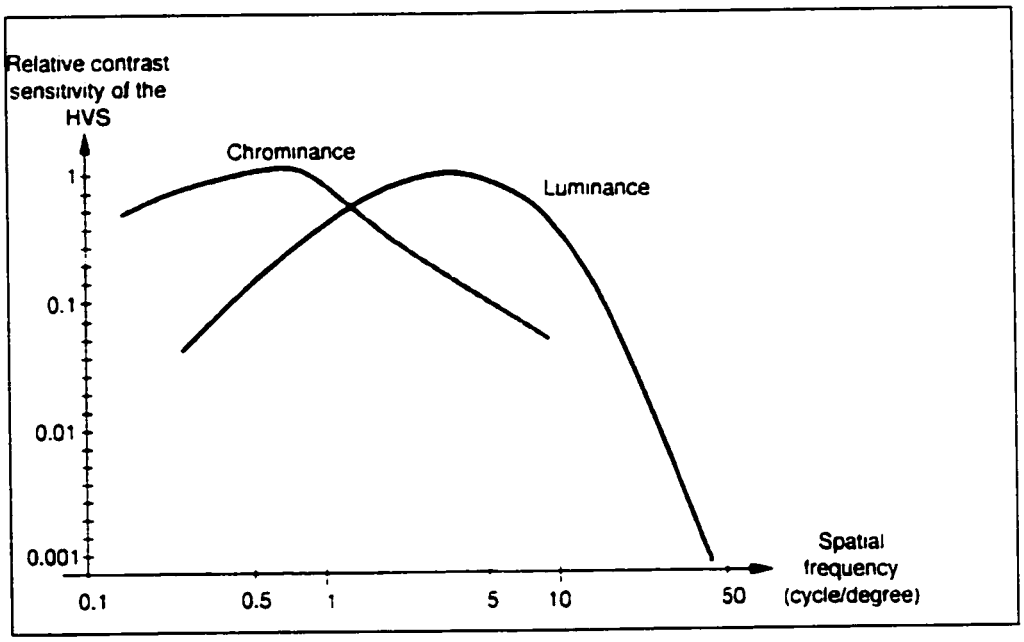


Fig 4.5 Relative contrast Sensitivity of the HVS with spatial frequency [Rob98]

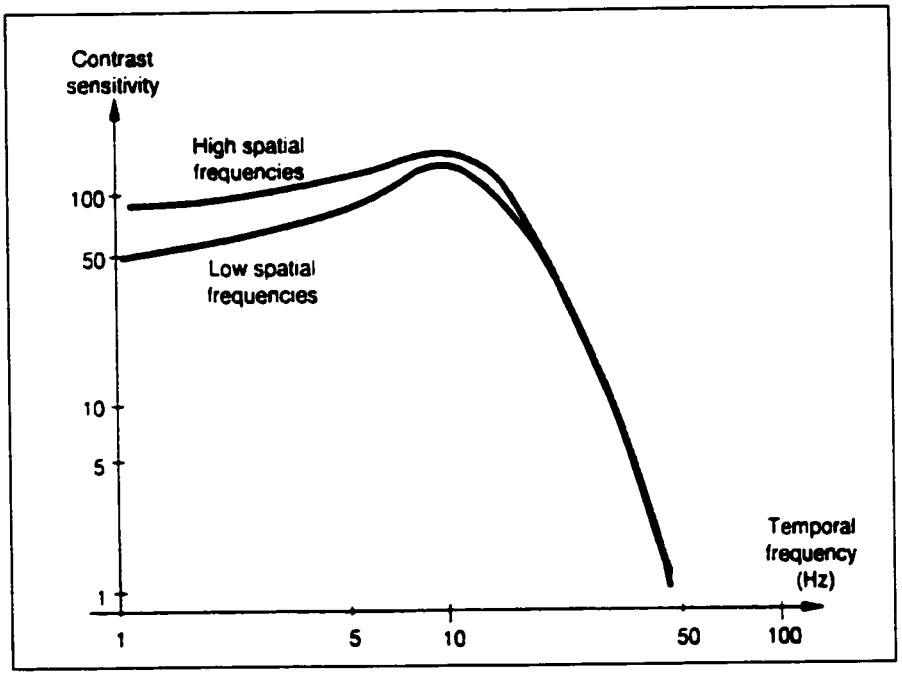


Fig 4.6 Variation of the HVS contrast Sensitivity to temporal frequency [Rob98]

2- Texture masking: The eye may not distinguish errors in textured areas or regions of the picture. And is very sensitive to distortions in uniform areas. The eye can see blocks in a blue-sky area, rather than the same blocks in a sports game with lots of details. In a textured region of the picture, the distortion from a DCT block is spread over many edges, which in turn prevents the detection of distortion a few pixels away from an edge.

3- Edge masking: This means that the contrast sensitivity is smaller in the neighborhood of an edge. Or in other words, errors are harder to see near the edges. This effect depends on the size, shape and duration of a stimuli [Meer97, Vass73]. Van der Meer [Meer97] states that the human visual system treats stimuli of different orientation or frequencies differently. The human visual system contains band-pass filters with a bandwidth of 1 octave, and an orientation of about 30 degrees. The visibility of distortion in a frequency band is independent from the distortion in other bands. [Olz86]. Van der Meer adds that the sensitivity of the human eye, which depends on the picture content, is called masking. The masking effects most often referred to are luminance, edge and temporal masking.

"Although edge masking was first introduced for true edge patterns, it has been shown to exist in each frequency band. It expresses itself as an elevation of the visibility threshold of a frequency orientation band when the contrast in the original band exceeds a certain value" [Daly93]. This is shown in fig 4.7.

4- Luminance masking: Visual threshold increase with background luminance. Van der Meer [Meer97] states that the contrast sensitivity increases with increased luminance. The important value here is the luminance as a function of the Grey background value. Fig. 4.8 shows luminance sensitivity as a function of background Grey value. Assuming a Grey level of 0 to 255 into a screen luminance L, and a correction for the characteristics of the display have been applied, then the relation between L and K can be expressed as:

$$L = 0.00025(K + 15)\ cd/m^2 \qquad [4.4]$$

But of course as different displays have different conversions, there can be different curves for luminance masking for different displays. As well, Van der Meer says "The sensitivity of the Human eye has a peak at mid-Grey levels".

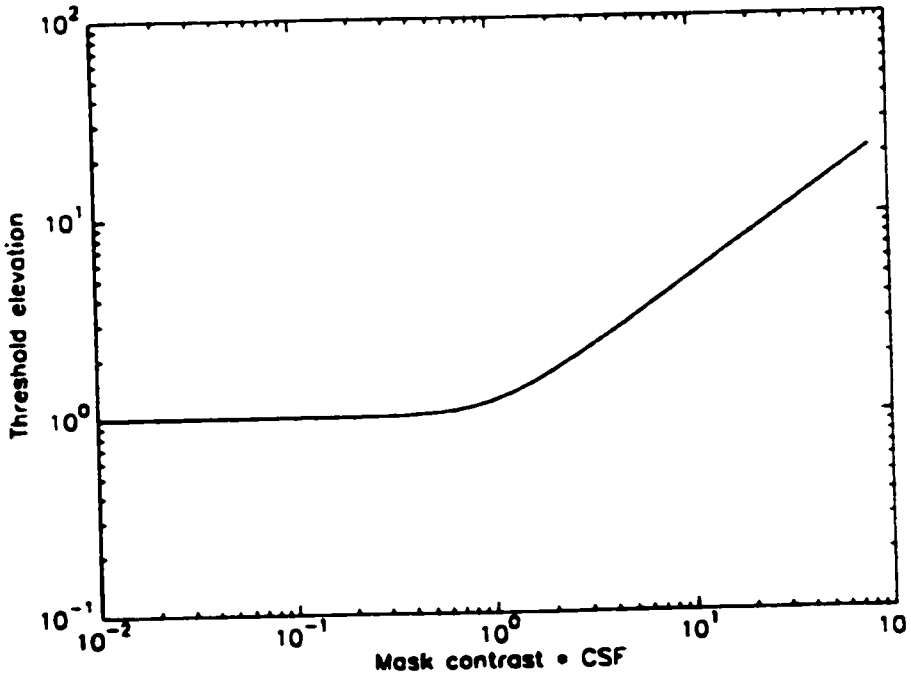


Fig 4.7 Threshold Elevation as a function of local contrast [Daly93]

5- Contrast masking: Where errors and noise in light regions of the picture are more difficult to detect. This refers to the reduction of visibility of one picture detail because of the presence of another.

As for Temporal Redundancy, the HVS is characterized by:

1- Temporal frequency sensitivity: Temporal masking results in a drop of sensitivity near temporal discontinuities"[Meer97]. Where flicker could be visible if frequency below 50Hz. Detection threshold of narrow lines increase four folds at temporal discontinuities [Car96].

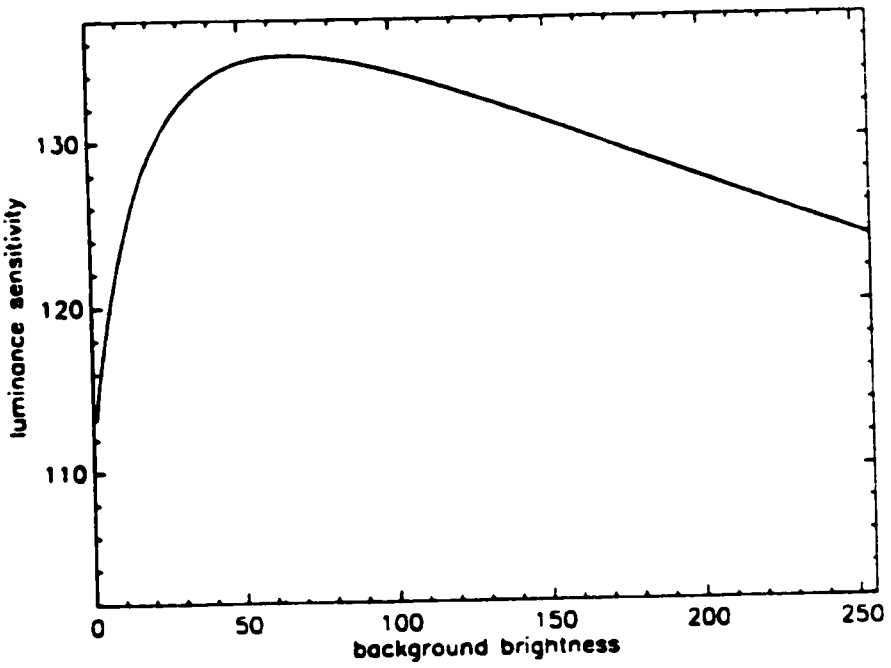


Fig 4.8 Luminance sensitivity as a function of Grey Value [Meer97]

3- Luminance masking: This refers to the fact that high luminance levels increase flicker.

4- Spatial Frequency content: Low spatial frequencies reduce the eye's sensitivity to image flickering.

The response of a system to different frequencies is called "Modulation Transfer Function"(MTF). Therefore, it's desired to characterize any HVS model response in an MTF. But it's not easy to do that. What can be measured is the "Contrast Discrimination", or the ability of the eye to observe differences in contrast. These measurement lead to "Contrast Sensitivity Function (CSF) characterizing the properties of the Human visual system. (See Fig. 4.7)

4.2 Video Artifacts

4.2.1 Analog Signal Artifacts:

Recommendation CCIR567 [Rob99] specifies the testing methodologies for composite analog television signal. The transmission, distribution and processing creates a number of impairments. The different elements making a complete distribution system including Amplifiers, D/A converters, VTRs, cabling, links, ...etc create additive impairments, and degrade the performance of the whole network. The analog impairments are categorized into 3 areas:

1- Linear Distortions

2- Non-Linear Distortions

3- Noise

Analog Composite Impairments	Linear Distortion	Frequency Domain	Frequency Response	
			Group Delay vs. Frequency	
		Time Domain	Luminance	Long Time
				Field Time
				Line Time
				Short Time
			Chrominance	Chrominance to Luminance Delay
				Chrominance to Luminance Delay
	Non-Linear Distortion	Luminance	Luminance Non-Linearity	
			Chrominance to Luminance Inter-modulation	
		Chrominance	Amplitude	Gain Non-Linearity
				Differential Gain
			Phase	Phase Non-Linearity
				Differential Phase
	Noise	Random		
		Coherent		

Fig 4.9 Analog Systems artifacts [Rob99]

There are 2 types of testing done on analog systems:

1- Out of Service testing: These are complete full field test signals, similar to tests done during equipment development.

2- In Service testing: Operational testing using test signals within unseen lines, called Vertical Interval test signals (VITs).

These measurements can be done using signal generators, VITs inserters, and analyzing test equipment, waveform monitors & vectroscopes.

4.2.2 Digital Testing Concepts:

Testing of the new digital compressed MPEG system is a totally different challenge than the analog one. The quality of compressed video depends on the distortion introduced during the compression stage.

The test equipment used for analog systems is useless for compressed video systems. Among the MPEG elements affecting the quality of picture are: The data rate (which is one of the most critical elements), The GOP structure (I, B and P frames), field/frame adaptive prediction, motion estimation and compensation methods, slice size and buffer size [Rob98].

The overall compressed video quality is a function of:

1- The Source material: because the perception of materials with high level of details like sports for example is different than plain "talking head" material.

2- Encoder Quality: How well the encoder codes a video stream. This includes the quantization levels and tables, motion and predictive estimation techniques used, as well as buffer size and usage.

3- The Data Rate: The higher the data rate, the more details that can be implemented, and better quality.

The testing of bit reduction systems is divided into 2 parts. The first one is to test the entire MPEG transport stream, and make sure it conforms to the canonical standard stream. To check the existence of synchronization signal, PCR...and different data structures, tables and flags in the stream, and check the integrity of the data after removing all the Reed-Solomon bits, and extracting the data packets.

Then the second level of analysis would be to check the quality of video and audio within that transport stream. Because the MPEG2 transport stream could be perfect, but the DCT, Quantization errors, and the many other errors may have distorted the signal to the extent, it would become annoying to watch.

On another level Van der Meer [Meer97] argues that the base functions of the DCT don't reflect pure frequencies. And for higher spatial frequencies, many DCT coefficients occur in the same band. As well, a DCT coefficient could appear in several different frequency bands of the Human Visual System [Klein92]. The exact visibility threshold for each DCT coefficient can be measured by means of a psychophysical model to measure the smallest value that gives a visible signal. [Pete91, Pete92]

4.2.3 Digital Signal artifacts:

CCIR Report 1089 states the classification of impairments associated with bit- rate reduction systems. Robin [Rob99] classifies the digital artifacts as follows according to CCIR 1089:



Fig 4.11 Original Picture (Susie) (One of the famous MPEG streams)



Fig 4.12 Original Indian Before Compression, with multiple errors from compression

A- Impairments associated with Intra-field coding:

1- Slope Overload: The rise time of the original signal cannot be matched causing blurred edges.

2- Edge Business: The precise continuity of an edge in the original signal cannot be matched and appears noisy.



Fig 4.13 Edge noise effect.

3- Contouring: The uniformity or monotonicity of the original signal cannot be matched causing a layering effect.

4- Granular Noise: Finely detailed portions of the picture (for examples those below threshold level) are not available, and the picture appears noisy.

5- Blocking: The checkered small boxes or underlying block structure sometimes called tiling. This is the most common and visible artifact of DCT transformation.



Fig 4.14 Blocking or Tiling effect due to the DCT

Due to temporal prediction inaccuracy, these artifacts may occur for interfiled and inter-frame coding:

1- Temporal Slope Overload: That happens when edges of fast moving objects cannot be matched and becoming blurred during movement.

2- Granularity and edge business: In movement, where fine detailed areas show granular noise and edge business.

There are other types of impairments caused by temporal sub sampling:

1- Jerkiness: There is a discontinuity in motion for example, where the smoothness of movement cannot be matched by the system.

2- Temporal Aliasing: Where high temporal frequency components are folded back.



Fig 4.15 Aliasing effect

3- Loss of resolution in moving pictures: Where spatial resolution is reduced during movement.

Subjective Evaluation of basic Quality	Intra-field Coding	Slope Overload
		Edge Business
		Contouring
		Granular Noise
		Blockiness
	Inter-field of Inter-frame Prediction Inaccuracy	Temporal Slope Overload
		Edge Business
		Granularity
		Blockiness
	Inter-field of Inter-frame Coding Inaccuracy	Jerkiness
		Temporal Aliasing
		Loss of Resolution in Moving Pictures

Fig 4.10 Bit rate Reduction Impairments according to CCIR1089 [Rob99]

One of the most annoying artifacts in compressed video is called "Mosquito Noise". This is a characteristic of all DCT systems, and appears on sharp edges (such as titles). The edges generate coefficients across the block. The high frequency coefficients are quantized more coarsely than lower frequency components. So, the energy is spread spatially over the block [Meer97, Sym98, Rob98].

Hence, although the block edge can be masked, the distortion from the quantization can be visible a few pixels away from the edge. Another example is around moving objects in front of a homogeneous background.

Back to texture masking, where for a textured region, the distribution of the DCT block is spread over a number of edges, which prevents the detection of distortion a few pixels away from the edges. When the DCT coefficients are not masked by texture, they become visible to the human eye as loss of resolution like un-sharp edges or blur.

And they could also be seen as artifacts like mosquito noise or blocking [Meer97, Sym98].

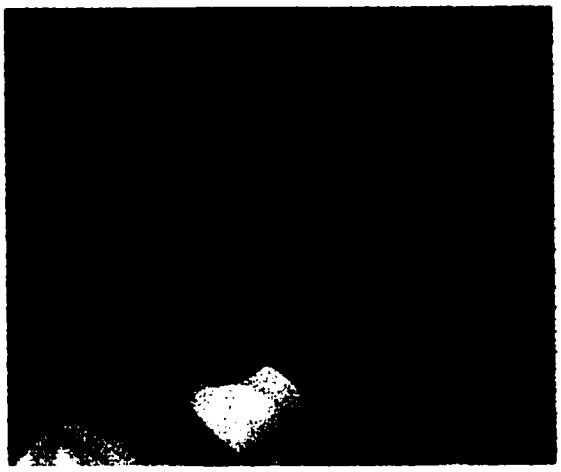


Fig 4.16 Blurring effect

Other artifacts are like:

- Dirty Window [Sym98]: This artifact occurs as streaks or noise that remains stationary, while objects appear to move behind. This looks like viewing

something behind a dirty window. Insufficient bits allocated to the code prediction vectors cause it.

- Wavy Noise [Sym98]: Similar to mosquito noise, this is caused by coarse quantization of high frequency coefficients. This is seen during show pans across a very detailed background scene. As an example, the crowd or spectators at a sporting event, the motion here causes the spreading to vary periodically as the details move across the DCT block.

4.3 The ITU-R.BT-500 standard and Subjective Quality

The first and most reliable technique to evaluate compressed video quality after reconstruction is through means of subjective testing. That is assessing the quality through the evaluation of a sample of viewers.

The CCIR-Rec.500, now known as ITU-R BT-500 sets the standard for all subjective measurements related to compressed video systems. The latest version of that standard is ITU-R BT.500-10 issued on March 2000. This standard describes the different methods and conditions to set up for subjective quality testing. These conditions include:

1- The general viewing conditions like luminance of inactive screen to peak luminance, display brightness and contrast, room illumination, monitor resolution, monitor contrast...etc. There are 2 standards that could be followed, one for lab. environment, and another one for home environment. This creates the potential split of quality standards.

2- The choice of source signals and test material. Particular test material should be used to address particular assessment problems.

3- The choice of observers. ITU-500 states that there should be at least 15 viewers, who are not experts. Prior to the testing, viewers should be screened for normal visual acuity

on the Snellen or Landolt chart [ITU500], and for normal color vision using specially selected charts.

4- The structure of the test session, length of test sequences followed by Grey screen for a number of seconds after each sequence...etc.

However, the most important concept introduced by the ITU-500 standard is the methods used for quality evaluation. They are described as follows:

1- **The Double Stimulus scale (DSIS) or (The EBU method):**

The meaning of double is the evaluation of a reference or clean video with no impairments, and then the same video sequence after being impaired by compression for example. In the DSIS method, the viewers are always shown the clean video source first, then the impaired one. Then they are asked to mark the quality of the second impaired stream, putting in mind the quality of the first over a scale of either 5 level grades or 1-100 scale. In a test session that may last for 30 minutes, the viewers are shown many impaired sequences covering all the required impairment combinations. After the session, the mean scores for each test condition and test picture is calculated.

2- **The Double Stimulus continuous Quality Scale (DSCQS) method**

Another method for evaluating a new system, or the effects of transmission impairments on video quality. This is a bit different than DSIS, in that the viewers are shown pairs of pictures, one is clean, and the other is impaired of the same source material. However, the choice of which comes first is random, so in one time they may see the clean source first, then the impaired one, and in another it could be vice versa. The viewers are asked to evaluate both clean and impaired sequences. The ITU describes a stream sequence of 10 seconds long, with 3 seconds of Grey in between each sequence and the next one.

B- The Single Stimulus (SS) methods:

In this method, a single image or sequence of video is presented, and the viewers give assessment of that one only without any reference stream. SS methods are divided into 3 categories:

B.1 Adjectival categorical judgment methods

In this method, viewers assign a video sequence to one of a set of categories, which are defined in semantic terms. Categories are related to detection of a certain attribute like establish impairment threshold for example. The list of categories is shown in table 4.1.

Five-grade scale			
Quality		Impairments	
5	Excellent	5	Imperceptible
4	Good	4	Perceptible, but not annoying
3	Fair	3	Slightly annoying
2	Poor	2	Annoying
1	Bad	1	Very annoying

Table 4.1 Quality and Impairment scales [ITU500]

B.2 Numerical categorical judgment methods

This method uses an 11-grade Numerical Categorical Scale (SSNCS). The study on this method is described in report (ITU-R BT.1082).

B.3 Non Categorical Judgment methods:

In this case, the viewers assign a value to each image sequence. There are 2 forms of this method:

B.3.a Continuous scaling: Where viewers assign each image sequence to a point on a line drawn between 2 semantic labels (as in table 4.1).

B.3.b Numerical Scaling: Where viewers assign each video sequence a number reflecting it's value on a specified dimension (image sharpness for example). This scale could be 0-100 for example.

3- Stimulus-comparison methods:

There are 3 stimulus comparison methods used in TV quality evaluation:

3.a Adjectival categorical judgment methods

In this method, viewers assign a relation between members of a pair to one in a set of categories, which are usually defined in semantic terms. The scale is a 7-step scale as shown in table 4.2

-3	Much Worse
-2	Worse
-1	Slightly Worse
0	Same
+1	Slightly Better
+2	Better
+3	Much Better

Table 4.2 Comparison Scale

3.b Non-categorical judgment methods:

Here, viewers assign a value to the relation between the members of the pair. The evaluation could be either in continuous scaling like a point between 2 labels "SAME- SLIGHTLY BETTER". Or it can be expressed in numerical value.

4- Single Stimulus Continuous Quality Evaluation (SSCQE)

In this method, viewers have a slider mechanism unit that is 10 cm, with the 5 grades scale, and a 0-100 numerical scale for example. The viewer would use this scale to continuously change the value depending on the scene, and the unit would capture the readings twice per second. This unit is shown in fig. 4.17

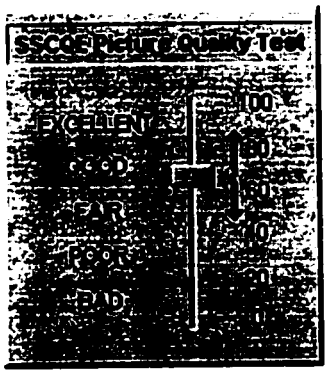


Fig 4.17 The slider mechanism evaluation device

The SSCQE describes conditions for the tests of TV material over longer periods of time that double stimulus systems. The proposed times here are program segments of 5 minutes long at least. The program segments should correspond to specific programming types like sports, news, drama, cartoons...etc. It also suggests a test session length of 30-60 minutes long.
The ITU-500 standard describes in detail the process of data interpretation and analysis to obtain comprehensible information from the data collected from viewers. They are mathematical and statistical techniques calculating the mean average value for each video sequence, then creating sort of confidence level of readings, by eliminating readings within far difference than the other viewer readings.
For each test presentation, the mean score value is calculated, as well as standard deviation and kurtosis coefficient [ITU-500]. From these results further mathematical analysis can be made to create relations between mean scores of streams to certain impairments studied.

4.4 Objective Quality Evaluation:

The objective methods of quality evaluations are made with close correlation to the ITU-500 recommendation in subjective testing. There are 3 agreed upon methods [Rob99]:

4.4.1 Picture Comparison method:

Sometimes this is called double stimulus or Full reference method. This method compares a source video, which is uncompressed (ITU-601 270 Mbits/sec) to the same sequence but impaired from a compression process. After decompression back to a 601 stream, the 2 streams are compared using mathematical processing over each picture. Since the system has the full information of both streams, it could present accurate results.

This system is usually used for MPEG encoders, decoders or Codec evaluations. It's also used in out of service testing of different transmission links like cable, fiber or satellite links. In this case, probably another measuring system must be available at both ends of the link, with the same streams, and synchronizing them to generate valid results. This technique is shown in Fig. 4.18

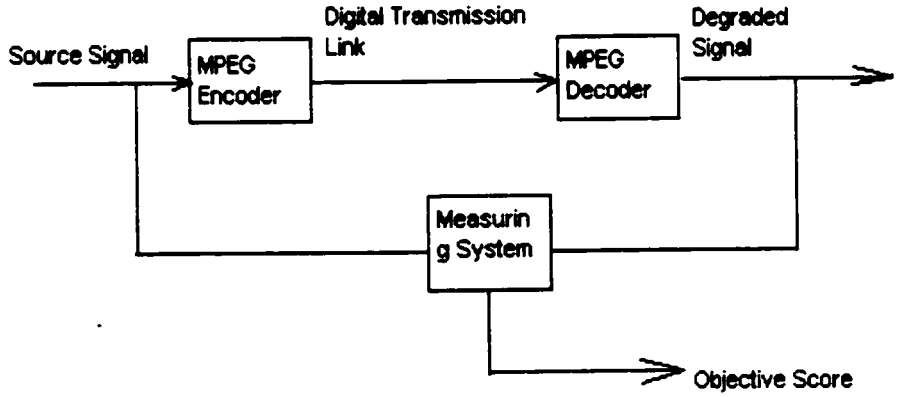


Fig 4.18 Picture Comparison or Full reference System

4.4.2 Feature Extraction Method:

Sometimes called Reduced Reference Method. In this method, the measuring system extracts a reduced amount of features from the degraded signal, and compares that to the same features from the original un-impaired signal. The system then compares the two, and generates an impairment measurement results. This system block diagram is shown in fig. 4.19.

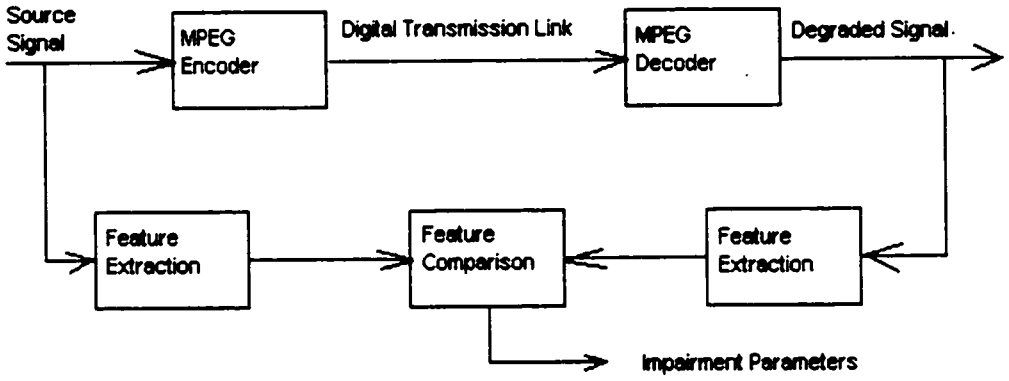


Fig 4.19 Feature Extraction or reduced Reference System

4.4.3 Single ended testing method:

Sometimes called Single Stimulus testing or no reference testing. This system analyzes the received signal based on known impairments or artifacts from MPEG compression, and generates a rating of the quality based on those impairments. This system is shown in Fig. 4.20

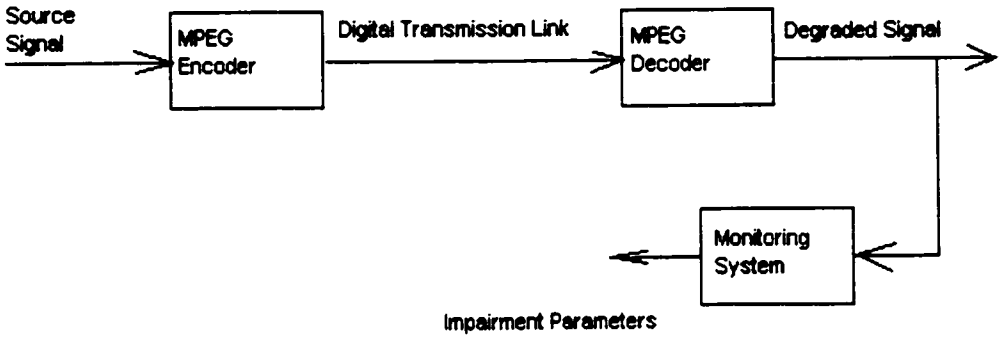


Fig 4.20 Single Stimulus or No Reference System

4.5 Quality Metric

To objectively evaluate the quality of reconstructed video after being compressed, or using equipment rather than subjective human viewers, we need to establish metrics for quality that are a function of both the original non-impaired and the compressed signals. One of the widely used quality measures is called the Mean Squared Error MSE:

$$MSE = \frac{1}{N_r * N_c} \sum_{x=0}^{N_c-1} \sum_{y=0}^{N_r-1} \left(p(x,y) - p\hat{}(x,y)\right)^2 \qquad [4.5]$$

Where p and p^ are the original and reconstructed video. Nr is number of rows, and Nc is the number of columns. And from this measure, there are 2 Signal-to-Noise ratio SNR measures that can be calculated:

$$SNR_p = 10 \log \frac{(255)^2}{MSE} \qquad [4.6]$$

And $$SNR_v = 10 \log \frac{\sigma_s^2}{MSE}$$ [4.7]

Assuming the picture was quantized by 8 bits per pixel, 255 is the maximum intensity, σ_s^2 is the variance of the original picture. SNRp is called peak-to-peak SNR and SNRv is variance SNR. These MSE and SNR are simple equations that reflect the distortion in the picture. However, they cannot reflect the frequency distribution of the errors, and thus don't reflect the human quality perception [Meer97]. The reason is human visual system is frequency dependent.
A way to compensate for that is using the "Parseval Theorem" and calculating the MSE in the Fourier domain to get the different frequency components [Katt91]:

$$WMSE = \frac{1}{N_r * N_c} \sum_{u=0}^{N_c-1} \sum_{v=0}^{N_r-1} \left(w(u,v)\left(P(u,v) - P\hat{}(u,v)\right)\right)^2 \quad [4.8]$$

Where w (u, v) are the weighing coefficients according to the CSF of the human eye. P (u, v) is the Fourier transform coefficients of the picture for frequencies (u, v). This function can be done in the DCT domain rather than the Fourier one, and in this case, the P (u, v) is replaced by the DCT coefficients c (u, v) and w (u, v), which are the visibility thresholds for the individual DCT coefficients.
The MSE and SNR metric is the simplest of quality metrics. Other metrics that consider the masking of the Human visual system are more complex. There are 2 classes of metrics Quality [Bass96]:

4.5.1 Metrics based on the Human visual system

In this case, the HVS is modeled, splitting the picture into frequency bands like the HVS. We then calculate the contrast from the pixel values from each band. The contrast can be defined

as "Weber contrast" [Olz86] for stimuli that are symmetric relative to the background luminance as:

$$C_w = \frac{\Delta L}{L} \qquad [4.9]$$

Where L is the luminance difference in the pattern, and L is the background luminance. One HVS system normalizes the Local Band-limited Contrast (LBC) to the visibility threshold [Wes95]. Other models express the signal in just noticeable difference (JND) [Jay93]. Adjusting the visibility threshold incorporates the edge and texture masking. The distortion is then measured in each pixel in each of the frequency bands. One of the models referred to by Van der Meer [Wes95] measures JND via the function of Masked LBC difference $MLBC_{K,I}(x, y)$:

$$JND_{K,I}(x,y) = \Delta MLBC_{K,I}(x,y) = \frac{LBC_{K,I}(x,y) - LBC^{*}_{K,I}(x,y)}{TE_{K,I}(x,y)} \qquad [4.10]$$

Where $LBC_{K,I}(x, y)$ and $LBC^{*}_{K,I}(x, y)$ are the local band contrast for the original and impaired video sequence. TE is a threshold elevation function as shown in fig 4.7.

After measuring visibility of difference between original and impaired picture for each pixel each frequency band, we should try to get a measure for the quality of the whole picture. One way to do that is by pooling of errors means (PEM). [Wes95] describes a method for calculating that:

$$PEM = \left(\sum_{x,y} \left| \sum_{k,I} \left| JND_{K,I}(x,y) \right|^{\alpha} \right|^{\beta} \right)^{\gamma} \qquad [4.11]$$

In this case, α, β, γ are constants that affect the relation between JND and the PEM. They could be used to make the correlation between this metric and the subjective measurement values closer.

4.5.2 Metrics based on Visual Artifacts:

This is the other approach for measuring quality, by measuring artifacts in compressed video like blocking, blurring, false contour...etc. The subjective evaluation is based on a combination of those artifacts affecting compressed video.

Miyahara [Miya92] developed a model that extracts five features from a stream of video. A simple model is made after that. Then it's tuned to a maximum correlation to subjective evaluation.

Other models reflected another combination of features. These features could be related to spatial distortions, changes in motion energy. Sometimes these featured are combined linearly in a metric. In other cases like stated by [Meer97]. [Lin95] used a non-linear back-propagation neural network to combine different features in a quality metric. This area of research is still open for new approaches.

Chapter 5

Algorithms of Video Quality

The models for digital video quality evaluation fall under the 3 categories described earlier:

1- Double Stimulus, Full Reference

2- Feature Extraction / Reduced Reference Models

3- Single Stimulus / No Reference Models

There has been extensive work done in this area, and many algorithms introduced. Below is a description of some of the already introduced models. There has been effort done in the past 2 years in the single stimulus or no reference models. This is for the need of TV networks to do on-line monitoring of their programming. Research in this area by [Cav01, Ham01, Kne01 & Tan01].

5.1 Double Stimulus Systems (Full and Reduced Reference FR/RR models:

5.1.1 JND Model (Just Noticeable Difference), Sarnoff Model, USA

The process of digital video reproduction entails the introduction of distortion and artifacts through the whole process. This includes the processes of initial capture (by camera for example), encoding, transmission, decoding and display.

The JND model introduced by Sarnoff makes predictions of perceptual ratings of a degraded video stream compared to a reference or clean one of the same material. It's based on a spatio-temporal human visual model. The difference of the 2 signals is quantified in units of the human Just Noticeable Difference (JND) [Sar97, Lubin95].

The basic block diagram of the system is shown in Fig 5.1.

The reference signal is un-distorted and the test sequence is a degraded signal, which passes through the "System Under test". This in turn could be one of a number of sources of distortion like an encoder, the transmission channel, or the decoding.

The model produces a sequence of JND maps.

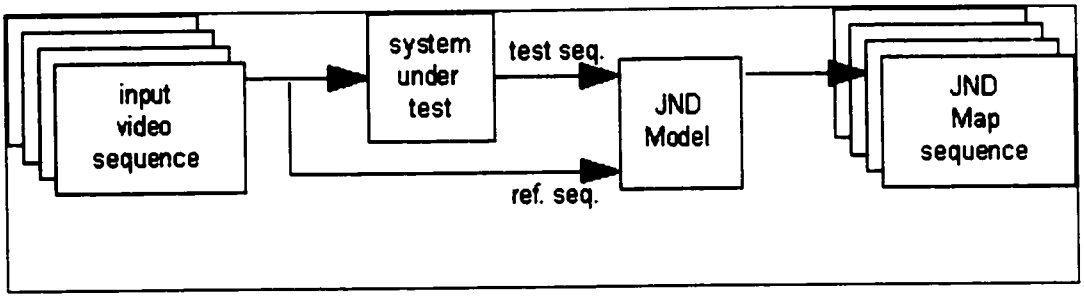


Fig 5.1 Block Diagram of JND Model [Sar97, Lubin]

Fig 5.2 shows the architecture of the model. As mentioned 2 sequences are entered to the system, one reference, and the other processed sequence. In the first stage the processing transforms the Y', Cb', Cr' into R', G' and B' gun voltages. After that point nonlinearity is applied to those values. This models the transform from R', G', B' gun voltages to model intensities (R,G and B) of the display (fractions of maximum luminance) [Sar97]. After nonlinearity, the vertical-electron beam is modeled by replacing the interline values in fields R, G, B by interpolated values from above and below. Then the vector (R, G, B) is subjected to a tri-stimulus coordinates (X, Y, Z). The luminance component Y of that vector is passed to the luminance processing part of the model.

And to ensure perceptual uniformity at each pixel, of the color space to isoluminant color difference, the model maps the pixels into CIELUV (An international standard uniform –color space). The chroma components u*, v* are passed to the chroma processing step.

After that, in Pyramid Decomposition, each luma value is subjected to a compressive nonlinearity. Then each luma field is filtered and down-sampled in a four level Gaussian Pyramid [Burt83] to model the psychophysically and physiologically of visual signals into different spatial-frequency bands. This process efficiently generates a range of spatial resolutions for subsequent filtering operations. After that, there are similar operations like oriented filtering done at each pyramid level.

Next, the Normalization stage sets the overall gain with a time-dependent average luminance, to model the visual system's relative insensitivity to overall light level, and to represent such effects as the loss of visual sensitivity after a transition from a bright to a dark scene or insensitivity to distortion in busy areas. [Nach74].

After normalization, three separate contrast measures are calculated. In each case, the contrast is a local difference of pixel values divided by a local sum, appropriately scaled as a function of pyramid level so that the result is 1 when the image contrast is at the human detection threshold. This establishes the definition of JND, which is passed on to subsequent stages of the model.

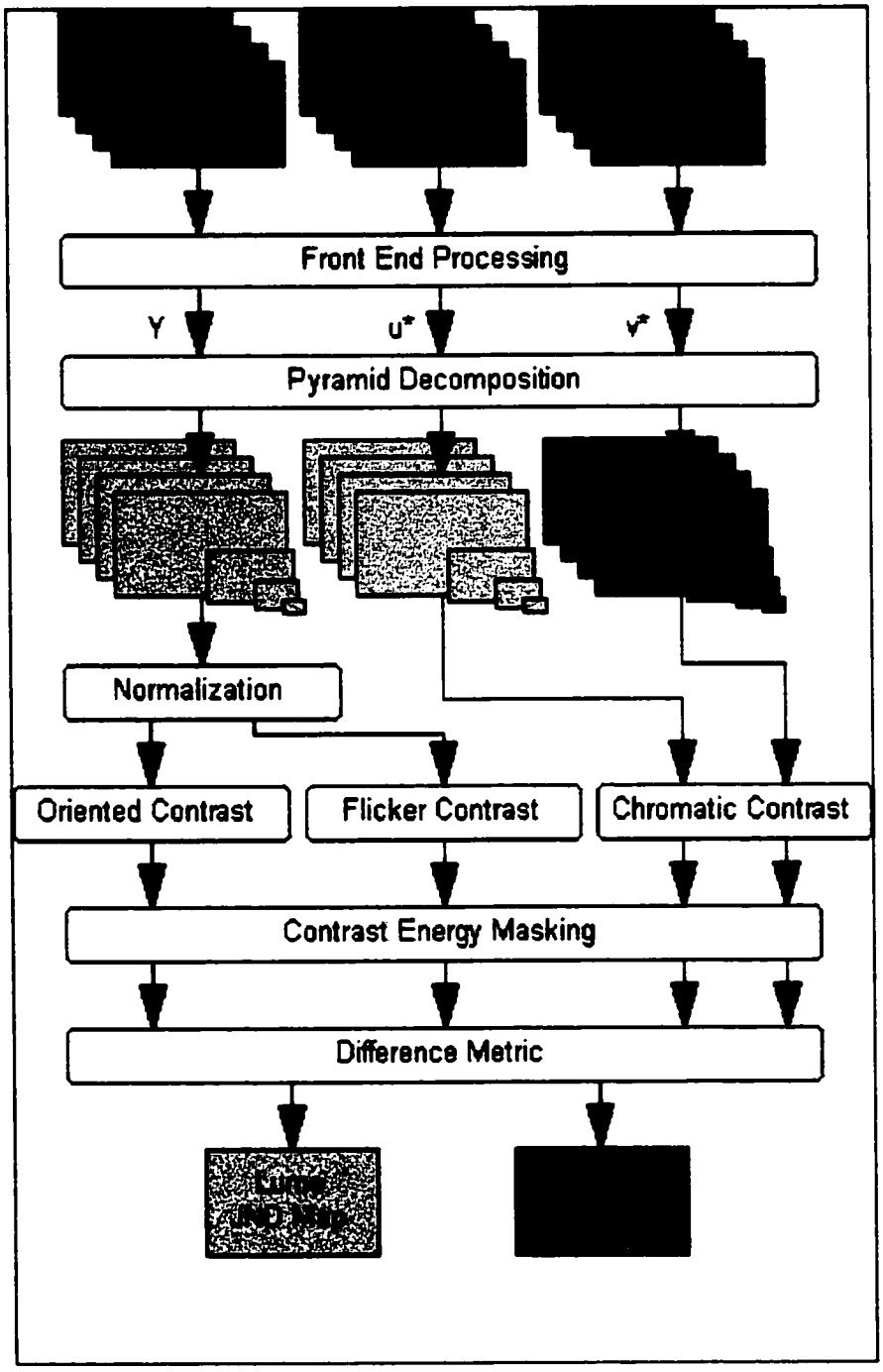


Fig 5.2 JND model Architecture [www.mpeg.org/MPEG/JND]

In the Contrast Energy Masking stage, each contrast image is subjected to a point non-linearity, the gain of which is controlled by the response across other resolution levels and channels. This gain-setting is included to model visual masking effects such as the decrease in sensitivity to distortions in "busy" image regions.

In the Difference Metric stage, outputs from the test and reference sequences are combined via a simple difference operator and then summed across pyramid levels and channels to return the number of JNDs in both luma and chroma.

Separate JND maps for luma and chroma can then be combined into one map. Summary statistics can also be obtained at this point.



Fig 5.3a Reference Image

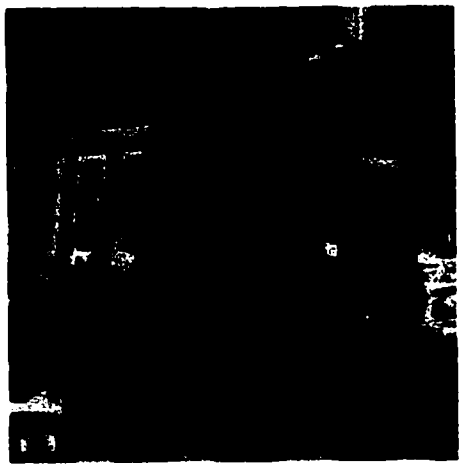


Fig 5.3b Processed Image

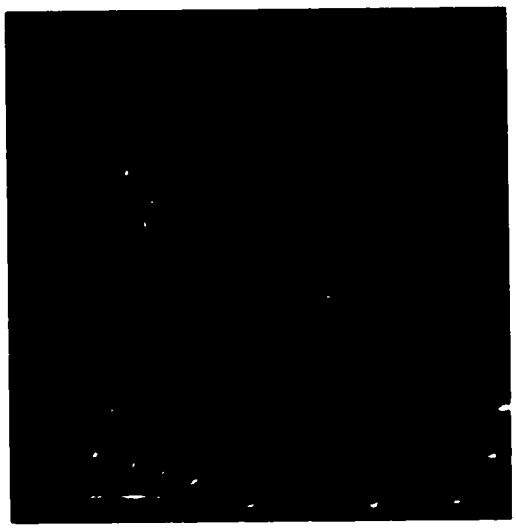


Fig 5.3c The difference map or JND map output according to the model

As well, the luma and chroma JND maps are reduced to one number, luma and chroma Picture Quality Ratings (PQRs). This is done by doing a histogram on the JND values for all the pixels over a threshold, then adopting a 90 percent value as the PQR value [Sar97]. The luma and chroma PQR numbers are combined by a linear combination of a sum and maximum to get the PQR for the field being processed.

5.1.2 The NTIA model, USA

This model is used for in service monitoring of video quality. It extracts some features from the input and output video scenes and compares them concluding the degradation in quality. These features are transmitted from transmission side to the receiving side by an ancillary data channel. This is shown in fig 5.4

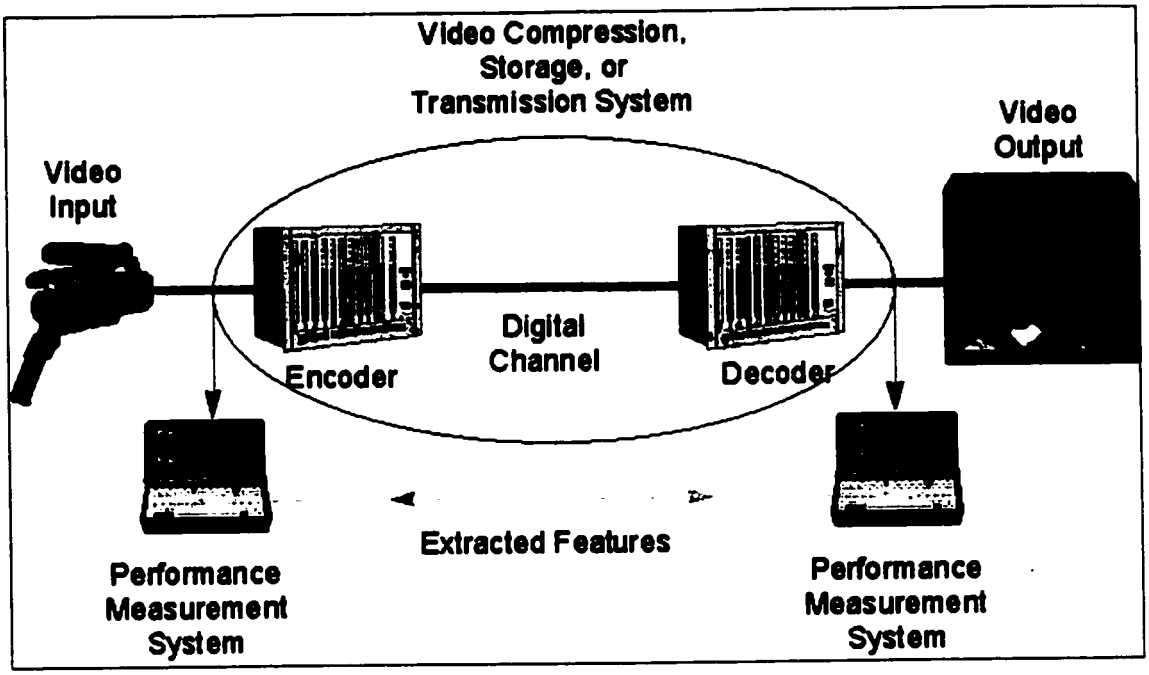


Fig 5.4 The Extracting Feature Approach for Quality Measurement [Wolf98]

For that, it uses a concept of Spatial - Temporal (S-T) regions from the pictures. Fig 5.5 gives a region of 3 pixels by 3 lines in a consecutive 3 frames as the S-T region considered here. The S-T region can vary. Large S-T regions can be used when the ancillary data channel bandwidth is small, while small regions are used when the ancillary data channel bandwidth is large. As more features are associated with smaller regions, this will need more bandwidth.

The way the system works is very straightforward. There is an input and an output calibration processor. The input processor estimates and adds the video transmission system delay to the input stream to synchronize with the output stream. This is achieved by correlating low bandwidth temporal features like motion features between the input and output calibration processors [Wolf98].

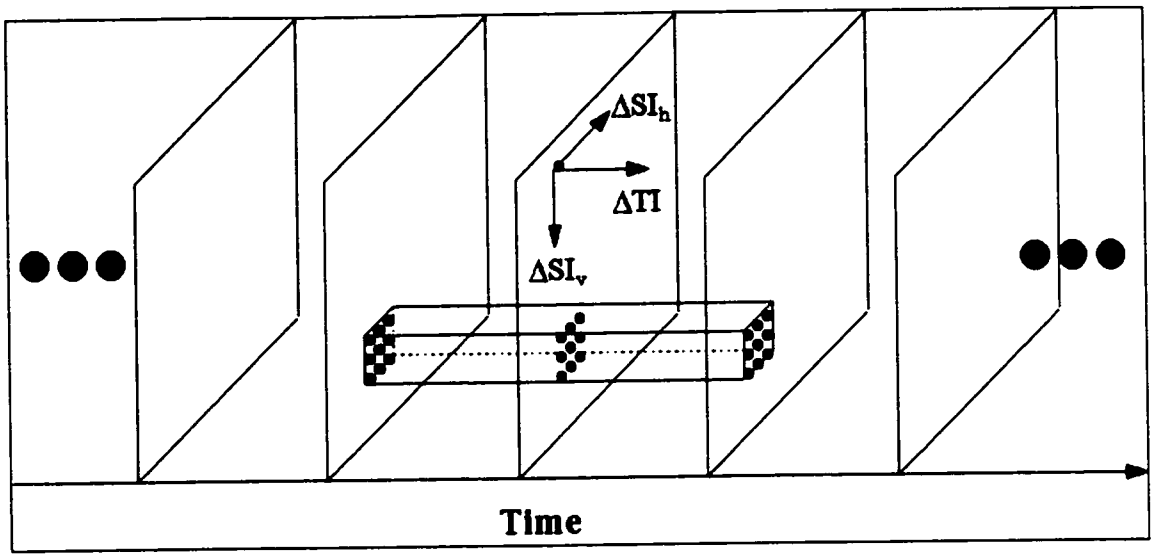


Fig 5.5 Sample Spatial Temporal region (S-T)

After this time alignment is done, the output calibration processor estimates the spatial shifts, gains and level offset of the transmission system, and applies the corrections to the output video stream.

After that, there are 4 programmable filters to extract features from the calibrated input and output video. They are spatial, temporal, spatial-temporal and chroma. The filters are programmable; meaning they can extract the features from variable S-T regions. The size of the S-T region can be changed based on the bandwidth of the ancillary data channel. The features chosen after extensive research to characterize the activity of motion, edges and colors.

The Spatial feature gives the activity of image edges. The digital video system can add edges creating edge noise, or reduce edges causing blurring. The temporal feature gives the temporal difference. The video system here can add motion causing blocking, or reduce motion causing frame repeats or picture freeze. In the chrominance feature, the video system can add color causing cross color or color artifacts, or reduce color causing color sub-sampling. The spatial temporal feature is a cross between the 2. An example of added spatial temporal artifact is mosquito noise, which is visible in stationary background around moving objects [Wolf98].

The benefit of this system is that it can do in-service measurement of quality without knowing the input video scenes, and needs only an ancillary data channel for the transmission of features.

5.1.3 Image Segmentation Model- CPqD-Brazil

This model assesses video quality using objective parameters based on image segmentation. Natural scenes are segmented into plane, edge and texture regions. A set of objective parameters are assigned to each of these contexts. Fig 5.6 shows the configuration of the objective parameters computation used.

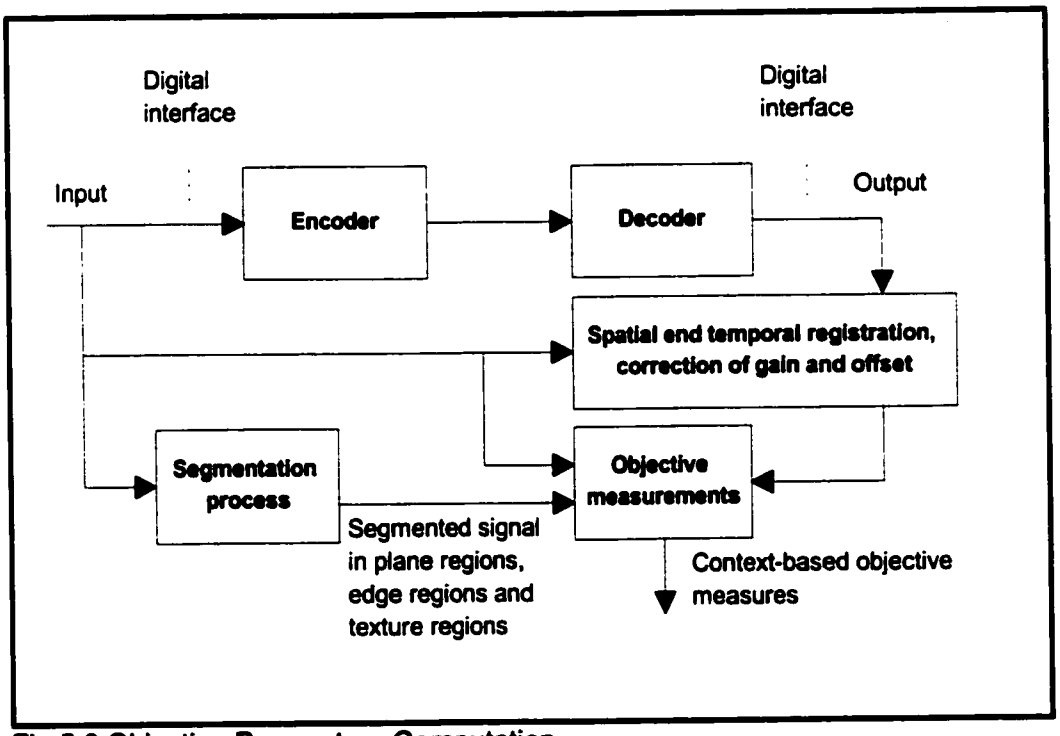


Fig 5.6 Objective Parameters Computation

Each objective parameter is computed separately within the context of plane, edge and texture regions. For example, a blocking distortion can be measured by an edge detector applied to the plane regions of the video scenes.

This type of algorithm usually needs a high computational complexity. However, this is reduced by using low complexity estimators, and by limiting their computation within the context of the scenes [Pess97].

There is a process of alignment needed, where the spatial and temporal registration between the input and output video, and the correction of gain and offset are estimated.

The objective parameters are computed by comparing the original to the impaired video sequences. The estimators are applied to fields rather than frames to ensure the reliability of the measurement in high level of motion scenes.

A perceptual based model, that predicts subjective rating, is defined by computing the relationship between objective measures and the results of subjective assessment tests applied on a set of natural scenes and MPEG-2 codecs. The scene-dependent perceptual models are defined in 2 steps:

A- The relationship between each parameters and the subjective impairment level is approximated by a logistic curve.

B- The final result is achieved by linearly combining the estimated impairment levels, where the weight of each impairment level is proportional to its' statistical reliability.

For spatial segmentation, there has been 3 algorithms developed:

1- Image segmentation based on edge detection using recursive filtering

This algorithm classifies each pixel based on the brightness variance computed within a neighborhood of the pixel. The resulting image is then smoothed by a median filter. The algorithm applies an edge detector on the image based on recursive filtering. The edge on the boundary of the plane regions is classified as belonging to the edge regions. The texture regions are the remaining regions of the picture.

Fig 5.7a,b show the mobile and calendar scene after segmentation. The plane regions are represented by white pixels, edge regions by gray pixels, and texture regions by black regions.



Fig 5.7a Part of Mobile & Calendar

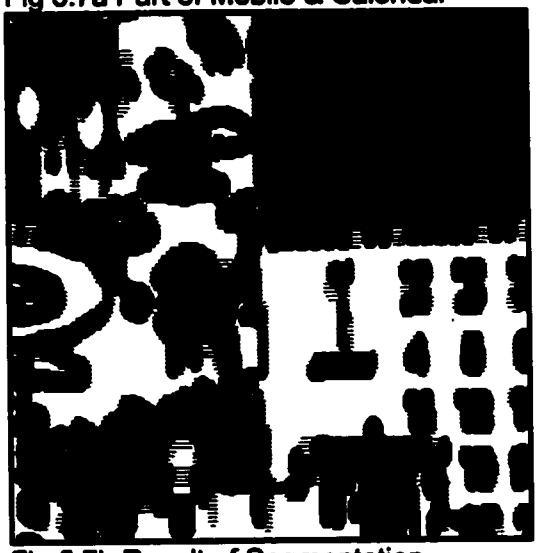


Fig 5.7b Result of Segmentation

2- Fuzzy Image segmentation based on spatial features:

This algorithm is divided into 2 steps. In the first step, the algorithm assigns a membership function to each of the 3 contexts under classification. For plane regions, the value of a pixel is defined inversely to the brightness variance within a neighborhood of the pixel. There is a "Morphological gradient" [Pess97] applied to define the membership function of the edge regions. The compliments of the fuzzy union between these 2 functions define the membership function of the texture regions. In the second step, each pixel is classified as belonging to context with highest value of membership among its' 3 membership values.

3- Image segmentation based on watershed:

In this algorithm, the luminance component is simplified by applying edge smoothing filter, increasing its' homogeneous regions. Then a "watershed algorithm" is applied to the morphical gradient of the simplified image. The watershed detects homogeneous regions, called catching basins. The plane regions are the catching basins with area greater than a threshold. The texture regions are given by the erosion of the complement of the plane regions, and the edge regions are the remaining ones.

Pessoa [Pess97] states that by using region-based objective measurements, more accurate predictions are obtained, compared on predictions based on global parameters.

5.1.4 DVQ model- NASA, USA

This algorithm accepts a pair of digital video sequences, and computes a measure of the magnitude of the visible difference between them. The metric is based on the Discrete Cosine Transform. It incorporates aspects of early visual processing, including light adaptation, luminance and chromatic channels, spatial and temporal filtering, spatial frequency channels, contrast masking, and probability summation. It also includes primitive dynamics of light adaptation and contrast masking. Fig. 5.8 shows a block diagram of the model.

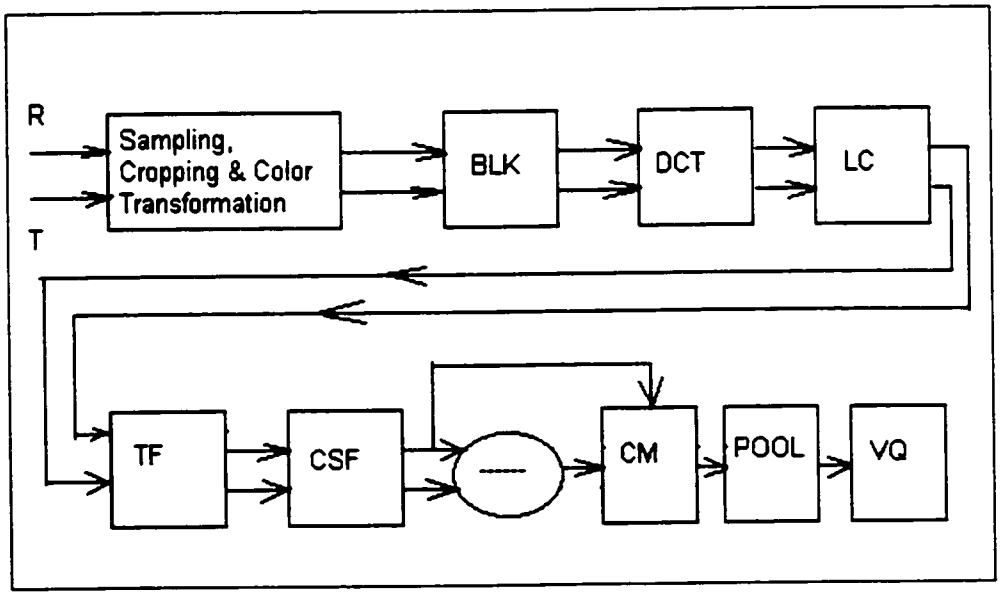


Fig 5.8 Overview of the DVQ Algorithm processing steps [Watson99]

The DVQ (Digital Video Quality) metric incorporates many aspects of human visual sensitivity in a simple image processing algorithm. One of the most complex and time consuming elements are the spatial filtering operations employed to implement the multiple, band-pass spatial filters that are characteristic of human vision.

This algorithm uses the Discrete Cosine Transform (DCT) for this decomposition into spatial channels. This provides a powerful advantage since efficient hardware and software are available for this transformation, and because in many applications the transform may have already been done as part of the compression process.

Watson [Watson99] describes the model as shown in fig 5.8 as follows; The input to the metric is a pair of color image sequences: reference (R), and test (T). The first step consists of various sampling, cropping, and color transformations that serve to restrict processing to a region of interest and to express the sequences in a perceptual color space. This stage also deals with de-interlacing and de-gamma-correcting the input video. The sequences are then subjected to a blocking (BLK) and a Discrete Cosine Transform (DCT), and the results are then transformed to local contrast (LC). Local contrast is the ratio of DCT amplitude to DC amplitude for the corresponding block.

After that, a temporal filtering operation (TF) is applied, which implements the temporal part of the contrast sensitivity function. This is accomplished through a suitable recursive discrete second order filter. The results are then converted to just-noticeable differences by dividing each DCT coefficient by its respective visual threshold. This implements the spatial part of the contrast sensitivity function (CSF).

After that, the two sequences are subtracted. The difference sequence is then subjected to a contrast masking operation (CM).

Finally the masked differences may be pooled in various ways to illustrate the perceptual error over various dimensions (POOL), and the pooled error may be converted to visual quality (VQ).

5.1.5 The Perceptual Distortion Metric (PDM), EPFL, Switzerland

This metric is based on a spatio-temporal model of the human visual system. It consists of 4 stages as shown in fig 5.9. The reference and processed or impaired sequence pass through those 4 stages. The first one converts the input to an opponent-color space. The second stage implements a spatio-temporal perceptual decomposition into separate visual channels of different temporal frequency, spatial frequency and orientation. The third stage models effects of pattern masking by simulating excitatory and inhibitory mechanisms according to a model of contrast gain control. The final stage processes the pooling and detection stage and calculates a distortion measure from the difference between the sensor outputs of the reference and the processed sequence [Wink99].

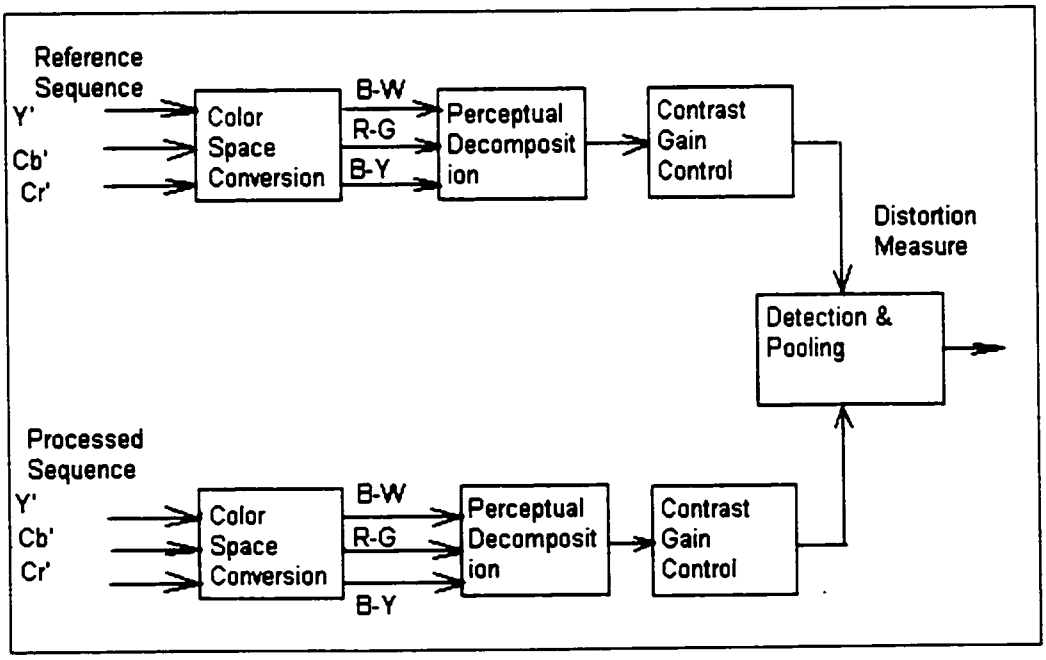


Fig 5.9 Block diagram of Perceptual Distortion Metric [Wink99]

5.1.6 Criticality Model, NHK, Mitsubishi, Japan

This model introduces a new concept called "Criticality". A definition of "criticality" is proposed as a quantitative measure of difficulty for MPEG-2 video coding, to analyze the picture quality of television programs statistically, depending on the complexity of the content, since picture quality in digital coding depends on picture characteristics as spatial detail and motion. [Nish97] Criticality is useful not only for estimating picture quality distribution characteristics but also for selecting a set of test sequences. By measuring frequency of occurrence of criticality and relating it to subjective picture quality, it becomes possible to obtain the statistical quality distribution of television programs in digital broadcasting.

The definition of criticality and the procedure to derive statistical picture quality distribution have been included in the latest revised ITU-R Recommendations BT.1210 (Test materials to be used in subjective assessment) and BT.1129 (Subjective assessment of standard definition digital television (SDTV) systems) [Nish00].

Criticality is defined as "the number of output bits per pixel from a hybrid DCT encoder with a fixed quantizer". Criticality is averaged over each whole frame. Figure 5.10 shows the configuration of the criticality measurement equipment based on the definition. This definition corresponds to the bit rate required to obtain an almost constant picture quality for various sequences, because a critical sequence requires a higher bit rate to maintain picture quality to the level of the non-critical sequences.

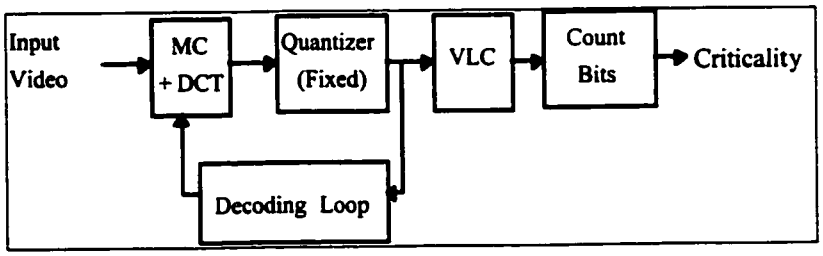


Fig 5.10 Configuration of criticality measurement [Nish00]

The procedure suggested by Nishida to evaluate picture quality distribution characteristics is shown in fig 5.11, and consists of 5 steps:

1- Measure criticality of test sequences used in subjective assessment.

2- Measure criticality distribution of broadcast programmes for a long time period.

3- Conduct a subjective assessment of picture quality of the system under test.

4- Derive a relationship between criticality and subjective picture quality for the test sequences.

5- Derive picture quality distribution characteristics (quality vs. frequency of occurrence) by combining the results of Step 4 (criticality vs. quality) and Step 2 (criticality vs. frequency of occurrence).

This method enables to evaluate the performance of coding systems in terms of frequency of occurrence of certain picture quality for overall broadcast programs. [Nish97, Nish00]

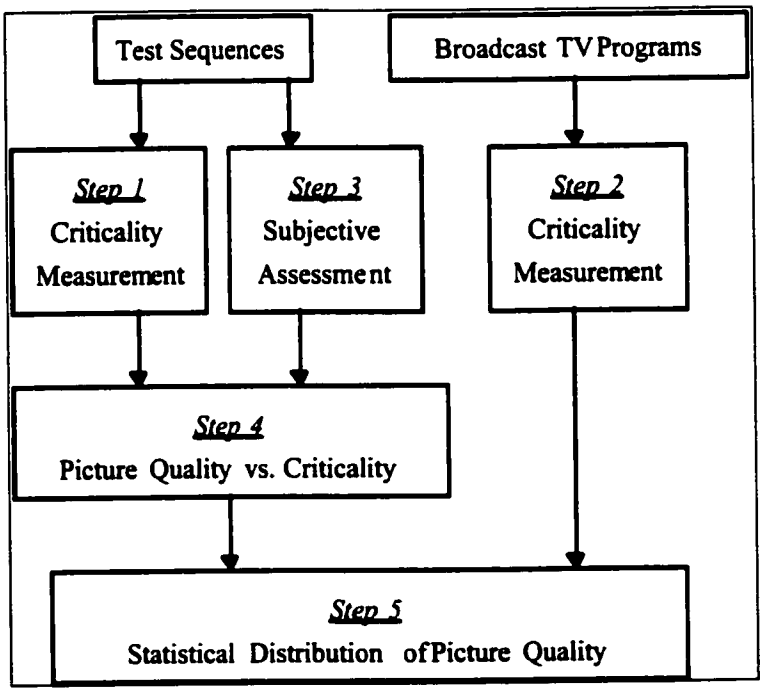


Fig 5.11 Procedure to derive picture quality distribution characteristics

5.1.7 KDD/Pixelmetrix Model, Japan

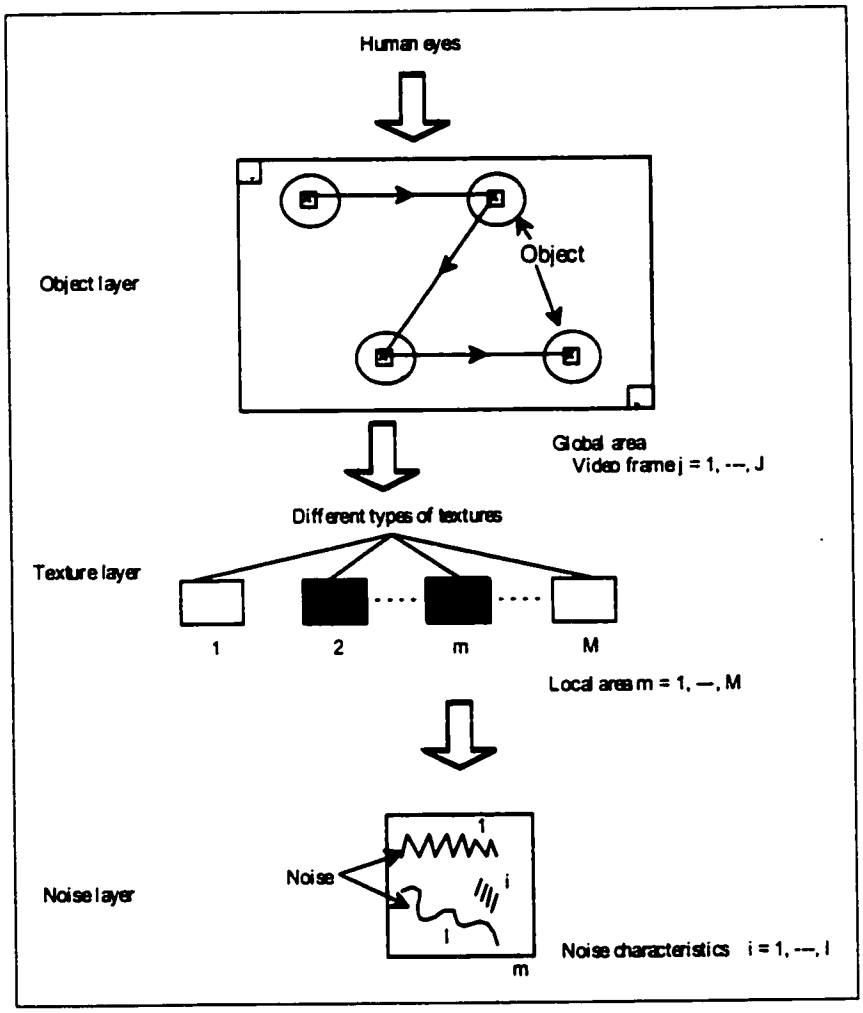


Fig 5.12 The 3 layered model for quality evaluation

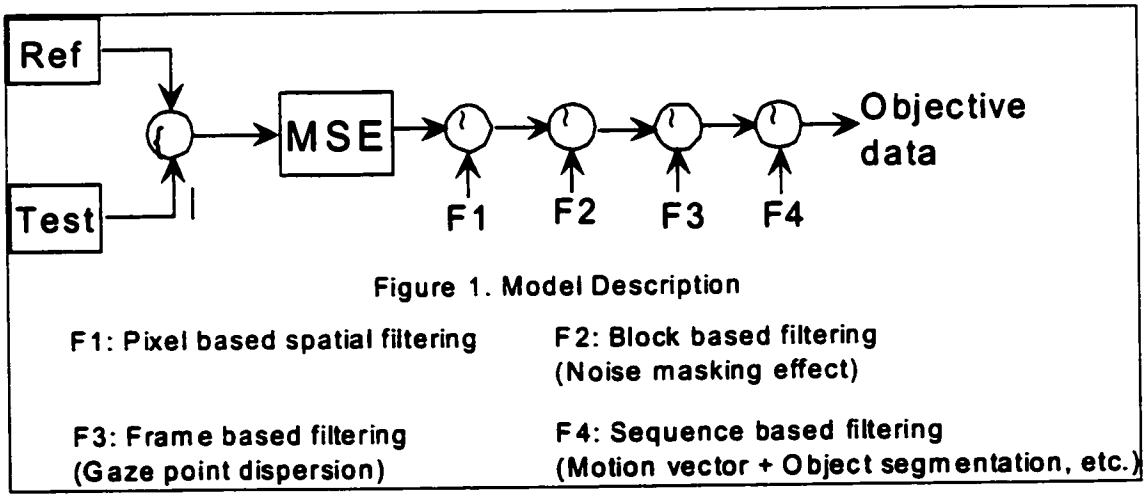


Fig 5.13a The KDD model

Figure 5.12, fig 5.13 shows the three-layered picture quality model as seen by the human eye. The human eye cannot watch a whole frame at a glance, but watch only a local spot area in a frame, which is around the gaze point of the human eye, and recognizes the texture and also quality of the area depending on the degree and characteristics of noise mixed in this texture [Ham00]. The whole frame is understood by moving the gaze point among objects for the whole frame at the same time. In this process, picture quality is determined by the noise over a frame. Therefore, to perform objective measurement of subjective picture quality, the macro to micro three-layered picture structures (object, texture and noise layers) are used, and a bottom-up noise weighting scheme is proposed which uses a particular weighting function at each layer taking into account human visual perception.

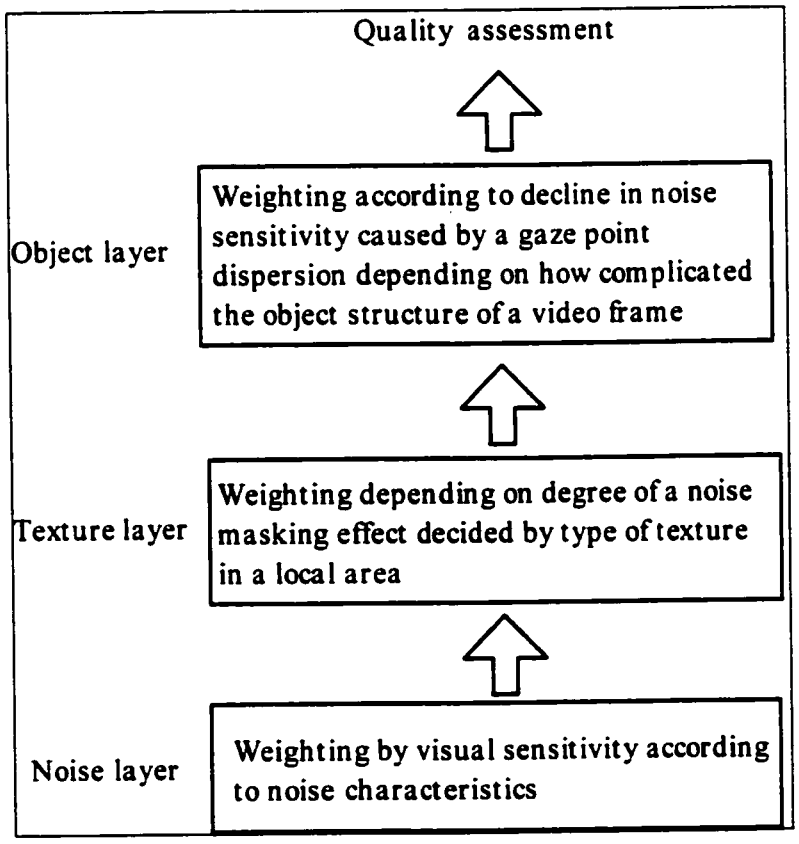


Fig 5.13b The 3 layer model of noise weighting [Ham00]

First, at the noise layer, common noise in a video compression process such as high frequency noise, low frequency noise, chroma noise, jerkiness, flicker ...etc are weighted depending on

their degrees and characteristics. After that, local spot areas are classified into several groups by their texture types. These groups include for example, forest, trees and a sports stadium in which noise are strongly masked, and "flat texture" such as a human skin and a sky in which noise are easily recognized. Thus, noises are weighted more or less according to their texture types.

Finally, at the object layer, the dispersion degree of the gaze point is predicted by measuring how complicated the structure is of objects in the video frame. Then, noises in the whole frame are weighted corresponding to a decline in noise sensitivity caused by this dispersion. [Ham00]

5.1.8 The KPN/Swisscom CT model, Netherlands

The Perceptual Video Quality Measure (PVQM) as developed by KPN/Swisscom CT uses the same approach in measuring video quality as the Perceptual Speech Quality Measure [Bee94] in measuring speech quality. The method was designed to cope with spatial, temporal distortions, and spatio-temporally localized distortions like found in error conditions. It uses ITU-R 601 input format video sequences (input and output) and re-samples them to 4:4:4, Y, Cb, Cr format.

A spatio-temporal-luminance alignment is included into the algorithm. Because global changes in the brightness and contrast only have a limited impact on the subjectively perceived quality, PVQM uses a special brightness/contrast adaptation of the distorted video sequence. The spatio-temporal alignment procedure is carried out by a kind of block matching procedure. The spatial luminance analysis part is based on edge detection of the Y signal, while the temporal part is based on difference frames analysis of the Y signal.

Since the Human Visual System (HVS) is much more sensitive to the sharpness of the luminance component than that of the chrominance components. As well, the HVS has a contrast sensitivity function (CSF) that decreases at high spatial frequencies. These basics of the HVS are reflected in the model. The system has 4 steps:

1- The PVQM algorithm provides a first order approximation to the contrast sensitivity functions (CSF) of the luminance and chrominance signals.
2- The edginess of the luminance Y is computed as a signal representation that contains the most important aspects of the picture. This edginess is computed by calculating the local gradient of the luminance signal (using a Sobel like spatial filtering) in each frame and then averaging this edginess over space and time.
3- The chrominance error is computed as a weighted average over the color error of both the Cb and Cr components with a dominance of the Cr component.
4- In the last step the three different indicators are mapped onto a single quality indicator, using a simple multiple linear regression, which correlates well the subjectively perceived overall video quality of the sequence. [VQEG00, Bee97]

5.1.9 Tapestries Model, UK

This model used a different approach by designing separate modules specifically tuned to certain type of distortions, the select one of the results reported by these modules as the final objective quality score.

It consists of a perceptual model and a feature extractor. The perceptual model simulates the human visual system, weighting the impairments according to their visibility. It involves contrast computation, spatial filtering, orientation-dependent weighting, and cortical processing. The feature extractor is tuned to blocking artifacts, and extracts this feature from the video. The perceptual model and the feature extractor each produce a score rating the overall quality of the HRC video.

Since the objective scores from the two modules are on different dynamic range, a linear translation process follows to transform these two results onto a common scale. One of these transformed results is then selected as the final objective score, and the decision is made based on the result from the feature extractor. [VQEG00]

5.2 No Reference Models

5.2.1 The IFN/R&S Algorithm, Germany

This model which was developed by the Institut für Nachrichtentechnik (IFN), Braunschweig Technical University and Rohde & Schwarz, Germany, is a single ended system that evaluates the quality of degraded video and measuring DCT artifact (i.e. blocking). The typical application of such an algorithm is in the on-line monitoring of MPEG2 video quality.

Woerner and Lauterjung [Woe01] state that during a research at the IFN by Prof. Ulrich Reimers; "It became apparent that the visual impression of the MPEG-2 inherent blocking structure has the greatest impact on the picture quality in an otherwise normal video stream". Other kinds of artifacts like edge blurriness and mosquito noise that are also visible have been found to be of lower importance.

The model consists of four main processing steps. The first one is the detection of the coding grid used. In the second step based on the given information the basic parameter of the method is calculated. In the third step, the result is weighted by some factors that take into account the masking effects of the video content like spatial activity and temporal activity. Because of the fact that the model is intended for monitoring the quality of MPEG-coding, the basic version produces two quality samples per second, as the Single Stimulus Continuous Quality Evaluation method (SSCQE, ITU-R BT rec. 500) does. The submitted version produces a single measure for the assessed sequence in order to predict the single subjective score of the DSCQS test used in this validation process. [VQEG00]

The theory of the algorithm is based on the values associated with the 8x8 blocks and 16x16 macroblocks of the DCT transform. The digital video quality level is calculated from vectors, which contain information about the averaged difference between adjacent pixels. Comparing all the pairs in a block as shown in fig 5.14, the encoding process usually reduces the difference between adjacent pixels. The exception of that are the pixels on the edges of the blocks or macroblocks.

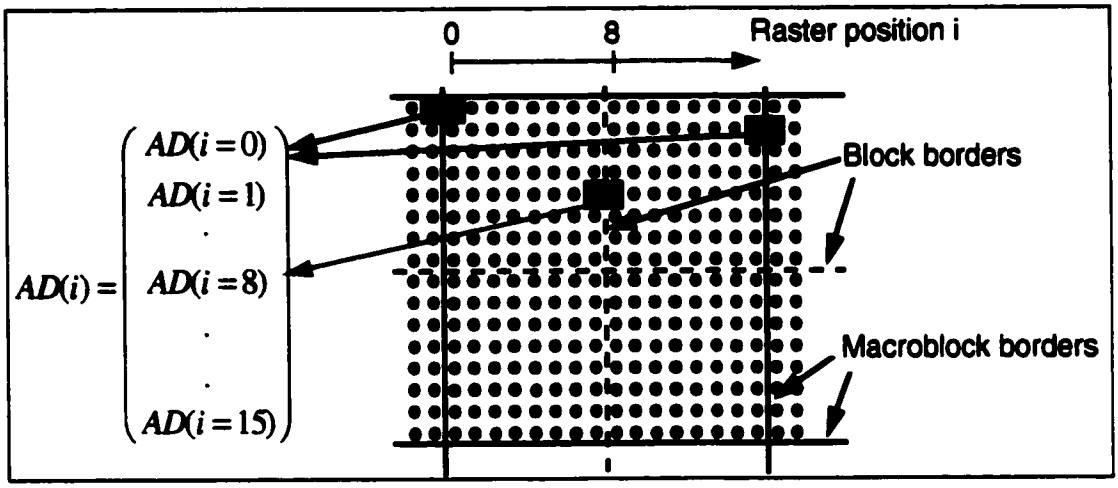


Fig 5.14 Sample calculation process for the amplitude differences of adjacent pixels [Lau98]

There are 16 vector elements built, which represent the average amplitude difference of all pixel pairs with the same relative distance within the macroblock grid. Fig 5.15 shows the result of the results of the calculation of the average differences on the original uncompressed stream of "Flowergarden". The values of all the pixel pairs are very close. The overall average represents the spatial activity of each frame.

After encoding the same sequence at a rate of 2 Mbits/sec then decoding it, the result is shown in fig 5.16. In this case, it's obvious that there are 2 values for the elements AD0, AD8 that are different from the rest of the block values. These values are at the edges of the block and macroblock. This affects the visibility of the blocking effect. In case the video was encoded at a higher rate for example, the difference between those 2 values and the rest of the block values will become less, hence reducing the blocking effect.

The spatial activity is calculated by averaging all pixel amplitudes differences regardless of their position within the frame. The temporal activity is calculated by averaging the amplitude difference of the same pixels in subsequent frames. In version 2.0 of the algorithm, this has been updated to use regions of pixels in subsequent frames. This made more accurate temporal

activity measurement, especially for slow moving camera or slow moving background scenes. For those scenes the previous system gave higher values for temporal activity.

There is masking applied to the values afterwards to come up with the objective weighted video quality measure that corresponds to the subjective evaluation.

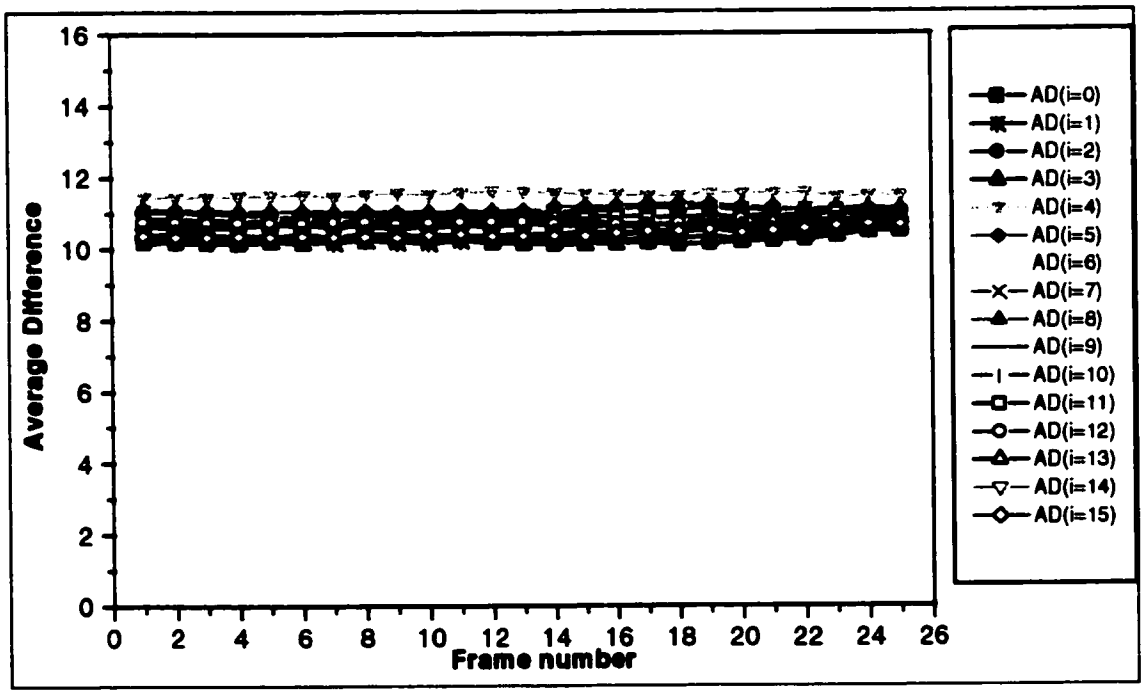


Fig 5.15 Averaged pixel amplitude differences, (Flowergarden stream uncompressed)

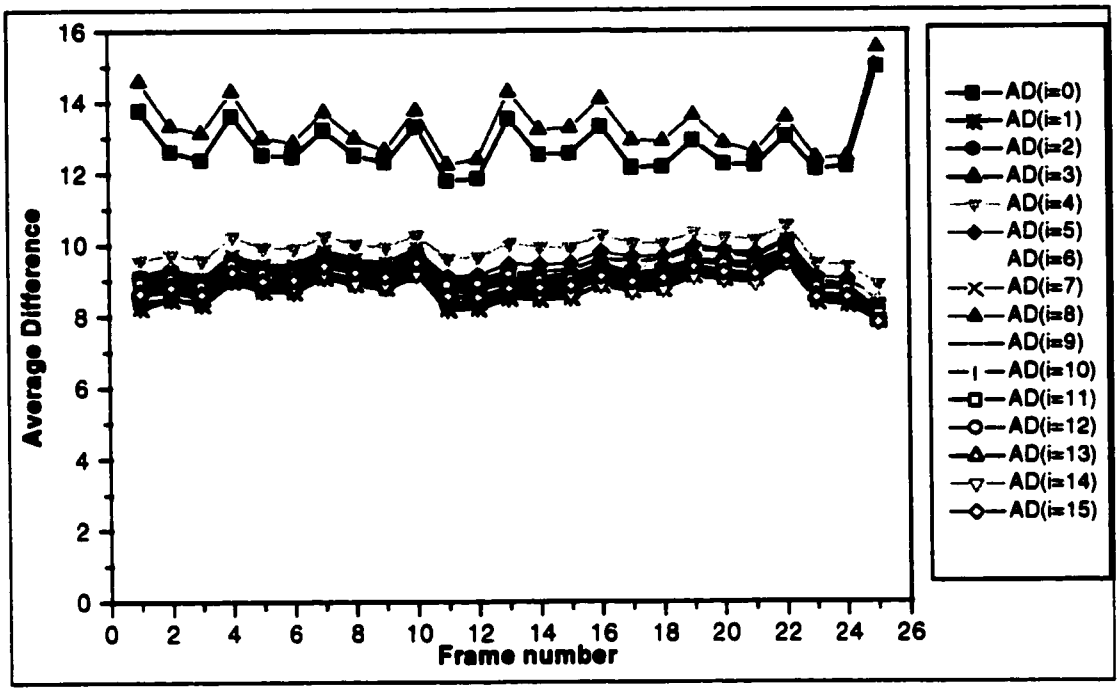


Fig 5.16 Averaged pixel amplitude difference (Flowergarden Sequence coded/decoded at 2 Mbits/sec)

The major benefit of this algorithm is that it doesn't need any reference signal to operate, and thus can be used in on-line monitoring of video networks, besides it can be put in any point in the network.

Chapter 6

Calibration of Objective to Subjective Values

6.1 The VQEG Work

One of the most exhaustive efforts done in the evaluation of digital video quality has been performed by the Video Quality Expert Group (VQEG). VQEG was formed in Oct. 1997 (CSELT, Turin, Italy) to make a framework for the evaluation of new objective methods for video quality. This would ultimately help the different ITU study groups develop new recommendation on the subject. A framework test plan was made in 1998 at the National Institute of Standards and technology, Gaithersburg, USA. The evaluation was made for 9 double-ended systems (mentioned earlier), and one single ended system (The IFN/R&S) system. They added one more metric which was the PSNR as a proponent.

The subjective testing was made according to the ITU-R BT.500 standard. Sessions where made into 20 minutes, with 2 sessions per day. Each stream was 8 seconds long with a 2 second Grey level in between different sources.

They followed the Double Stimulus Continuous Quality Scale (DSCQS). So, The viewer was shown each sequence then he would mark on a sheet on a scale 1-5 or 0 -100 the mark he felt corresponded to the quality value. The sequence of scenes always composed of a pair of the same scene. One of the streams in this pair would be the reference or clean signal, and the other would be the impaired signal. The viewer was not told which of the pairs was the clean and which was the impaired. This procedure is unlike the DSIS procedure, where the first scene of the pair was always the reference signal, and the second was the impaired one.

The score of each group was averaged for each sequence after excluding readings that were twice the deviation level from the mean value.

The Mean Opinion Score were calculated and DMOS or Differential Mean Opinion Score between reference and processed video quality values.

The VQEG divided their test material into 4 groups:

1- 50Hz or 625 line streams

2- 60Hz or 525 line streams

3- Low quality streams (768KB/sec – 4.5 MB/s)

4- High quality streams (3 MB/s – 50 MB/s)

This covered the different categories they considered. They used 20 test sequences (10- 50Hz 625 line, 10- 60Hz 525 line streams. The picture formats were different from MPEG2 with different profiles (MP@ML, SP@ML, 422p@ML and H263. They used 16 Hypothetical Reference Circuit) HRC to insert artifacts into each reference sequence.

6.1.1 Objective Model Evaluation Criteria:

The VQEG estimates a number of attributes are responsible for characterizing the performance of a video quality model. These are:

1- Prediction Accuracy

2- Prediction Monotonicity

3- Prediction Consistency

The outputs by the objective video quality rating model (the VQR's) are correlated with the viewer DMOS's (Differential Mean Opinion Score) in a predictable and repeatable fashion. The VQEG group in their objective test plan explored the possibility of the non-linearity in the relationship between predicted VQR and DMOS, as subjective testing can have nonlinear quality rating. However, later on it was accepted by VQEG and others like Woerner and Lauterjung [Woe01] that linear regression is acceptable as index value of the correlation.

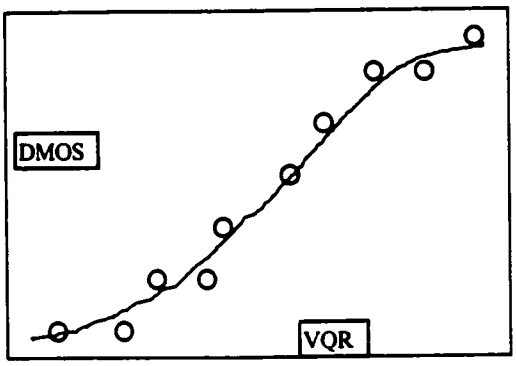


Fig 6.1 Relation between subjective and objective Measurement [VQEG-Ob98]

Fig 6.1 shows the relation graph between a subjective measured DMOS value and the predicted DMOS of the objective model. The scale could be 0-100 for both. This curve represents the correlation value for one specific objective model, and the points on the curve are the different video sequences used in both subjective and objective tests.

The nonlinear regression can be fitted to the [subjective/objective] data set. The 2nd, 3rd and 4th order polynomial nonlinear regression are discussed, besides the linear regression. Where the set of differences between measured and predicted DMOS is defined as the quality-error set Qerror[ ]:

$$Qerror[i] = DMOS[i] - DMOS_p[i]$$

The DMOS[I] is the measured differential mean opinion score for stream "I", and DMOSp[I] is the predicted DMOS by a specific objective model for the same stream i.

There were a number of metrics chosen as a measure for the above mentioned attributes [VQEG-Ob98].

5.2 Evaluation Metrics:

5.2.1 Metrics relating to Prediction Accuracy of a model

This attribute measures the ability of the model to predict the viewers' DMOS ratings with a minimum error "on average". [VQEG-Ob98]. There are a few metrics that can be used to measure the average error, with root-mean-square (RMS) error being a common one. To

include the known variance in subjective DMOS data, the simple RMS error can also be weighted by the confidence intervals for the mean DMOS data points. ITU500 describes a procedure to get 95% confidence level in the subjective measurements [ITU500]. The pearson linear correlation coefficient is a common metric that is related to the average error because lower average errors lead to higher values of the correlation coefficient, and it's a very simple relation to realize by mathematical or spreadsheet programs.

Metric1: The Pearson linear correlation coefficient between DOS_p and DOS, including a test of significance of the difference. Where DOS is differential opinion score for subjective measurement, and DOSp is for predicted objective measurement.

Metric2: The Pearson linear correlation coefficient between $DMOS_p$ and DMOS.

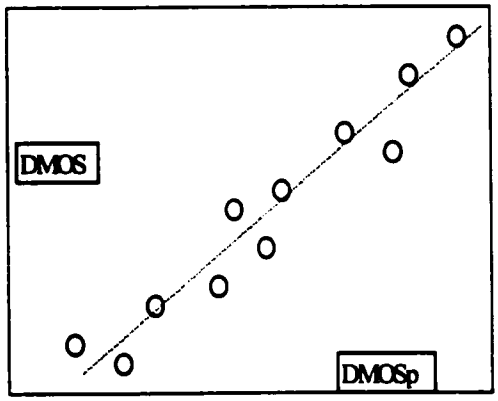


Fig 6.2 Model with greater accuracy [VQEG-Ob98]

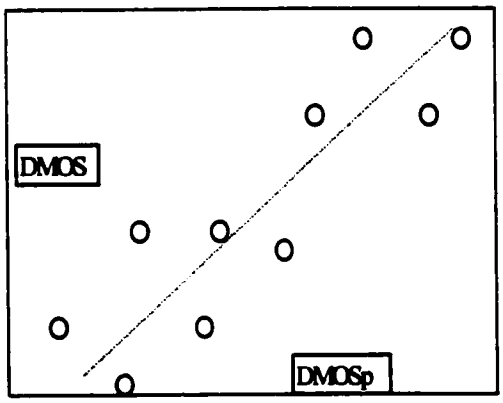


Fig 6.3 Model with lower accuracy [VQEG-ob98]

6.2.2 Metrics relating to Prediction Monotonicity of a model

An objective model's $DMOS_p$ values should ideally be completely monotonic in their relationship to the matching DMOS values [VQEG-Ob98]. The model should predict a change in $DMOS_p$ that has the same sign as the change in DMOS. Figures 6.4 and 6.5 show the hypothetical relationships between $DMOS_p$ and DMOS for two models of varying monotonicity. Both relationships have approximately the same prediction accuracy in terms of RMS error, but the model of Figure 6.4 has predictions that monotonically increase. The model in Figure 6.5 is less monotonic and falsely predicts a decrease in $DMOS_p$ for a case in which viewers actually see an increase in DMOS.

The Spearman rank-order correlation between $DMOS_p$ and DMOS is a sensitive measure of Monotonicity. [VQEG-Ob98].

Metric3: Spearman rank order correlation coefficient between $DMOS_p$ and DMOS.

A pair-wise comparison of pairs of HRC's on a scene by scene basis has also been proposed for examining the correlation between subjective preferences and objective preferences, and merits further investigation by the VQEG for inclusion in these tests.

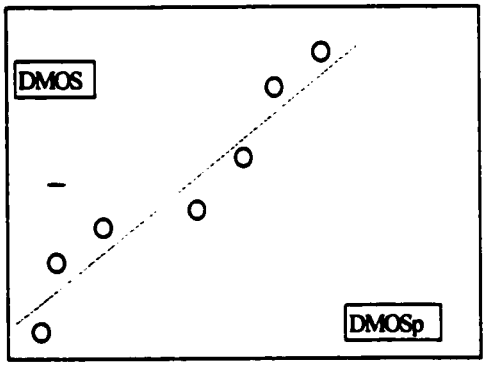


Fig 6.4 Model with more Monotonicity [VQEG-Ob98]

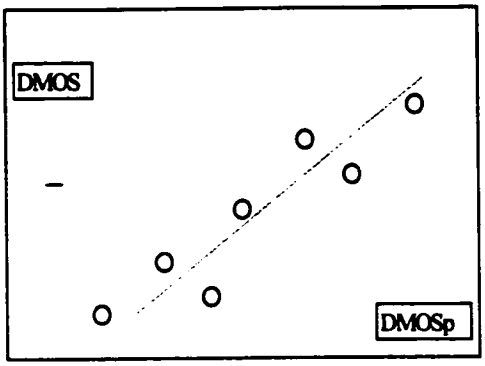


Fig 6.5 Model with less Monotonicity [VQEG-Ob98]

6.2.3 Metrics relating to Prediction Consistency of a model

This attribute means that the objective quality model's ability to provide consistently accurate predictions for all types of video sequences and not fail for a subset of sequences.

Figures 6.6 and 6.7 show models with approximately equal RMS errors between predicted and measured DMOS, but in Figure 6.6 it has accurate predictions for the majority of sequences but has large prediction error for the two points in the middle of the figure. In the second Figure 6.7 this model has a balanced set of prediction errors, although it is not as accurate as the first model, but it performs "consistently" by providing reasonable predictions for all the sequences. The model's prediction consistency can be measured by the number of outlier points, like having an error greater than a given threshold such as one confidence interval, [ITU500]. A smaller outlier fraction means the model's predictions are more consistent. Another metric that relates to consistency is Kurtosis [VQEG-ob98], which is a dimensionless quantity that relates only to the shape of the error distribution and not to the distribution's width. Two models may have identical RMS error, but the model with an error distribution having larger "tails" to the distribution will have a greater Kurtosis

Metric4: Outlier Ratio of "outlier-points" to total points N.

$$OutlierRatio = \frac{(Total\ number\ of\ Outliers)}{N}$$

where an outlier is a point for which:

$$Abs[Qerror[i]] \rangle 2 * DMOSS\tan dardError[i]$$

Twice the DMOS Standard Error is used as the threshold for defining an outlier point.

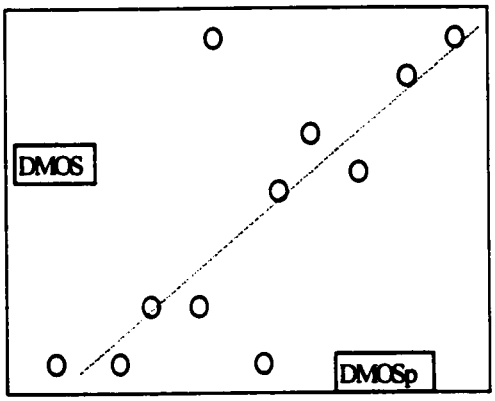


Fig 6.6 Model with large outlying errors [VQEG-Ob98]

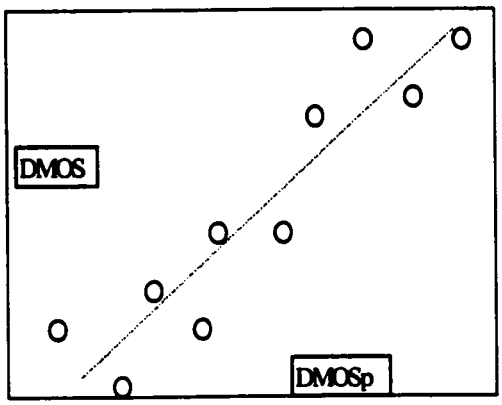


Fig 6.7 Model with consistent errors [VQEG-Ob98]

6.3 The VQEG Test and Conclusions:

The VQEG initially adopted these metrics and another one for the Analysis of variance ANOVA [VQEG00]. In the testing phase, they used software models on workstations like Sparc, silicon graphics and the streams were entered as a luminance and chrominance ITU601 component

stream on the algorithms, after being normalized and aligned in a special way [VQEG00, VQEG-ob98].

There were 2 other double-ended algorithms that reported poor results due to some technical problems. However, it's probable that these problems have been remedied afterwards with new releases of the algorithms.

The most important finding of the report was that there was "No conclusive Result"! The aim of VQEG was to get the best system performance. However, most of the algorithms gave very close results and correlation values (around 70s-80s) with no one to be a clear best. Another finding was that the different quadrants they used (525 line, 625 line, low and high quality sequences), results gave a wide range of values for the same model. Thus there was low consistency of results.

There was another observation in one of the correlation tables in the VQEG report [VQEG00 pp106] that for some models, the correlation for any 2 quadrants for a model gave better results when the values were combined to get the (all) data sets results. For example if we see the 50Hz (625 line) sequences correlation values and 60 Hz (525 line streams) correlation values, the correlation value for all the sequences should give a value somewhere in between the 2. However in a few models, it showed that it gave "Better" correlation. Thus the correlation value was sometimes affected by the number of points in the curve. And with more points they got better correlation rather than the more accurate one.

The other very interesting finding that the simplest model added, which was the PSNR reported remarkably good results overall (around 80%).

6.4 The Suggested In-Field Calibration standard

The excellent work done by groups like the VQEG has created a framework for development of such solutions and answers for broadcasters. And one solutions possible is to suggest a form of In-Field calibration performed by the TV networks or organizations involved in the transmission

and distribution of digital MPEG2 video to adopt. The [ITU-500] standard has suggested 2 levels of subjective tests to be done. One for the laboratory level and a lower less stringent level for home viewer level. The second one can be adopted by this standard to stand for the quality level expected in the customer's environment as their benchmark.

The most critical part in this new standard is the use of test sequences for the evaluation. At this age of hundreds of TV stations, the concept of specialized networks is becoming more dominant. Thus the material varies considerably between the following classes of networks:

1- Cartoon Networks

2- Drama Networks

3- News Networks

4- Sports Networks

5- Talk Show Networks ...etc.

The type of audience for these types of programming maybe different. Their judgment differs based on age group, psychological, mood and other factors that are not measured in a very straightforward manner.

The other but more relevant issue is that the complexity of spatial, temporal activity, luminance and chrominance complexity of these programs. The data rates for these kinds of programs are a critical factor for the quality.

One difference demonstrated by the VQEG report that the 50Hz vs. 60Hz programs may give different results, so this is another factor to be considered.

The other factor that was discounted as a variable, thus making the modeling a bit easier, was the type of MPEG2 encoders used to perform the compression of programs.

As mentioned before, video quality objective techniques are used for some applications including as one to evaluate the differences in CODECs and encoders performance. ITU-R BT.813 recommendation "Methods for Objective Picture Quality Assessment in Relation to Impairments from Digital Coding of Television Signals" [ITU-813] demonstrates the differences

in encoder types and testing methodologies. Thus this is another factor for TV networks to consider. Different networks will have different encoders, which yield different quality of video. As well, with time, those networks will change those encoders over time to newer technology encoders that will be more efficient in addressing critical areas within the video stream to give better data bandwidth in those areas, without exceeding the constraint of bit rate allocated. This would be in either constant bandwidth or statistical multiplexing [Kuhn00] systems.

Coombs [Coom72] states that: "It's natural that an echelon or hierarchy of standards will evolve if a measurement system is to show lineage or traceability to a common source". There will be standards for higher end reference entities like laboratory and R&D, manufacturers for example. And there will be lower working standards for operational units. In our case would be TV stations and networks.

Therefore a suggested calibration standard or validation for networks would include:

1- Subjective Testing of streams according to the condition of the Network or station (Data rates, test material, encoders)
2- Objective testing of the same streams by the model they have. The IFN/R&S in this case.
3- Adapting the results to create the correlation relationship.

This is shown in fig. 6.8. And if there was another double stimulus system used, the same is shown in fig. 6.9.

6.4.1. Subjective Measurement of quality for video sequences.

The choice of material should be a simulation of the regular programming material that this Network broadcasts. For example if it's a sports network, emphasis should be made on the different artifacts caused in the high spatial, temporal, luminance and chrominance fields in different sports, when constructing the HRCs and video sequences used.

The test should follow one of the standard listed in the ITU-R BT.500 (ITU-500 version 10 at the moment issued on March/2000), when performing this.

Following the second standard that addresses viewer rather than the lab test, may address the customers of network better.

For the continuous monitoring of video quality, if a single stimulus standard were followed, the duration of the sequences would be a critical issue. This should be chosen somewhere between 5 minutes and 10 seconds, which are the times for double and single ended tests respectively [ITU-500], if this system would be used for both single and double ended system.

The reading of values may need to be changed accordingly. At this time DSCQS, the measurement is taken once for the whole stream of 8 or 10 seconds, where as for single ended systems, the reading is taken from the evaluation unit twice per second. An optimum time is probably around 40 sec.- 1 minute. This way it would be long enough to do the single stimulus system and at the same time can offset any time lost in the beginning of the sequence when the system starts the measurement. It will have adequate results for the double stimulus systems as well.

A Network can perform these subjective tests at a number of facilities or independent labs available. There are many of those in each country or at least continent. This is probably the most expensive part of the test. As subjective tests take some days, and need quite a lot of resources and cost. However this will not be done on a regular basis. One set of such measurements may be used for years.

Of course when using subjective test sequences with HRCs representing the different artifacts and issues to be addresses, it should played from a digital machine that doesn't have compression inside like a D1, D2, D3 or D5 tape machines. The other solution is to use a digital disk recorder that doesn't have compression. The reason for that is that we shouldn't let the test itself affect the quality measurements.

6.4.2- Objective testing:

The same streams used could be run at the objective quality evaluation system under evaluation. This needs to be very simple by just running the streams from the tape or disk machine into the unit under test, and acquiring the results. This will be similar to the usual way they would get their quality monitoring results. Representing the correlation curve afterwards using one of the standard metrics would give the correlation for that system. A very simple correlation relation that can be used is the Pearson linear regression technique. Needless to say, the same kind of tape or disk machine should be used for this test as well to standardize the same conditions for both subjective and objective tests.

The process of calibration is shown in the following graphs. For single stimulus systems, which would be used for the continuous monitoring of quality, this is shown in fig. 6.8. Fig 6.9 shows the same for double-ended systems.

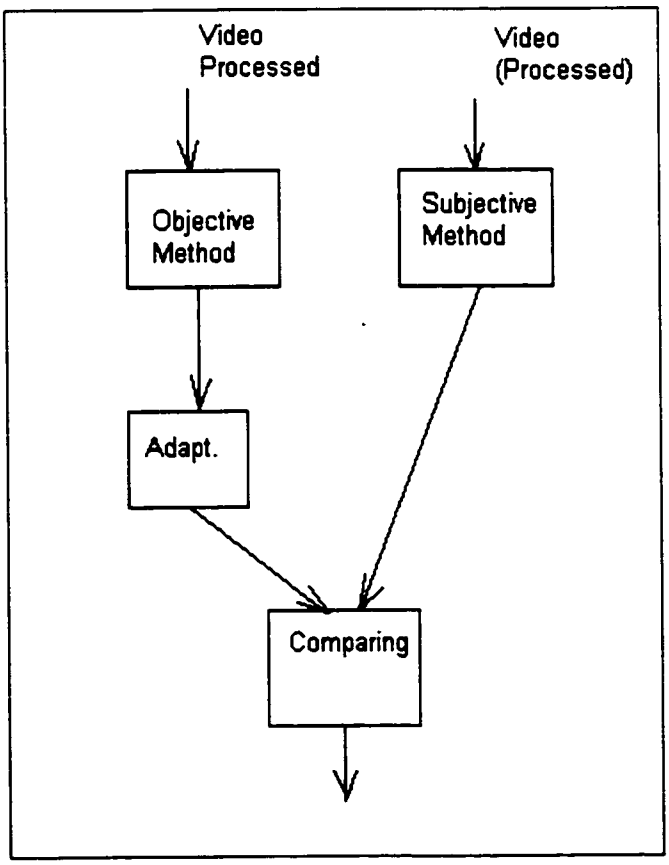


Fig 6.8 Single ended systems evaluation

This kind of procedure would be useful for documenting quality and what quality values obtained really mean in relation to subjective values. This knowledge could be helpful in a number of ways.

1- In case there is an update on the video quality system algorithm itself, and the network would run those objective tests on the updated system, and hopefully get a better correlation result.

2- In case the network have a number of quality evaluation systems, and need to get a common reference of the different quality values and scales used by different algorithms.

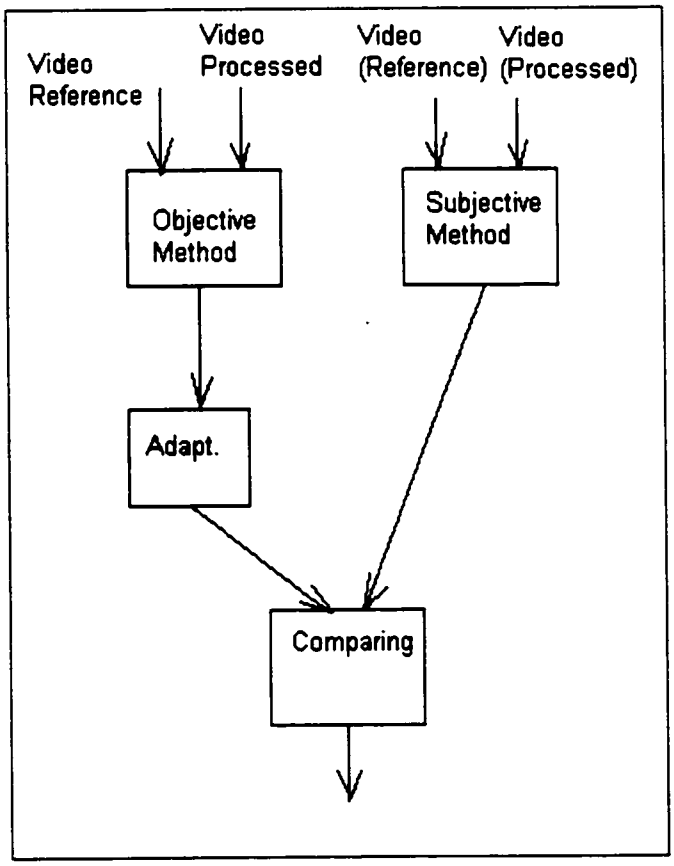


Fig 6.9 Double Ended systems evaluation

3- It can be expanded as well to include both single stimulus and double stimulus systems. In case the network has double stimulus system to do out of service test of the transmission links for example, and another single stimulus system to do the continuous online monitoring of quality.

4- In case there is a change in the condition of the network. That means if the network changes the type of encoders/decoders (CODECs) used for example. There maybe a need to re-design the test again to get the proper characterization of the network. However, in this case, there will be a need to re-create test sequences and artifacts simulating the new network.

5- To better characterize the performance of the network and closely follow similar standards put for analog systems performance. Such performance is usually important and could become included in such quality standards like ISO-9000.

6.5 Practical Experiment

As a demonstration of this simple technique, it would have been extremely expensive to do all the subjective tests to demonstrate this technique. Thus, used the already available subjective results from the VQEG report. I used a subset of the VQEG previous sequences with the permission and help of Communication Research Center (CRC-Canada) to get the results for the objective part of the experiment. CRC was one of the labs assigned with subjective testing during the VQEG tests.

6.5.1 Subjective Evaluation

I used the 60 Hz (525 lines) sequences that were encoded in MPEG2, and with digital type artifacts, and used the already available subjective measurement of those streams [VQEG00]. Thus this would simulate a TV station or network in North America (525 line TV system) with such material and set of artifacts represented in HRCs. The list of video sequences used and the corresponding encoding is shown in table 6.1. As well the set of HRCs used are listed in table 6.2.

Assigned number	Sequence	Characteristics	Source
13	Baloon-pops	Film, saturated color, movement	CCETT
14	NewYork 2	Masking effect, movement)	AT&T/CSELT
15	Mobile&Calendar	Available in both formats, color, movement	CCETT
16	Betes_pas_betes	Color, synthetic, movement, scene cut	CRC/CBC
17	Le_point	Color, transparency, movement in all the directions	CRC/CBC
18	Autumn_leaves	Color, landscape, zooming, water fall movement	CRC/CBC
19	Football	Color, movement	CRC/CBC
20	Sailboat	Almost still	EBU
21	Susie	Skin color	EBU
22	Tempete	Color, movement	EBU

Table 6.1. 525/60 format sequences

The first frame for all video sequences is enclosed in Appendix. A. They consisted a complete range of a typical TV station material. Each stream represented a type of material with certain critical factors to both DCT transform, and transmission. These characteristics included: color, movement, sharp building edges, waterfalls and water movement, skin color, animation and graphics.

ASSIGNED NUMBER	BIT RATE	Resolution	METHOD	COMMENTS
14	2 Mb/s	¾	mp@ml	This is horizontal resolution reduction only
13	2 Mb/s	¾	sp@ml	
12	4.5 Mb/s		mp@ml	With errors TBD
11	3 Mb/s		mp@ml	With errors TBD
10	4.5 Mb/s		mp@ml	
9	3 Mb/s		mp@ml	
5	8 & 4.5 Mb/s		mp@ml	Two codecs concatenated
2	19-19-12 Mb/s		422p@ml	3rd generation

Table 6.2. Test conditions (HRCs)

I have used 10 video streams of 60 Hz (525 line streams), with 8 Hypothetical Reference Circuit (HRCs), which were a subset of the VQEG streams. They were a mix of high quality and low quality from the VQEG definitions. All were MPEG2 compressed, but had different profiles and resolution. Their bit rates ranged from 2 Mbits/sec., and 19 Mbits/sec. As practical use for standard definition streams, this is a reasonable range for a full range of conditions for a TV station. Typical rates for broadcast are between 2 and 6 Mbits/sec. The combination of video stream and HRCs created about 80 streams, giving a decent impression of a station programming variety.

6.6. The test setup

I have used a very simple setup including:

1- A D-1 tape machine to playback the video streams. This is the same tape machine used for the subjective evaluation tests at the CRC.

2- A Sony 19" Monitor connected to the tape machine.

3- ITU-601 interface. This is the 270 Mbits/sec. interface for uncompressed video. It's common interface to most new tape machines with digital outputs, and many hard disk servers. The IFN/R&S system can accept either an ITU-601 uncompressed video or a compressed video in the ASI (Asynchronous Serial Interface) or SPI (Serial Parallel Interface) formats.

4- The IFN/R&S video quality unit (Single Stimulus Model).

5- A Computer Laptop with video quality software for the R&S/IFN system (called Quality Explorer Software).

6- RS-232 interface between the R&S/IFN system and the Laptop. This interface is used to acquire the results recorded from the unit in word and excel formats. The setup is shown in fig 6.10.

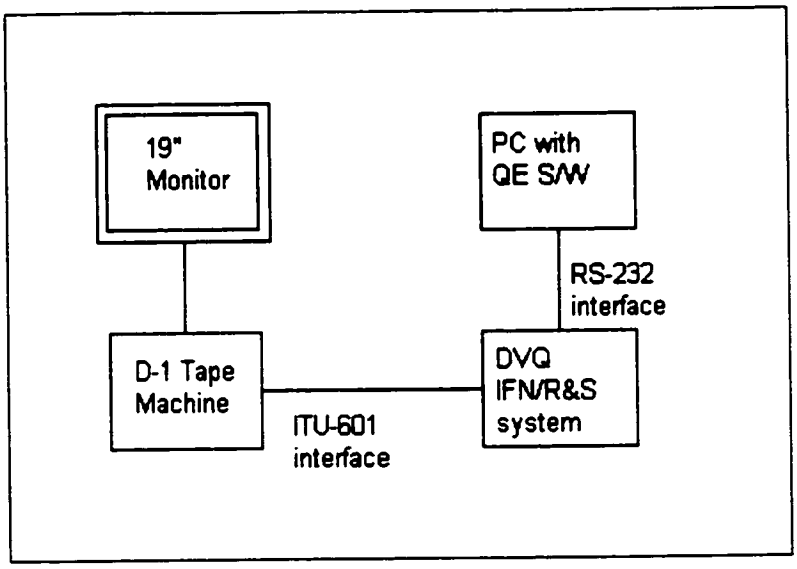


Fig 6.10 Test Setup for the Objective Measurement.

The test procedure was simple:

1- Start the D1 tape from the first frame of each test sequence of the Sequence/HRC combination (80 sequences).

2- Press the start button of the DVQ. This started the algorithm from the first step, which is calculating the picture resolution. After this step, which takes a few seconds, the algorithm is able to calculate the quality of the video sequence on a scale from 0-100.

3- Start the Quality Explorer (R&S) software on the computer to start acquiring the quality results. This is basically a curve of all measured values (twice per second) or once every 400 ms. As well, an excel sheet of all the measured values.

4- Obtain the values from the excel sheets, and compare them with the acquired graph, and obtain the average value per sequence. The results are shown in the Excel sheet of the results. There is one column representing the mentioned objective values. Another one for subjective values as per [VQEG00].

There was some alignment done at the end after obtaining the results. The reason was, as the stream ended, the tape would go to Grey. When that happened, the temporal activity went to 100%. Therefore, in some streams, I deleted values of the last one or 2

frames were the transition occurred for a 100% temporal change. So, the quality values were as best as possible corresponding to the measured video stream.

6.7 The test results and analysis

The results are shown in the enclosed figures. The Pearson linear regression was performed between subjective and objective values. The correlation obtained using linear regression was 76.54%. This correlation can be obtained very easily mathematically. I've tried also using polynomial regression to see the changes of correlation in these cases as per [VQEG-ob98]. The correlation changes a bit to 76.7% for a second order polynomial, 77.14% for a third order polynomial, and 77.48% for a 4^{th} order polynomial.

However, the linear regression is widely acceptable [VQEG00]. The result is reasonable, and correlates to the values obtained by Woerner and Lauterjung [Woe01] of 82.79%.

Furthermore, it's not exactly comparing apples to apples when considering the VQEG results, as they included analog & H-263 streams. HRCs included artifacts for analog artifacts, and others. However, looking at results in a more general way, correlations for those streams and artifacts ranged roughly from 70s to 80s.

Thus for MPEG2 streams and a sample of algorithms used in a typical TV station, this algorithm performs well, taking in consideration it's a single stimulus system.

There were other factors affecting the results, that should be considered in other Networks designing their test procedures, and eventually including in a standard:

1- The video streams were only 8 seconds long. Since this is a single stimulus system, it took a few seconds to calculate the resolution, and start calculating the quality. Therefore, the average number of values measured per stream were around 9 or 10 values. Almost the first half of the stream was not measured. Therefore, for future testing specifically designed for single stimulus systems, the streams should be longer than

that, as the first 3-4 seconds will be discarded. If the length of the stream is much higher, that will give a better confidence in the measurement values.

2- This testing was done only on a subset of the test sequences i.e. 60 Hz streams. The discarding of 50Hz streams may have changed the overall results a bit. However, in ordinary situations, a network would only be broadcasting one of those standards.

3- The test was performed in a manual fashion by 2 people. One person started the D1 tape machine. The other started the DVQ unit and software. The delay of start of stream may induce a small error factor as well. Again this error can be reduced when the streams are longer, and thus these milliseconds in the beginning will not affect the result. The other solution is in new software development, when one start from the software would initialize the unit, and stop on the condition of temporal activity reaching 100%.

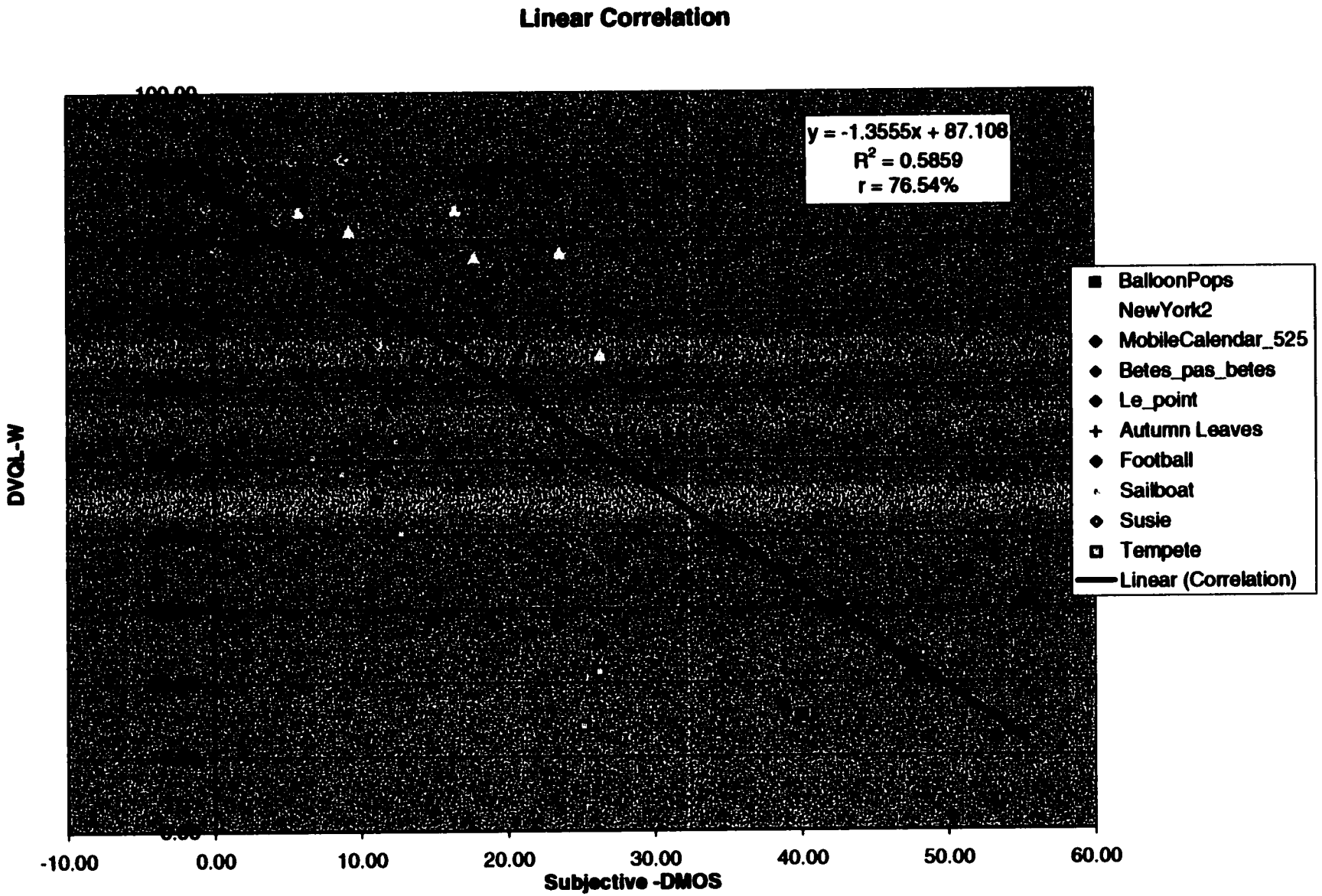
4- I had to disregard some values because the unit took almost the whole 8 seconds to calculate the resolution, so the measure of the last second or so didn't represent the whole stream quality value.

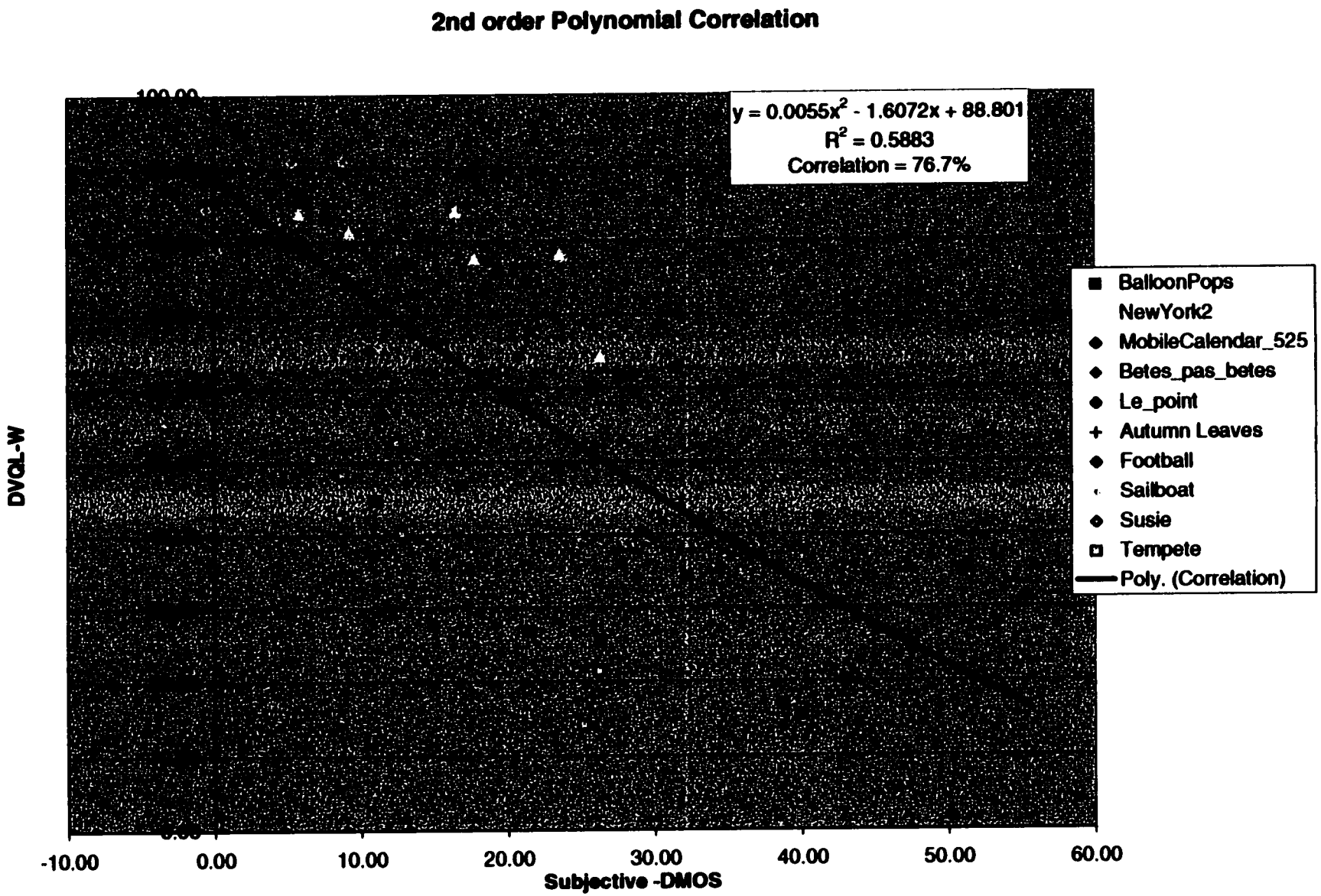
Correlation of Subective to Objective Measurement of Video Quality

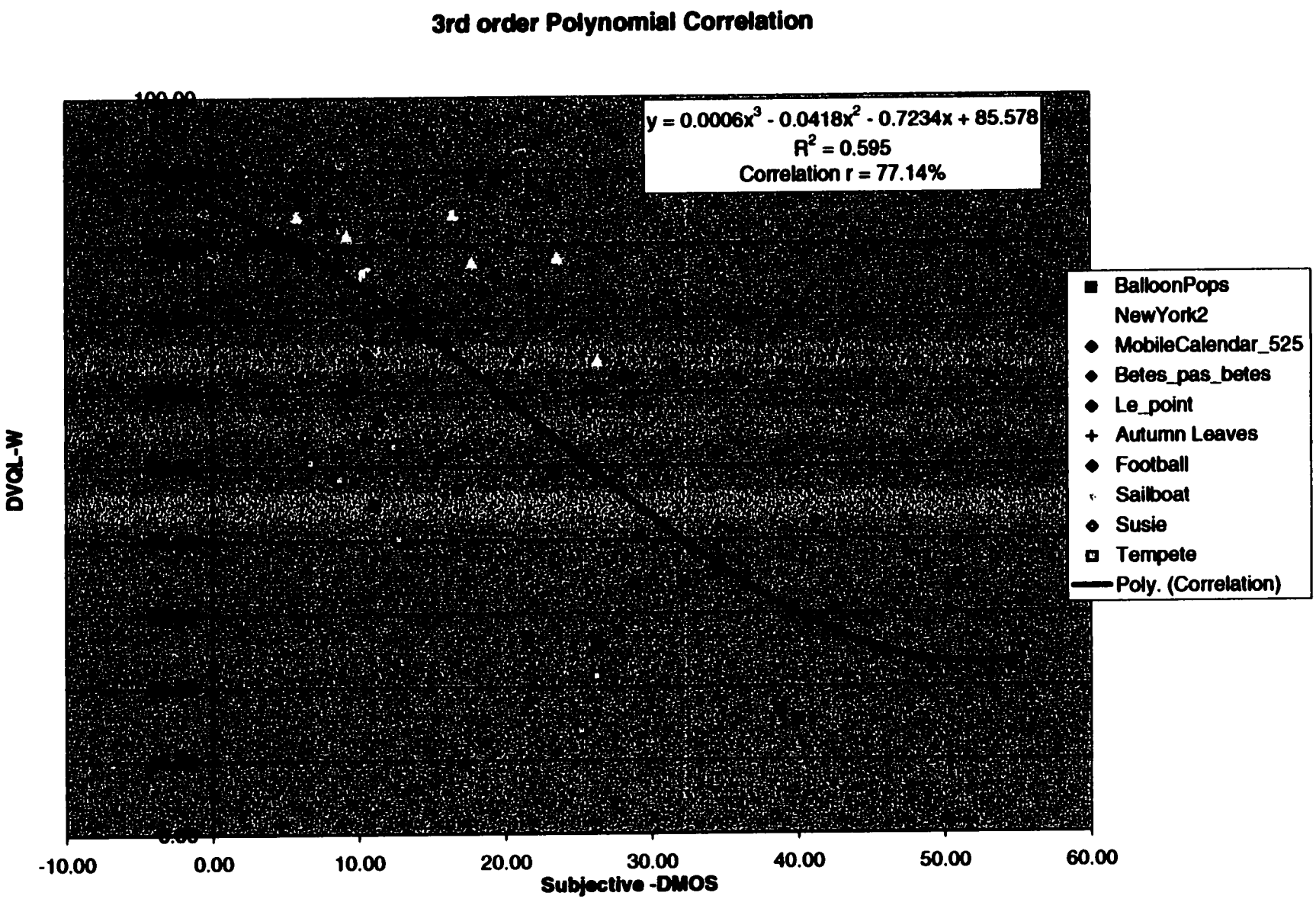
Sequence/HRC	source/HRC no.	Objective DVQL-W	DMOS Subjective
HRC2_BalloonPops	Src13/Hrc2	79.50	5.69
HRC5_BalloonPops	Src13/Hrc5	44.70	11.06
HRC9_BalloonPops	Src13/Hrc9	25.56	26.31
HRC10_BalloonPops	Src13/Hrc10	46.80	16.80
HRC11_BalloonPops	Src13/Hrc11	16.90	38.73
HRC12_BalloonPops	Src13/Hrc12	26.82	21.56
HRC13_BalloonPops	Src13/Hrc13	34.20	32.19
HRC14_BalloonPops	Src13/Hrc14	15.60	40.01
HRC2_NewYork2	Src14/Hrc2	90.20	2.16
HRC5_NewYork2	Src14/Hrc5	82.30	7.97
HRC9_NewYork2	Src14/Hrc9	81.30	9.25
HRC10_NewYork2	Src14/Hrc10	83.90	5.82
HRC11_NewYork2	Src14/Hrc11	77.60	17.76
HRC12_NewYork2	Src14/Hrc12	84.10	16.47
HRC13_NewYork2	Src14/Hrc13	64.20	26.37
HRC14_NewYork2	Src14/Hrc14	78.22	23.60
HRC2_525_MobileCalendar	Src15/Hrc2	88.90	17.76
HRC5_525_MobileCalendar	Src15/Hrc5	74.60	21.36
HRC9_525_MobileCalendar	Src15/Hrc9	60.55	25.95
HRC10_525_MobileCalendar	Src15/Hrc10	73.78	18.57
HRC11_525_MobileCalendar	Src15/Hrc11	61.90	28.59
HRC12_525_MobileCalendar	Src15/Hrc12	75.78	19.39
HRC13_525_MobileCalendar	Src15/Hrc13	31.44	55.29
HRC14_525_MobileCalendar	Src15/Hrc14	63.00	31.64
HRC2_Betes_pas_betes	Src16/Hrc2	92.90	2.34
HRC5_Betes_pas_betes	Src16/Hrc5	86.56	3.90
HRC9_Betes_pas_betes	Src16/Hrc9	86.89	5.57
HRC10_Betes_pas_betes	Src16/Hrc10	89.20	19.47
HRC11_Betes_pas_betes	Src16/Hrc11	80.40	14.04
HRC12_Betes_pas_betes	Src16/Hrc12	85.60	6.19
HRC13_Betes_pas_betes	Src16/Hrc13	73.00	13.74
HRC14_Betes_pas_betes	Src16/Hrc14	83.44	7.70
HRC2_Le_point	Src17/Hrc2	90.10	1.99
HRC5_Le_point	Src17/Hrc5	69.30	10.71
HRC9_Le_point	Src17/Hrc9	51.33	15.58
HRC10_Le_point	Src17/Hrc10	67.90	13.63
HRC11_Le_point	Src17/Hrc11	49.20	23.38
HRC12_Le_point	Src17/Hrc12	65.22	10.61
HRC13_Le_point	Src17/Hrc13	1.80	50.16
HRC14_Le_point	Src17/Hrc14	34.10	28.80

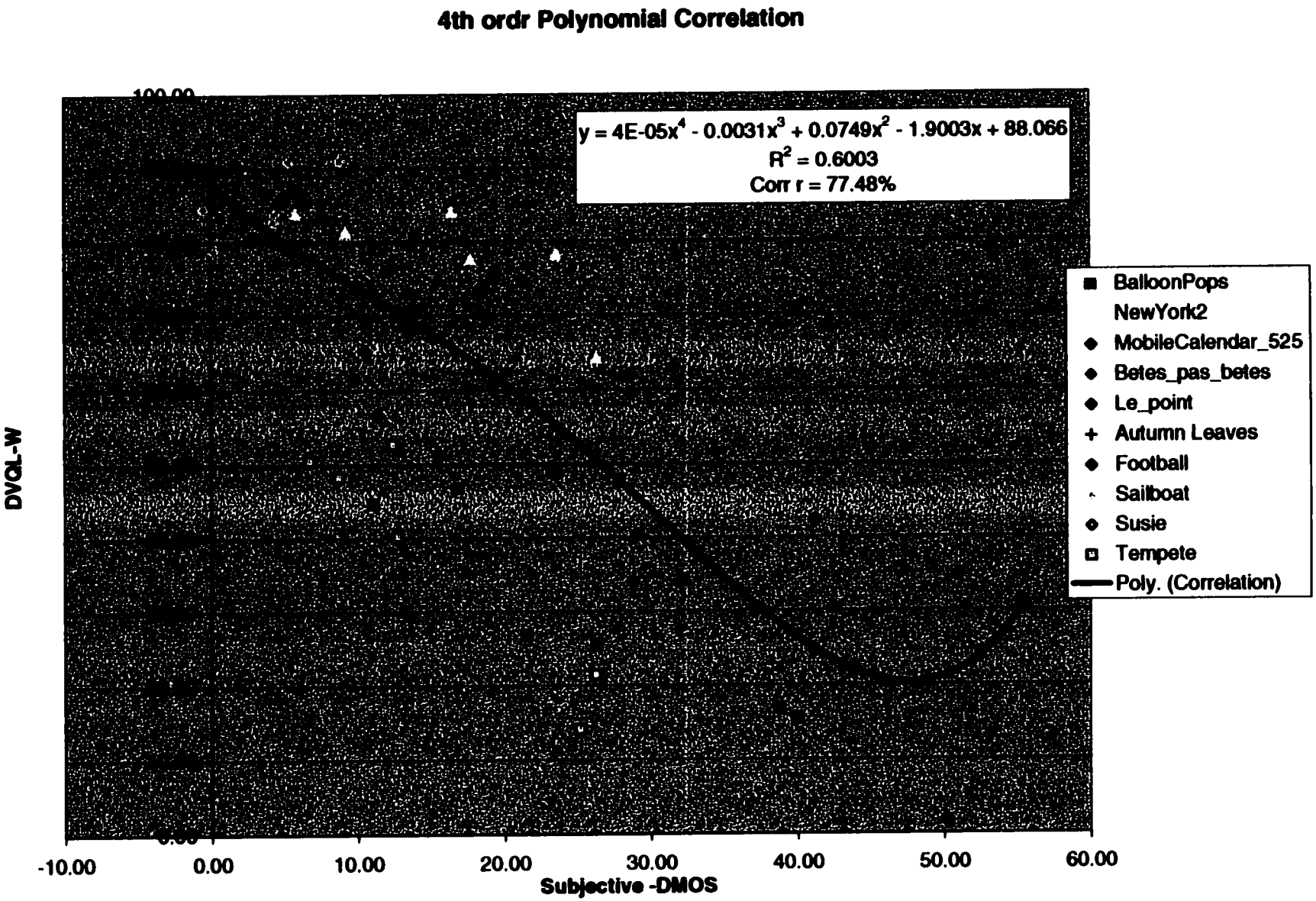
Sequence/HRC	source/HRC no.	Objective DVQL-W	DMOS Subjective
HRC2_AutumnLeaves	Src18/Hrc2	86.67	5.96
HRC5_AutumnLeaves	Src18/Hrc5	81.80	12.58
HRC9_AutumnLeaves	Src18/Hrc9	79.34	8.54
HRC10_AutumnLeaves	Src18/Hrc10	85.33	7.64
HRC11_AutumnLeaves	Src18/Hrc11	80.89	3.54
HRC12_AutumnLeaves	Src18/Hrc12	85.30	6.25
HRC13_AutumnLeaves	Src18/Hrc13	64.18	20.80
HRC14_AutumnLeaves	Src18/Hrc14	74.70	15.54
HRC2_Football	Src19/Hrc2	80.00	4.38
HRC5_Football	Src19/Hrc5	56.55	11.40
HRC9_Football	Src19/Hrc9	40.50	29.11
HRC10_Football	Src19/Hrc10	61.40	9.80
HRC11_Football	Src19/Hrc11	34.90	50.94
HRC12_Football	Src19/Hrc12	59.45	28.64
HRC13_Football	Src19/Hrc13	42.33	41.21
HRC14_Football	Src19/Hrc14	30.42	42.48
HRC2_Sailboat	Src20/Hrc2	84.44	-0.50
HRC5_Sailboat	Src20/Hrc5	82.70	4.29
HRC9_Sailboat	Src20/Hrc9	78.70	0.22
HRC10_Sailboat	Src20/Hrc10	90.80	5.36
HRC11_Sailboat	Src20/Hrc11	83.80	4.38
HRC12_Sailboat	Src20/Hrc12	91.00	8.80
HRC13_Sailboat	Src20/Hrc13	65.80	11.17
HRC2_Susie	Src21/Hrc2	89.60	6.41
HRC5_Susie	Src21/Hrc5	88.22	-0.66
HRC9_Susie	Src21/Hrc9	80.30	2.74
HRC10_Susie	Src21/Hrc10	88.80	-1.06
HRC11_Susie	Src21/Hrc11	76.10	12.23
HRC12_Susie	Src21/Hrc12	82.00	8.06
HRC13_Susie	Src21/Hrc13	75.40	3.30
HRC2_Tempete	Src22/Hrc2	72.30	4.31
HRC5_Tempete	Src22/Hrc5	48.30	8.74
HRC9_Tempete	Src22/Hrc9	50.50	6.75
HRC11_Tempete	Src22/Hrc11	40.20	12.76
HRC12_Tempete	Src22/Hrc12	52.67	12.44
HRC13_Tempete	Src22/Hrc13	14.11	25.19
HRC14_Tempete	Src22/Hrc14	21.50	26.25

-0.76545


Linear Correlation
y = -1.3555x + 87.108
R2 = 0.5859
r = 76.54%
DVQL-W
Subjective -DMOS
-10.00
0.00
10.00
20.00
30.00
40.00
50.00
60.00
BalloonPops
NewYork2
MobileCalendar_525
Betes_pas_betes
Le_point
Autumn Leaves
Football
Sailboat
Susie
Tempete
Linear (Correlation)


2nd order Polynomial Correlation
y = 0.0055x² - 1.6072x + 88.801
R² = 0.5883
Correlation = 76.7%
BalloonPops
NewYork2
MobileCalendar_525
Betes_pas_betes
Le_point
Autumn Leaves
Football
Sailboat
Susie
Tempete
Poly. (Correlation)
DVQL-W
-10.00
0.00
10.00
20.00
30.00
40.00
50.00
60.00
Subjective -DMOS


3rd order Polynomial Correlation
y = 0.0006x3 - 0.0418x2 - 0.7234x + 85.578
R2 = 0.595
Correlation r = 77.14%
DVQL-W
Subjective -DMOS
-10.00
0.00
10.00
20.00
30.00
40.00
50.00
60.00
BalloonPops
NewYork2
MobileCalendar_525
Betes_pas_betes
Le_point
Autumn Leaves
Football
Sailboat
Susie
Tempete
Poly. (Correlation)


4th ordr Polynomial Correlation
y = 4E-05x^4 - 0.0031x^3 + 0.0749x^2 - 1.9003x + 88.066
R^2 = 0.6003
Corr r = 77.48%
DVQL-W
Subjective -DMOS
-10.00
0.00
10.00
20.00
30.00
40.00
50.00
60.00
BalloonPops
NewYork2
MobileCalendar_525
Betes_pas_betes
Le_point
Autumn Leaves
Football
Sailboat
Susie
Tempete
Poly. (Correlation)

Chapter 7
Conclusion and Future Work

The objective of this Thesis was to study a generalization method of the ITU recommendations and VQEG work to develop a simple application for In-field calibration procedure for TV networks to validate the video quality subjective/objective correlation for an objective video quality monitoring system based on their individualized set of Network characterization.

A simulation of a typical TV network variety of material and set of different data rates and inserted artifacts allocated. This was a subset of 60Hz low and high quality MPEG2 compressed video programming material in different MPEG profiles. And made a decent simulation from the VQEG source material and HRCs used in their report [VQEG00], and also by Woerner and Lauterjung [Woer00]. The real subjective testing for networks would have to follow a different set of procedures described in chapter 5 to ensure the test follows a standard like the ITU-500 recommendation in a level corresponding to viewer or operational level as described in the recommendations.

The objective test used a simple setup as per fig 6.10.

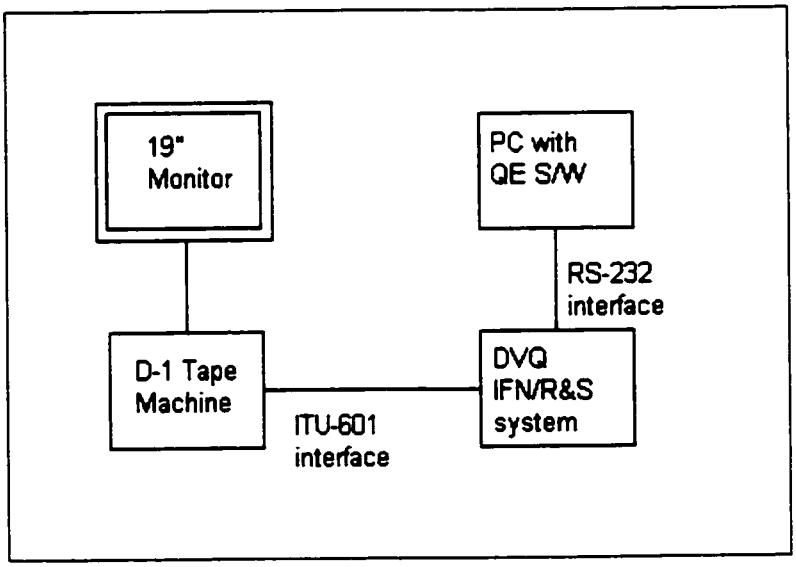


Fig 6.10 Test Setup for the Objective Measurement

Using the Pearson linear regression between subjective DMOS values and obtained objective predicted values would give the correlation relationship.

This technique will characterize the Network correlation based on it's specific characterization such as type of material, the encoders and decoders used, transmission links and set of expected artifacts in this specific case. The technique could be further expanded to characterize other quality evaluation systems used in the Network, to create an integrated quality system with known performance.

The wide spread of digital TV and MPEG2 compressed video as per the ATSC terrestrial standard in North America and the DVB standard in Europe and the rest of the world for terrestrial, cable and satellite, it becomes very important to start developing new standards for quality measurements, and create that "Echelon of Standards" for the whole industry. As the subjective measurement is the main reference of how we perceive video quality, all standards have to refer to that. Creating a standard that puts in consideration the individual circumstances and differential factors for each network is important to create a reference of quality of MPEG2 compressed video from the viewer perspective, which eventually is the real customer for Broadcasting.

There is a lot of work in the area of video quality and its' standards. A lot of work is done in the metrics for video quality and development of new metrics that perform better than earlier models. Current models are also enhancing their original models. Thus being able to calibrate any enhanced model as described, would be critical for the users (TV Networks) to see the difference in performance in a measured way.

An area like nonlinear approach to combine different features in a video stream to come up with video quality value is a particularly interesting field. Lin [Lin95] has used a back-propagation neural network for that. It's my personal view that using non-linear methods like Neural Network

and maybe Fuzzy Logic could open the door to customizing the quality evaluations "weighting" process even more to follow subjective evaluation even more closely. This is an application that Neural Network has good advantage in because of its' training process capability. Some TV networks for example need to calibrate that reading to their Golden eye evaluations, not regular viewers as per ITU-500 recommendation.

As per other research groups like VQEG and T1A1 in the USA, the VQEG will do similar round of video quality evaluations for reduced reference and no reference algorithms, since there has been a few developed in the past couple of years. Their work as mentioned earlier will help the ITU, SMPTE and other international bodies create more recommendations and standards in the field.

There is further work in the development of encoders to utilize low bandwidth data rate and get better quality.

The other area that is opening a huge field is the spread of HDTV broadcast, and the need to develop quality models and metrics for those formats. At the moment, trying to process uncompressed decoded Standard Definition video in the data rates of 270 Mbits/sec. has been difficult. Thus going the next step of uncompressed video streams of 1.55 Gbits/sec. is still quite challenging. However, there are efforts in this area at the moment.

The other area that needs to be addressed more, and there are efforts in, is the analysis of audio quality embedded in MPEG2 streams. Especially going to Dolby-AC3 in the ATSC, there are 6 audio channels (referred to as 5.1 channels), and getting into such analysis will be important.

8- References:

1- Ara01 — Arabatti, A.; Myler, H.; "An Objective Quality Measure Based on Subjective Measure for JPEG Compression"; Teranex Inc, Orlando, FL, USA, VQEG meeting, Orlando, FL, July/2001

2- Bai01 — Baina, J.; Bretillon, P.; Goudezeune, G.; "Reference Model for Quality Meter Description"; TDF, France, VQEG meeting, FL, USA, July/2001

3- - Bass96 — Basso, A.; Dalgic, I.; Tobagi, F.; and Van den Brander Lambrecht, C: "Study of MPEG-2 coding performance based on perceptual quality metric"; in PCS'96 International Picture Coding Symposium, vol. 1, pp. 263-268, Mar. 1996

4- Bee94 — Beerends, J.G.; "Modelling a cognitive effects that play a role in the perception of speech quality, in Speech Quality Assessment, Workshop papers, Bochum, pp1-9, November, 1994

5- Bee97 — Beerends, J. G., "Objective Measurement of Video Quality", KPN research, Netherlands, ITU study group 12-contribution 7, Feb./1997

6- Bha97 — Bhaskaran, V. & Konstantinides, K.: "Image and Video Compression Standards" – Algorithms and Architectures – Second Edition, Kluwer Academic Publishers, Norwell, Mass., USA, 1997

7- Bru99 — Bruin, R.; Smits, J., "Digital Video Broadcasting; Technology, Standards and Regulations", Artech House, Norwood, MA, USA, 1999

8- Burt83 — Burt, P.J.; Adelson, E.; "The Laplacian pyramid as a compact image code"; IEEE Transactions on Communications, COM-31, 532-540, 1983

9- Car96 — Carney, T.; Klein, S.; and Hu, Q; "Visual masking near spatiotemporal edges", in Human Vision and Electronic Imaging (B.E. Rogowitz and J. Allebach, eds.) vol. 2657, pp393-402, SPIE, 1996

10- Cas96 — Castleman, K. R., "Digital Image Processing", Prentice Hall Inc., New Jersey, USA, 1996

11- Cav01 — Caviedes, J.; Jung, J.;"No Reference Metric for a Video Quality Control Loop"; Philips Research, NY,USA/ Philips France, VQEG meeting, Orlando, FL, July/2001

12- Coom72 — Coombs, C. F.; "Basic Electronic Instrument handbook";

Mcgraw-Hill Book Co., USA, 1972

13- Daly93 — Daly, S; "The visible differences predictor, an algorithm for the assessment of image fidelity"; in Digital Image and Human Vision (A. Watson, ed.), ch 7, pp 179-208, MIT Press, Cambridge, Massachusetts, 1993

14- Eff98 — Effelsberg, W.; Steinmetz, R., "Video Compression Techniques", dpunkt-verlag, Heidelberg, Germany 1998

15- Eva95 — Evans, B., "Understanding Digital TV; The route to HDTV", IEEE press, NY, USA, 1995

16- Fen98 — Fenimore, C.; Libert, J.; Wolf, S.; "Perceptual Effects of Noise in Digital Video Compression"; NTIA, CO, USA, SMPTE meeting, Oct 28-31/1998

17- Fib87 — Fibush, D. K., " A Guide to Digital Television Systems and Measurements", Tektronix Inc., USA, 1987.

18- Fis95 — Fisher, Y., "Fractal Image Compression", Springer-Verlag, NY, USA, 1995

19- Fra98 — Fanzen, N. et al, " Verification of Normalization Processing for VQEG Video Test Data", Tektronix Measurement Div., 18, May/1998

20- Ham00 — Hamada, T., "Video Codec Evaluation Scheme and Implementation Based on Characteristics of Human Visual Perception", KDD R&D laboratories, Japan, 2000

21- Ham00_2 — Hamada, T., "Proposal for an Object Based Model Considering Object Complexity", KDD Labs/Pixelmetrix, Japan, VQEG meeting, Ottawa, March/2000

22- Ham01 — Hamada, T.; "Digital Video/Audio NR+RR Monitoring System Based on Motion Compensated Interframe/ Intraframe Objective Parameters"; KDD Corp., Japan, VQEG meeting, Orlando, FL, USA, July/2001

23- Ham97 — Hamada, H; Namba, S; "Objective Picture Quality Scale for Digital Compressed Picture for Broadcast"; NHK Research Labs.; VQEG meeting, Oct./1997

24- Has97 — Haskell, B. G., "Digital Video: an introduction to MPEG-2", Chapman and Hall, USA, 1997

25- ITU-500 — ITU-R Recommendation BT.500-10: "Methodology for the

Subjective Assessment of the Quality of Television Pictures", ITU, geneva, Switzerland, March/2000

26- ITU-813 ITU-R Recommendation BT.813: "Methods for Objective picture quality assessment in relation to impairments from digital coding of Television Signals", ITU, Geneva, Switzerland, 1992

27 - Jay93 Jayant, N.; Johnston, J.; and Safranek, R.; "Signal compression based on models of human perception"; Proceedings of the IEEE, vol. 81, pp. 1385-1422, Oct. 1993

28 - Katt91 Katto, J.; Onda, K.; and Yasuda, Y.; "Variable bit-rate coding based on human visual system"; Signal Processing: Image Communication, vol. 3, pp. 321-331, Sept. 1991

29- Kaw01 Kawada, R; "A method for estimating PSNR of coded pictures by using of embedded invisible markers"; KDDI Labs Inc., Japan, VQEG meeting, Boulder, CO, May/2001

30- Klein92 Klein, S.; Silverstein, A; Carney, T.; "Relevance of human visionto JPEG-DCT compression"; in Human Vision, Visual Processing, and Digital Display III (B.E. Rogowitz, ed.), vol 1666, pp. 200-215, SPIE, 1992

31- Kne01 Knee, M, "A Single-Ended Picture Quality Measure for MPEG-2", Snell & Wilcox, UK, VQEG meeting, Orlando, FL., USA, July 2001

32- Kuh00 Kuhn, M.; Antkowiak, J.;" Statistical Multiplex - what does it mean for DVB-T", FKT Fachzeitschrift fur Fernsehen, Film und elektronische Medien, Germany- April/2000

33- Lamb96 Van den Branden Lambrecht, C; " A working spatio-temporal model of the human visual system for image restoration and quality assessment applications"; in International Conference on Acoustics, Speech and Signal Processing ICASSP' 96, 1996

34- Lau98 Lauterjung, J.; "Picture Quality Measurement"; Rohde & Schwarz GmbH, Munich, Germany. Proceedings of the International Broadcasting Convention IBC 1998, Amsterdam, Netherlands, 1998

35- Lau99 Lauterjung, J. 1999; "First Results of digital video quality measurement in DVB networks. Proceeding of the Radio Show, Montreux, Switzerland, 1999

36- Lin95 Lin, F.; Mersereau, R; "A constant subjective quality MPEG encoder"; in International Conference on Acoustics, Speech and Signal Processing ICASSP'95, pp. 2177-2180, 1993

37- Lub00 Lubin, J.; "The Use of Psychophysical Data and Models in the Analysis of Display System Performance"; Sarnoff Corp., USA, VQEG meeting, Munich, Germany, June/2000

38- Lut99 Luther, C.; Inglis, A., "Video Engineering", McGraw Hill Inc., 3rd edition, USA 1999.

39- Mau92 Mausl, R., "Refresher topics – Television technology", Rohde & Schwarz GmbH, Germany 1992

40- Meer97 Van der Meer, P., "Variable Bit Rate Compressed Video", Delft Univ. Press, Delft, Netherlands, 1997

41- Miya92 Miyahara, M.; Kotani, K; & Algazi, V.; "Objective Picture Quality Scale (PQS) for image coding" SID digest, pp859-862, 1992

42- Nach74 Nachmias, J.; Sansbury, R.; "Grating contrast: Discrimination may be better than detection"; Vision Res. 14, 1039-1042, 1974

43- Nish00 ITU, study group 9;"Objective Picture Quality Evaluation For the Digital Broadcasting By Introduction of Criticality"; Geneva, May 15-19, 2000

44- Nish97 Nishida, Yukihiro; "Objective Picture Quality Evaluation For the Digital Broadcasting By Introduction of Criticality"; NHK Research Labs, Japan, VQEG meeting 14-16 Oct./1997

45- Olz86 Olzak, L.; Thomas, J.; "Seeing spatial pattern", in Handbook of perception and Human Performance (K. Boff, L. Kaufman, and J. Thomas eds.), ch.7, Wiley, New York, 1986.

46- Pan99 Panasonic Technology Inc., "The Video Compression Book", 1999

47- Pess97 Pessoa, A; ITU-T Study Group 12, Contribution12-39, "Video Quality Assessment using objective parameters based on image segmentation"; December 1997

48- Pete91 Peterson, H.; Peng, H.; Morgan, J. and Watson, A.;" Quantization of color image components in the DCT domain"; in Human Vision, Visual Processing and Digital Display II (B.E. Rogowitz, M. Brill, and J. Allebach, eds.), vol. 1453, pp. 210-222, SPIE, 1991

49- Pete92	Peterson, H.; "DCT basis functions visibility in RGB space"; in Society for Information Display Digest of technical Papers (J. Morreale, ed.), Society for Information Display, 1992
50- Poy96	Poynton, C.; "A Technical Introduction of Digital Video", John Wiley & Sons Inc., USA 1996.
51- Rob98	Robin, M.; Poulin, M., "Digital Television Fundamentals: Design and installation of video and audio systems", McGraw Hill, USA 1998.
52- Rob99	Robin, M., "Testing MPEG compressed Video", December/1999 issue, page 80-88, Broadcast Engineering, USA, Dec./1999.
53- Sag89	Saghri, J; Cheatham, P; Habibi, A; "Image Quality measure based on a human visual system model"; Optical Engineering, vol 28, pp. 813-818, July 1989.
54- Sar97	Sarnoff Corp., "Sarnoff JND Vision Model Algorithm Description and Testing", Princeton, NJ, USA, August 4, 1997
55- Saw99	Saw, Y., "Rate-Quality Optimized Video Coding", Kluwer Academic Publishers, Mass. USA, 1999
56- Sug01	Sugimoto, O. et al; "Automatic Objective Picture Quality Measurement Method using invisible Markers Signals as Reduced References", KDDI R&D Labs Inc., Saitama, Japan, VQEG meeting, FI, USA, July/2001
57- Sym98	Symes, P., "Video Compression", McGraw Hill, USA 1998
58- Tan01	Tan, K.T.; Ghanbari, M; "Blockiness Measurement for MPEG Video";Singaore Polytechnic/ Univ. of Essex, VQEG meeting, Orlando, FL, July/2001
59- Tek97	Tektronix Inc., "A Guide to Picture Quality Measurements for Modern Systems", 1997.
60- Tek98	Tektronix Inc., "A Guide to MPEG Fundamentals and Protocol Analysis (Including DVB and ATSC)", 1998.
61- Vass73	Vassilev, A; "Contrast Sensitivity near borders: Significance of test stimulus form, size and duration", Vision Research vol. 13, pp. 719-730, 1973
62- Vor92	Voran, S.; "The Effect of Multiple Scenes on Objective Video

Quality Assessment", NTIA, CO, USA, T1A1.5 meeting, Document# T1A1.5/92-136, July 13/1992

63- VQE00 VQEG, ITU, "Final Report from the Video Quality Expert Group on the Validation of Objective Models of Video Quality Assessment", March 2000

64- VQEG- ob98 ITU study group 12; "Evaluation of new methods for objective testing of video quality: objective test plan"; 1998

65- Wat99 Watkinson, J., "MPEG-2" , focal press, UK, 1999

66- Wes95 Westen, S.; Lagendijk, R.; Beimond J.; "Perceptual Image Quality based on a multiple channel HVS model"; in International Conference on Acoustics, Speech and Signal Processing ICASSP' 95, vol 4, pp 2351-2354, May 1995

67- Wes97 Westwater, R. & Furht, B., "Real-Time Video Compression – Techniques & Algorithms", Kluwer Academic Publishers, Norwell, Mass., USA, 1997

68- Wink99 Winkler, S.; "A Perceptual Distortion Metric for Digital Color Video", Human Vision and Elctronic Imaging IV, Proceedings Volume 3644, B.E. Rogowitz and T.N. Pappas eds., pages 175-184, SPIE, Bellingham, WA, 1999

69- Woe00 Woerner, A., "DVQ- An objective picture quality measurement not requiring a reference (NR)", Rohde & Schwarz GmbH, T1A1.1/2000-044 committee meeting, Charleston, Jan 29th, 2001

70- Woe01 Woerner, A.; Lauterjung, J; "A real time single ended algorithm for objective quality monitoring of compressed video signals", Rohde & Schwarz GmbH, Munich, Germany, VQEG meeting, Orlando, FL, USA, July, 2001

71- Wolf92 Wolf, S.; "An Automated Technique for Measuring Transmitted Frame Rate (TFR) and Average Frame Rate (AFR)", NTIA, CO, USA, T1A1.5 meeting, July, 13/1992

72- Wolf97 Wolf, S.; et al, "Objective and Subjective Measure of MPEG Video Quality"; NTIA, CO, USA, SMPTE meeting, Nov 21-24/1997

73- Wolf97 Wolf, S.; Webster, A.; "Subjective and Objective Measures of Scene Criticality", NTIA/ITS, Boulder, CO, USA, ITU meeting, Turin Italy Oct./1997

74- Wolf98 Wolf, S.; Pinson, M.; "In-Service Performance Metrics for

MPEG-2 Video Systems"; NTIA, CO, USA, IAB, Nov. 12-13/1998

75- Wolf99 Wolf, S.; Pinson, M.; "Spatial-Temporal distortion metrics for in-service quality monitoring of any digital video system"; NTIA, Boulder, CO., USA, Spie meeting, Sept, 11/1999

9- Acronyms and Glossary:

16 VSB	Vestigial sideband modulation with 16 discrete amplitude levels.
8 VSB	Vestigial sideband modulation with 8 discrete amplitude levels. VSB is an analog modulation technique used to reduce the amount of spectrum needed to transmit information through cable TV, or over-the-air broadcasts used in the NTSC (analog) standard. 8-VSB is the U.S. ATSC digital television transmission standard.
AES	Audio Engineering Society.
Aliasing	Defects or distortion in a television picture or audio. Defects are typically seen as jagged edges on diagonal lines and twinkling or brightening. In digital video, aliasing is caused by insufficient sampling or poor filtering of the digital video.
Anchor frame	A video frame that is used for prediction. I-frames and P-frames are generally used as anchor frames, but B-frames are never anchor frames.
ANOVA	ANalysis Of Variance
ANSI	American National Standards Institute.
Artifacts	Undesirable elements or defects in a video picture. Most common in digital are macroblocks, which resemble pixelation of the video image, and pops and clicks in audio.
ASI	Asynchronous serial interface
Aspect Ratio 16:9	Aspect ratio of widescreen DTV formats for all HDTV and some SDTV (Standard Definition) video. "16" unit width corresponds to "9" unit height, proportionally, regardless of the actual size of the screen.

Aspect Ratio 4:3	Aspect ratio of the NTSC TV screen, with "4" unit width corresponding to "3" unit height, proportionally, regardless of the actual size of the screen.
ATM	Asynchronous Transfer Mode. A digital signal protocol for efficient transport of both constant-rate and bursty information in broadband digital networks. The ATM digital stream consists of fixed-length packets called "cells," each containing 53 8-bit bytes—a 5-byte header and a 48-byte information payload.
ATSC	Advanced Television Standards Committee
Bit Rate	The rate at which the compressed bit stream is delivered from the channel to the input of a decoder.
Block	A block is an 8-by-8 array of pixel element values or DCT coefficients representing luminance or chrominance information.
B-pictures	Bidirectional picture. Pictures that use both future and past pictures as a reference. This technique is termed bidirectional prediction. B-pictures provide the most compression. B-pictures do not propagate coding errors as they are never used as a reference.
Byte-aligned	A bit in a coded bit stream is byte-aligned if its position is a multiple of 8-bits from the first bit in the stream.
CAT	Conditional access table
CBR	Constant Bit Rate, Operation where the bit rate is constant from start to finish of the compressed bit stream.
CCIR	Comite Consultatif International des Radiocommunications
cd/m2	Candles per square meter. A unit for measuring luminance.
CDTV	Conventional Definition TV. This term is used to signify the

	analog NTSC television system as defined in ITU-R Recommendation 470. See also standard definition television and ITU-R Recommendation 1125.
Channel	A digital medium that stores or transports a digital television stream.
CIF	Common Intermediate Format: 288 lines X 352 pixles/ line (Luminance), 144 lines X 176 pixels/line (Chrominance)
Codec	Coder-decoder. A device that converts analog video and audio signals into a digital format for transmission. Also converts received digital signals back into analog format.
COFDM	Coded orthogonal frequency division multiplexing. COFDM can transmit many streams of data simultaneously, each one occupying only a small portion of the total available bandwidth. The DTV standard used in Europe.
Compression	Reduction in the number of bits used to represent an item of data.
CRC	Communications Research Center (Canada)
DCT	Discrete Cosine Ttransform; A mathematical transform that can be perfectly undone and which is useful in image compression.
DDR	Digital Disk Recorder. A video recording device that uses a hard disk drive or optical disk drive mechanism. Disk recorders offer nearly instantaneous access to recorded material.
Decoded Stream	The decoded reconstruction of a compressed bit stream.
Decoder	An embodiment of a decoding process.
Decoding (process)	The process defined in the Digital Television Standard that reads an input coded bit stream and outputs decoded pictures or audio samples.
D-frame	Frame coded according to an MPEG-1 mode which uses

	DC coefficients only.
Digital Betacam	A development of the original analog Betacam which records digitally on a Betacam-style cassette by Sony.
DMOS	Differential Mean Opinion Score
DMOSp	Differential Mean Opinion Score – predicted
Dolby AC-3	Digital Dolby. The approved 5.1 channel (surround-sound) audio standard for ATSC digital television, using approximately 13:1 compression. Six discreet audio channels are used: Left, Center, Right, Left Rear (or side), Right Rear (or side), and a subwoofer -- LFE, "low frequency effects" -- (considered the ".1" as it is limited in bandwidth).
DSCQS	Double Stimulus Continuous Quality Scale
DTS	Decoding time stamp
DVB-C	Digital Video Broadcasting-Cable
DVB-S	Digital Video Broadcasting-Satellite
DVB-T	Digital Video Broadcasting-Terrestrial
DVCR	Digital video cassette recorder
EIT	Event information table
Elementary stream (ES)	A generic term for one of the coded video, coded audio or other coded bit streams. One elementary stream is carried in a sequence of PES packets with one and only one stream_id.
Encoder	An embodiment of an encoding process.
Encoding (process)	A process that reads a stream of input pictures or audio samples and produces a valid coded bit stream as defined in the Digital Television Standard.
Entropy coding	Variable length lossless coding of the digital representation of a signal

	to reduce redundancy.
Event	An event is defined as a collection of elementary streams with a common time base, an associated start time, and an associated end time.
Field	For an interlaced video signal, a "field" is the assembly of alternate lines of a frame. Therefore, an interlaced frame is composed of two fields, a top field and a bottom field.
FR	Full Reference
Frame	A frame contains lines of spatial information of a video signal. For progressive video, these lines contain samples starting from one time instant and continuing through successive lines to the bottom of the frame. For interlaced video a frame consists of two fields, a top field and a bottom field. One of these fields will commence one field later than the other.
GOP	Group of pictures; A group of pictures consists of a sequence of I, B and P frames
HDTV	High definition television has a resolution of approximately twice that of conventional television in both the horizontal (H) and vertical (V) dimensions and a picture aspect ratio (HxV) of 16:9. ITU-R Recommendation 1125 further defines "HDTV quality" as the delivery of a television picture which is subjectively identical with the interlaced HDTV studio standard.
High Level	A range of allowed picture parameters defined by the MPEG-2 video coding specification which corresponds to high definition television.
HRC	Hypothetical Reference Circuit
Huffman coding	A type of source coding that uses codes of different lengths to represent symbols which have unequal likelihood of occurrence.
I Frames	Pictures that are coded using information present only in the picture itself

	and not depending on information from other pictures. I-pictures provide a mechanism for random access into the compressed video data. I-pictures employ transform coding of the pel blocks and provide only moderate compression.
IEC	International Electrotechnical Commission.
IRT	Institut Rundfunk Technische (Germany)
ISO	International Organization for Standardization.
ITU	International Telecommunications Union
JEC	Joint Engineering Committee of EIA and NCTA.
JND	Just noticeable difference. Quality system by Sarnoff labs.
Level	A range of allowed picture parameters and combinations of picture parameters.
Macroblock	The four 8 by 8 blocks of luminance data and the two (for 4:2:0 chroma format), four (for 4:2:2 chroma format) or eight (for 4:4:4 chroma format) corresponding 8 by 8 blocks of chrominance data coming from a 16 by 16 section of the luminance component of the picture. Macroblock is sometimes used to refer to the pel (Pixel Elements) data and sometimes to the coded representation of the pel values and other data elements defined in the macroblock header.
Mbps	1,000,000 bits per second.
MOS	Mean Opinion Score
MOSp	Mean Opinion Score, predicted
Motion vector	A pair of numbers which represent the vertical and horizontal displacement of a region of a reference picture for prediction.
MP@HL	Main profile at high level.
MP@ML	Main profile at main level.
MPEG-1	Refers to ISO/IEC standards 11172-1 (Systems), 11172-2 (Video), 11172-3 (Audio), 11172-4 (Compliance Testing), and 11172-5 (Technical Report).

MPEG-2	Refers to ISO/IEC standards 13818-1 (Systems), 13818-2 (Video), 13818-3 (Audio), 13818-4 (Compliance).
MSE	Mean squared error
NIT	Network Information Table
NR	No (or Zero) Reference)
NTSC	National Television Standard Code. 525 line color system.
Packet	A packet consists of a header followed by a number of contiguous bytes from an elementary data stream. It is a layer in the system coding syntax.
Padding	A method to adjust the average length of an audio frame in time to the duration of the corresponding PCM samples, by continuously adding a slot to the audio frame
PAL	Phase Alternating Line color TV system.
PAT	Program associated table
Payload	Payload refers to the bytes which follow the header byte in a packet. For example, the payload of a transport stream packet includes the PES_packet_header and its PES_packet_data_bytes or pointer_field and PSI sections, or private data. A PES_packet_payload, however, consists only of PES_packet_data_bytes. The transport stream packet header and adaptation fields are not payload.
PCR	Program Clock Reference; A time stamp in the transport stream from which decoder timing is derived.
PCR	Program clock reference
PEM	Pooling of error means
PES	Packetized elementary stream.
PES Stream	A PES stream consists of PES packets, all of whose payloads consist of data from a single elementary stream, and all of which have the

	same stream_id.
Picture	Source, coded or reconstructed image data. A source or reconstructed picture consists of three rectangular matrices representing the luminance and two chrominance signals.
PID	Packet identifier; A unique integer value used to associate elementary Streams of a program in a single or multi-program transport stream.
Pixel	"Picture element" or "pel." A pixel is a digital sample of the color intensity values of a picture at a single point.
P-pictures	Pictures that are coded with respect to the nearest previous I or P-picture. This technique is termed forward prediction. P-pictures provide more compression than I-pictures and serve as a reference for future P-pictures or B-pictures. P-pictures can propagate coding errors when P-pictures (or B-pictures) are predicte from prior P-pictures where the prediction is flawed.
PQR	Picture quality rating
Profile	Predicted Pictures. A defined subset of the syntax specified in the MPEG-2 video coding specification
Program	A program is a collection of program elements. Program elements may be elementary streams. Program elements need not have any defined time base; those that do have a common time base and are intended for synchronized presentation.
PS	Program Segment
PSI	Program Specific Information; PSI consists of normative data which is necessary for the demultiplexing of transport streams and the successful regeneration of programs.
PSIP	Pronounced "P-SIP" - "Program and system information protocol."

	A part of the ATSC digital television specification that enables a DTV receiver to identify program information contributed by content providers and use it to create sophisticated electronic program guides.
PSNR	Peak signal to noise ratio
PTS	Presentation time-stamp; A field that may be present in a PES packet header that indicates the time that a presentation unit is presented in the system target decoder.
PTS	Presentation time stamp
QAM	Quadrature Amplitude Modulation
QCIF	Quarter Common Intermediate Format. Has half of the CIF resolution.
QPSK	Quadrature Phase Shift Keying
Quantizer	A processing step which intentionally reduces the precision of DCT coefficients
RR	Reduced Reference
RST	Running status table
SCR	System Clock Reference. A time stamp in the program stream from which decoder timing is derived.
Scrambling	The alteration of the characteristics of a video, audio or coded data stream in order to prevent unauthorized reception of the information in a clear form. This alteration is a specified process under the control of a conditional access system.

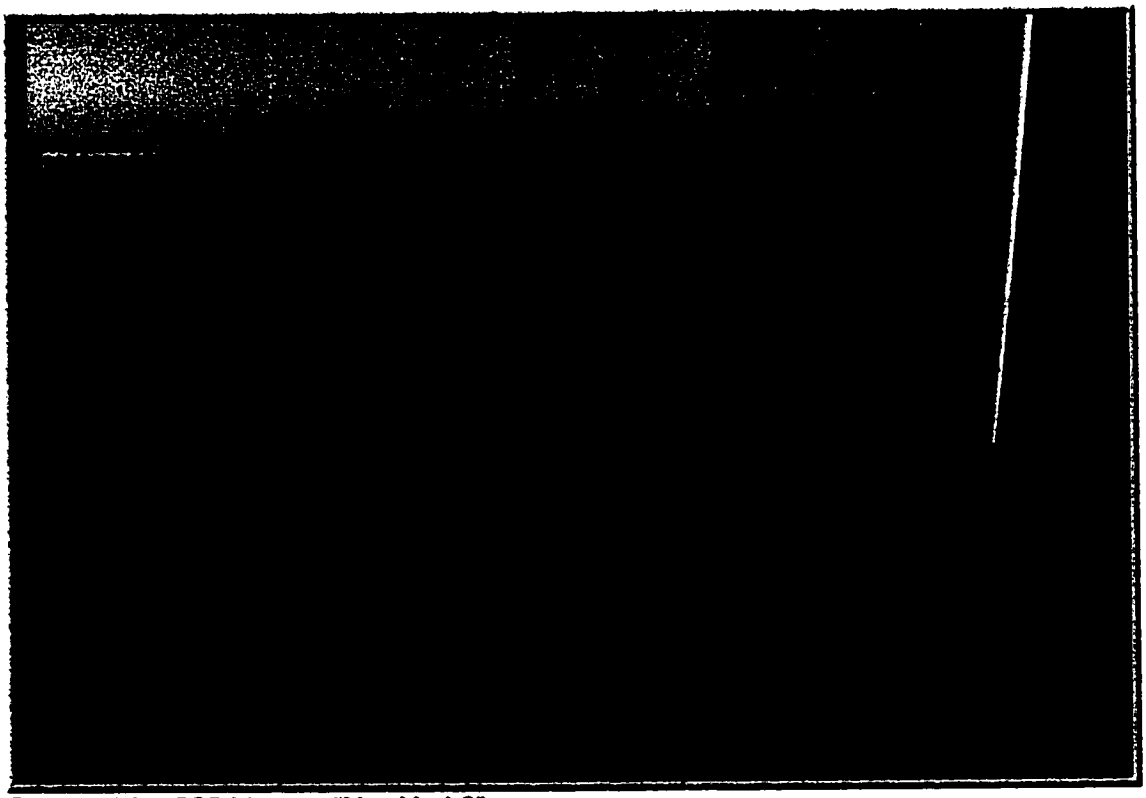
SDTV	Standard Definition Television. Digital formats that do not achieve the video quality of HDTV, but are at least equal, or superior to, NTSC pictures. SDTV may have either 4:3 or 16:9 aspect ratios, and it includes surround sound. Variations of fps (frames per second), lines of resolution, and other factors of 480p and 480i make up the 12 SDTV formats in the ATSC standard. This equivalent quality may be achieved from pictures sourced at the 4:2:2 level of ITU-R Recommendation 601 and subjected to processing as part of the bit rate compression. The results should be such that when judged across a representative sample of program material, subjective equivalence with NTSC is achieved. Also called standard digital television. See also conventional definition television and ITU-R Recommendation 1125.
SECAM	Systeme Electronique Coloeur avec Memoire
Slice	A series of consecutive macroblocks.
SMPTE	Society of Motion Picture and Television Engineers.
SPI	Synchronous parallel interface
SRC	Source Reference Channel or Circuit
SSCQE	Single Stimulus Continuous Quality Evaluation
TDT	Time and data table
TOT	Time offset table
VBR	Variable Bit Rate Operation where the bit rate varies with time during the decoding of a compressed bit stream.

Video Sequence	A video sequence is represented by a sequence header, one or more groups of pictures, and an end_of_sequence code in the data stream.
VQEG	Video Quality Experts Group
VTR	Video Tape Recorder

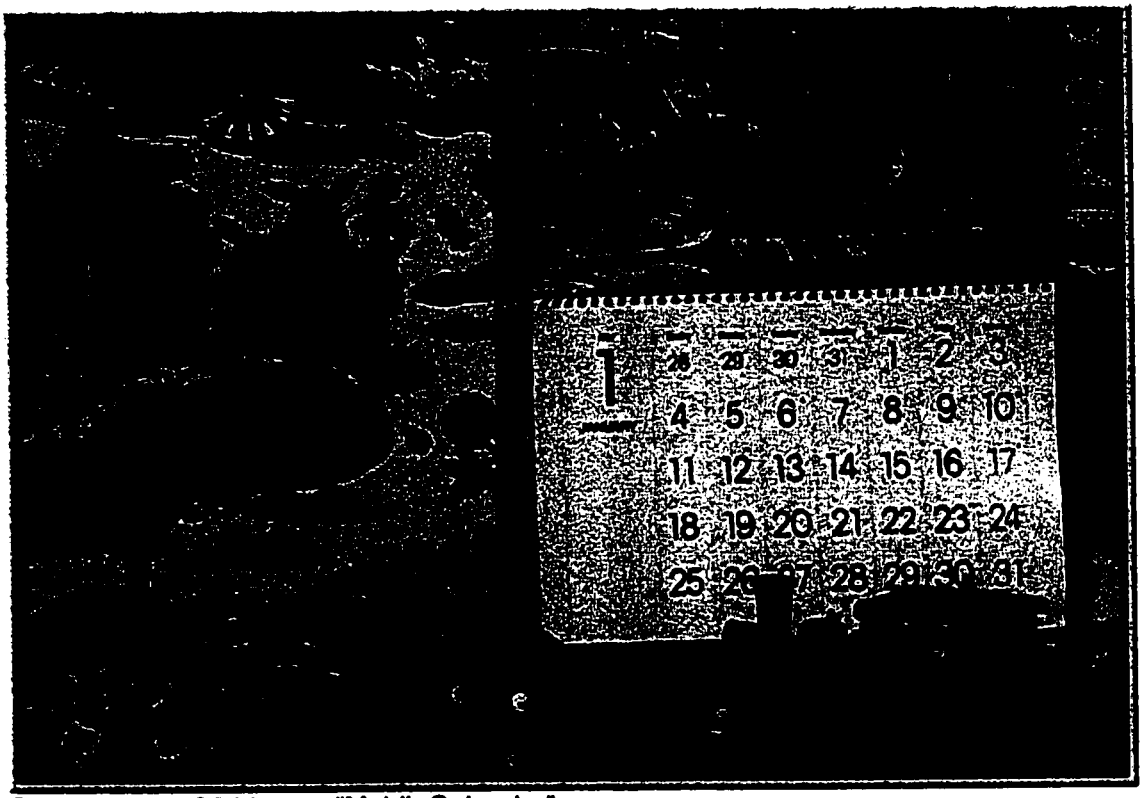
Appendix A



Source 13 – 525 Lines – "Baloon Pops"



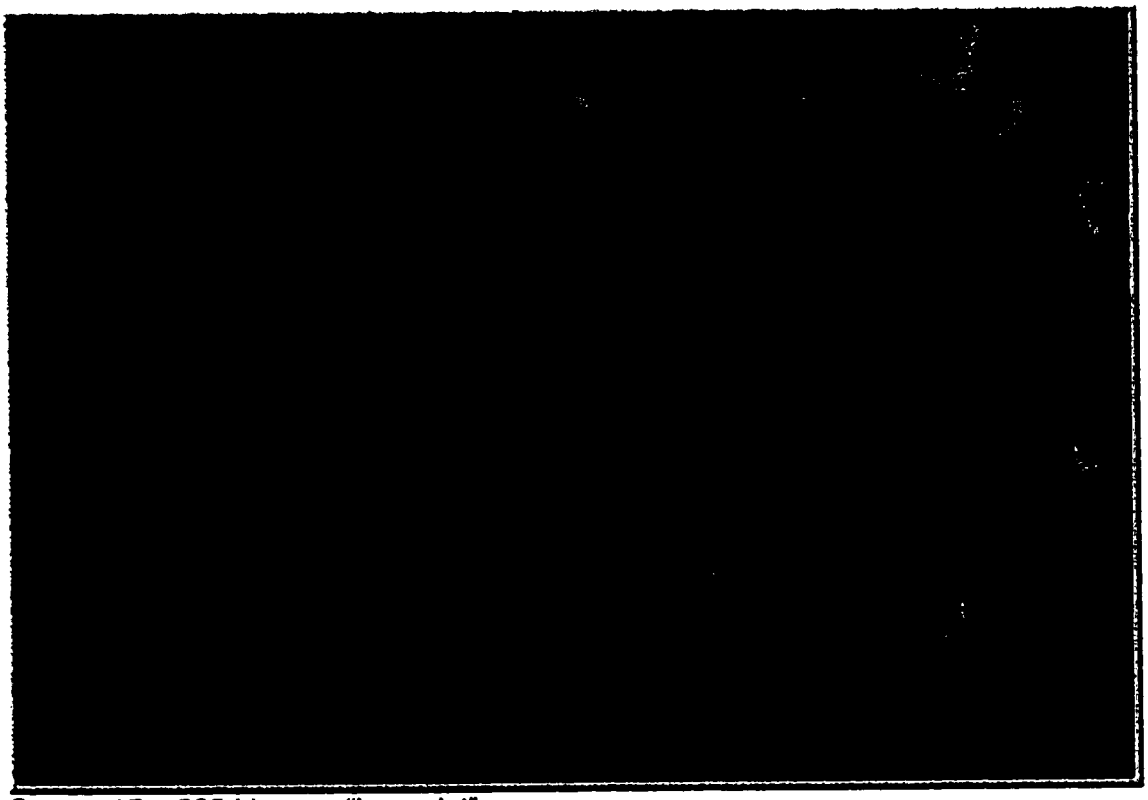
Source 14 – 525 Lines – "NewYork2"



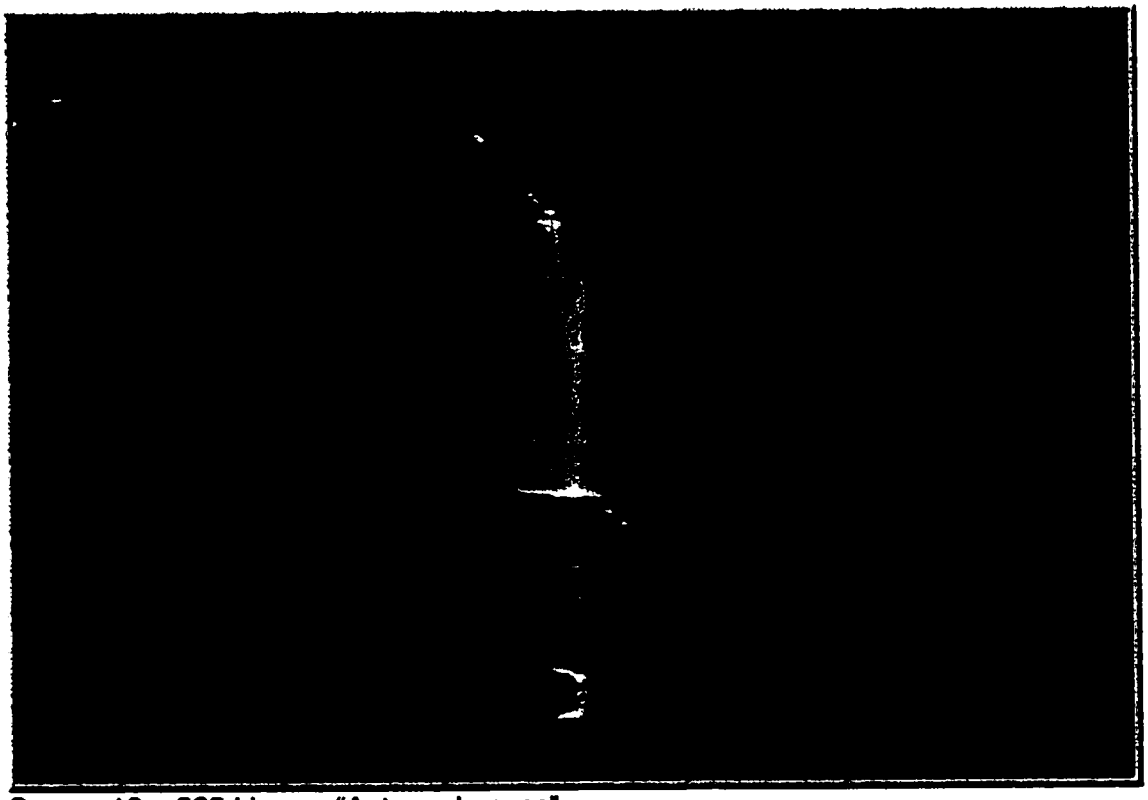
Source 15 – 525 Lines – "MobileCalendar"



Source 16 – 525 Lines – "Betes_pas_betes"



Source 17 – 525 Lines - : "Le_point"



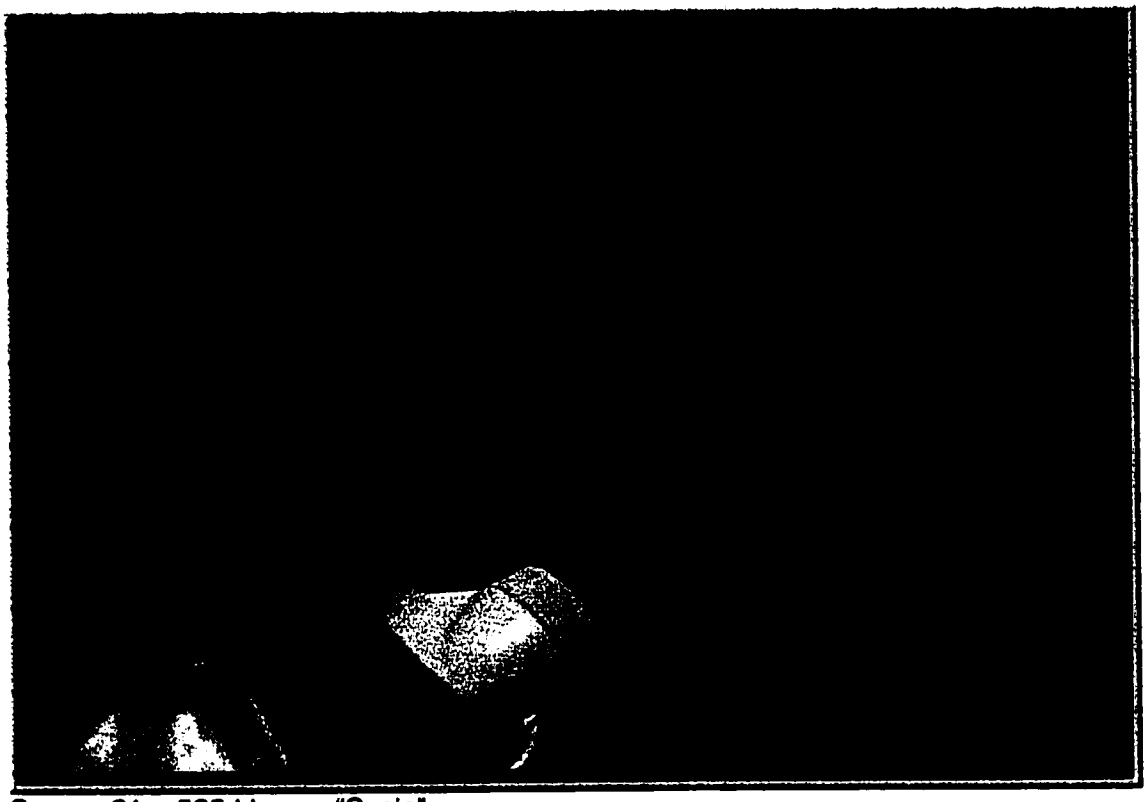
Source 18 – 525 Lines – "Autumn Leaves"



Source 19 – 525 Lines – "Football"



Source 20 - 525 Lines – "Sailboat"



Source 21 – 525 Lines – "Susie"



Source 22 – 525 Lines – "Tempete"

MQ

66049

UMI

MICROFILMED 2002

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

Université d'Ottawa • University of Ottawa

PERMISSION DE REPRODUIRE ET DE DISTRIBUER LA THÈSE

PERMISSION TO REPRODUCE AND DISTRIBUTE THE THESIS

NOM DE L'AUTEUR / NAME OF AUTHOR:	HASSINE, Jameleddine
ADRESSE POSTALE / MAILING ADDRESS:	185 Dalhousie Street Ottawa, Ontario K1N 7C7
GRADE / DEGREE: ~~Master of Computer Science~~ M.C.S.	**ANNÉE D'OBTENTION / YEAR GRANTED** 2001
TITRE DE LA THÈSE / TITLE OF THESIS: FEATURE INTERACTION FILTERING AND DETECTION WITH USE CASE MAPS AND LOTOS	

L'auteur permet, par la présente, la consultation et le prêt de cette thèse en conformité avec les règlements établis par le bibliothécaire en chef de l'Université d'Ottawa. L'auteur autorise aussi l'Université d'Ottawa, ses sumccesseurs et cessionnaires, à reproduire cet exemplaire par photograhie ou photocopie pour fins de prêt ou de vente au prix coûtant aux bibliothèques ou aux chercheurs qui en feront la demande.

Les droits de publication par tout autre moyen et pour vente au public demeureront la propriété de l'auteur de la thèse sous réserve des règlements de l'Université d'Ottawa en matière de publication de thèses.

The author hereby permits the consultation and the lending of this thesis pursuant to the regulations established by the Chief Librarian of the University of Ottawa. The author also authorizes the University of Ottawa, its successors and assignees, to make reproductions of this copy by photographic means or by photocopying and to lend or sell such reproductions at cost to libraries and to scholars requesting them.

The right to publish the thesis by other means and to sell it to the public is reserved to the author, subject to the regulations of the University of Ottawa governing the publication of theses.

N.B LE MASCULIN COMPREND ÉGALEMENT LE FÉMININ

21/01/2002

DATE

(AUTEUR) SIGNATURE (AUTHOR)



Université d'Ottawa • University of Ottawa

Université d'Ottawa • University of Ottawa

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

HASSINE, Jameleddine

AUTEUR DE LA THÈSE - AUTHOR OF THESIS

M.C.S.

GRADE - DEGREE

School of Information Technology and Engineering

FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

Feature Interaction Filtering and Detection with Use Case Maps and Lotos

Luigi Logrippo

DIRECTEUR DE LA THESE - THESIS SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

F. Bordeleau

A. Felty

J.-M. De Koninck, Ph.D.

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

SIGNATURE

DEAN OF THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

Feature Interaction Filtering and Detection with Use Case Maps and LOTOS

Jameleddine Hassine

Thesis submitted to
the school of graduate studies and research
in partial fulfillment of
the requirements for the degree of

Master of Computer Science

Under the auspices of the Ottawa-Carleton
Institute of Computer Science



University of Ottawa,
Ottawa, Ontario, Canada
February 2001

Copyright © Jameleddine Hassine, Ottawa, Canada, 2001

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-66049-4

Canada

To My Mother

Abstract

Telephony systems have evolved from the Plain Old Telephony System providing only the basic functionality of making phone calls, to sophisticated systems in which many features have been introduced, providing network subscribers more control on the call establishment process. However, these facilities are confronted with a major obstacle known as the feature interaction problem.

A feature interaction occurs when at least one feature is prevented from performing its functionality or when the system functions incorrectly due to the presence of features.

In the first part of the thesis, we present a model for describing telephony features at the requirements stage. This model is built using the Use Case Maps Notation (UCM). Based on this model, we propose a method to filter feature interactions at the requirements stage. This preliminary evaluation allows the detection process to focus on feature combinations where interactions are possible and therefore reduces the cost of the detection process.

In the second part of the thesis, a Feature Interaction Detection System is developed for detecting feature interactions between switch based and IN features. This method aims to detect interactions occurring at the abstract specification level and resulting in violation of feature properties. This technique in based on the Formal Description technique LOTOS and uses Abstract Data Types to detect those violations. Our method detects feature interaction by executing the system specification. The designer can reach those interaction points either by a step by step execution or using the goal oriented execution technique.

It is concluded that UCM and LOTOS are useful in specifying the telephony system with features and for detecting feature interactions at the abstract specification level.

Acknowledgment

I would like to thank all the people who helped me in completing this thesis.

First, I would like to express my deepest appreciation to my supervisor, Prof. Luigi Logrippo, for his encouragement, guidance and financial assistance through my graduate studies. His accuracy in reviewing the drafts of the thesis has greatly improved the contents and its presentation.

Second, I also owe special thanks to Dr. Masahide Nakamure for the inspiring and fruitful discussions we have had together.

I would also like to thank the members of University of Ottawa LOTOS group for all their advice and assistance. In particular, I would like to thank Daniel Amyot, Jacques Sincennes and Rossana M. Castro Andrade for their useful comments about some parts of the thesis.

Finally, I would like to express my gratitude for the financial support from the following sources: The Tunisian Government under the National scholarship program, Communications and Information Technology Ontario (CITO), Natural Sciences and Engineering Research Council of Canada (NSERC) and Nortel Networks.

TABLE OF CONTENTS

CHAPTER 1 7

INTRODUCTION: 7

MOTIVATION AND BACKGROUND 7

1.1 INTRODUCTION 7
1.2 THE FEATURE INTERACTION PROBLEM 8
1.2.1 Example of feature interaction 9
1.2.2 Addressing feature Interaction 10
1.3 CONTRIBUTION OF THE THESIS 12
1.3.1 Contribution 1: Feature Interaction Filtering at the Requirements Stage 12
1.3.2 Contribution 2: Feature Interaction Detection Method 12
1.4 ORGANIZATION OF THE THESIS 13

CHAPTER 2 15

RELATED WORK: 15

FEATURE INTERACTION DETECTION AT 15

REQUIREMENT AND SPECIFICATION STAGES 15

2.1 ADDRESSING FEATURE INTERACTIONS IN THE REQUIREMENTS STAGE 15
2.2 FORMAL SPECIFICATION METHODOLOGIES FOR TELEPHONY SYSTEMS 18
2.2.1 Finite State Machines 19
2.2.2 Petri Nets 19
2.2.3 LOTOS 21
2.2.4 SDL 21
2.3 ADDRESSING FEATURE INTERACTION AT SPECIFICATION LEVEL USING LOTOS 22

CHAPTER 3 26

DESCRIBING REQUIREMENTS 26

WITH USE CASE MAPS 26

3.1 INTRODUCTION 26
3.2 STAGE1: REQUIREMENT DESCRIPTION 27
3.2.1 Chisel Diagram Notation 28
3.2.2 Use Case Maps 31
3.2.2.1 Introduction 31
3.2.2.2 UCM Notation Elements 32
3.2.2.3 Advanced Notation Elements 34
3.2.2.4 Philosophy of UCMs 35

3.2.2.5 UCM Tools 35
3.3 STAGE2: MESSAGE SEQUENCE CHART 36
3.4 BENEFITS OF UCMS AND THEIR RELATION WITH MSCS 36

CHAPTER 4 39

FEATURE INTERACTION FILTERING AT 39

REQUIREMENT STAGE 39

4.1 MOTIVATION 39
4.2 DESCRIPTION OF SERVICES AT REQUIREMENT STAGE 40
4.2.1 Service Decomposition 41
4.2.2 UCM Call Model 41
4.2.2.1 Introduction 42
4.2.2.2 UCM service description 43
4.2.2.3 UCM Call Model 44
4.2.2.4 Stub Description 45
4.2.2.5 Plug-ins of POTS 47
4.3 FEATURE DESCRIPTION USING THE UCM MODEL 48
4.4 FEATURE ADDITION 49
4.4.1 Originating feature: INTL (IN Teen Line) 50
4.4.2 Terminating feature: CND (Calling Number Delivery) 50
4.4.3 Terminating feature: TCS (Terminating Call Screening) 51
4.4.4 Terminating feature: INFB (IN Free Phone Billing) 52
4.4.5 Terminating feature: CFBL (Call Forwarding Busy Line) 54
4.5 FEATURE INTERACTION FILTERING METHOD 54
4.5.1 Stub Configuration Vector 55
4.5.2 Feature Composition 58
4.5.3 Feature Interaction targeted 58
4.5.4 Filtering rules 60
4.5.5 Completeness 62
4.5.6 Feature Interaction Filtering Method 66
4.6 APPLICATION 71
4.6.1 Results 72
4.7 COMPARISON BETWEEN OUR METHOD AND THE METHOD PRESENTED IN [29] 73
4.8 CONCLUSION AND LIMITATIONS 74

CHAPTER 5 74

SPECIFYING FEATURES 74

USING LOTOS 74

5.1 INTRODUCTION 74
5.2 LOTOS ABSTRACT DATA TYPES 75
5.3 THE CONTROL COMPONENT 77
5.3.1 LOTOS Process 77
5.3.2 LOTOS Action 78
5.3.3 LOTOS Behaviour Expressions 78
5.3.4 LOTOS Operators 79
5.4 MAPPING FROM UCM TO LOTOS 83
5.5 THE LOTOS SPECIFICATION 87
5.5.1 Design of the Abstract Data Types 87
5.5.2 Feature Specification 89

LOTOS specification of CND (Call Number Delivery): 90

CHAPTER 6 92

FEATURE INTERACTION 92

DETECTION METHOD 92

6.1 FORMAL DEFINITION OF FEATURE INTERACTION 92
6.2 FEATURE INTERACTION DETECTION METHOD 94
6.2.1 Deriving Features Properties 96
6.2.2 UCM creation and Mapping to LOTOS 98
6.2.3 Introducing observation points in the specification 98
6.2.3.1 Information related to connection 99
6.2.3.2 Information related to Billing 101
6.2.3.3 Information related to Signals 102
6.2.3.4 Information related to Display 102
6.2.4 Design of the Abstract Data Types 103
6.2.4.1 Storing Connection information 103
6.2.4.2 Controlling signals 106
6.2.4.3 Storing Billing information 108
6.2.4.4 Display information 110
6.2.5 Feature Interaction Detection System Architecture 111
6.2.6 Design of the process FI Detector 114
6.3 DETECTION OF INTERACTIONS 116
6.4 FI SCENARIO GENERATION: GOAL ORIENTED EXECUTION (GOE) 119
6.5 APPLICATION 121
6.5.1 Interactions between OCS and CFBL 121
6.5.2 Interactions between INFB and CFBL 122
6.5.3 Interactions between INFB and OCS 124
6.6 RESULTS 125
Comparing our results with the benchmark Feature Interaction 125
6.7 CONCLUSION 129

CHAPTER 7 130

CONCLUSION & FUTURE WORK 130

APPENDIX 133

SPECIFICATION OF THE ABSTRACT DATA TYPES 133
SOME FEATURE SPECIFICATIONS 140

BIBLIOGRAPHY 145

LIST OF FIGURES

FIGURE 1: ILLUSTRATION OF A TELEPHONY FEATURE INTERACTION 10
FIGURE 2: FINITE STATE MACHINE 19
FIGURE 3: A SIMPLE PETRI NET 20
FIGURE 4: THREE STAGES METHODOLOGY 27
FIGURE 5: CHISEL DIAGRAM FOR POTS 29
FIGURE 6: UCM BASIC PATH 32
FIGURE 7: BOUND UCM 33
FIGURE 8: OR-FORKS/JOINS AND AND-FORKS/JOINS 34
FIGURE 9: STATIC STUBS HAVE ONLY ONE PLUG-IN (SUB-UCM) 34
FIGURE 10: DYNAMIC STUBS MAY HAVE MULTIPLE PLUG-IN 35
FIGURE 11: MSC EXAMPLE 36
FIGURE 12: CAUSAL SEQUENCE OF A UCM 37
FIGURE 13: A REFERENCE MODEL OF A PROTOCOL SYSTEM 40
FIGURE 14: UCM SERVICE MODEL 42
FIGURE 15: UCM CALL MODEL (ROOT MAP) 43
FIGURE 16: USE CASE MAPS FOR BASIC CALL MODEL 46
FIGURE 17: INTL PLUG-INS 49
FIGURE 18: CND PLUG-INS 50
FIGURE 19: TCS PLUG-INS 51
FIGURE 20: INFB PLUG-INS 52
FIGURE 21: CFBL PLUG-INS 53
FIGURE 22: COMPOSITION OPERATOR "." 55
FIGURE 23: FEATURE COMPOSITION (1) 56
FIGURE 24: FEATURE COMPOSITION (2) 57
FIGURE 25: NON DETERMINISM (INTERACTION) 59
FIGURE 26: $F \oplus G$:INTERACTION FREE CASE 1 61
FIGURE 27: $F \oplus G$ INTERACTION FREE CASE 2 61
FIGURE 28: $F \oplus G$ INTERACTION FREE CASE 3 62
FIGURE 29: UCM REFINEMENT 65
FIGURE 30: FEATURE CAUSING THE FORWARD/ROUTING OF THE CALL (B) 66
FIGURE 31: FEATURE TO WHICH C SUBSCRIBES 66
FIGURE 32: CFBL (B) SC SUB VECTOR 70
FIGURE 33: CND (C) SC SUB VECTOR 70
FIGURE 34: SKETCH OF MULTI-LEG CALL MODEL 73
FIGURE 35: SEQUENCE 84
FIGURE 36: OR-FORK 84
FIGURE 37: GUARDED BEHAVIOUR 85
FIGURE 38: AND-FORK 86
FIGURE 39: OR-JOIN 86

FIGURE 40: AND-JOIN 87
FIGURE 41: FEATURE INTERACTION DETECTION CYCLE 95
FIGURE 42: GRAPHICAL REPRESENTATION OF THE TOP-LEVEL OF THE SPECIFICATION 113
FIGURE 43: FI DETECTOR PROCEDURE 116
FIGURE 44: SYSTEM WITH A THIRD PARTY AS A FEATURE SUBSCRIBER 118
FIGURE 45: EXAMPLE OF A SCENARIO INVOLVING TCS AND CFBL 119

LIST OF TABLES

TABLE 1: TRANSLATION GUIDELINES FROM CHISEL TO UCM 48
TABLE 2: DISTRIBUTION OF THE FEATURES BETWEEN USERS 67
TABLE 3: FILTERING RESULTS 71
TABLE 4: THE MAIN ABSTRACT DATA TYPES 88
TABLE 5: FEATURE INTERACTION DETECTION RESULTS 125
TABLE 6: THE MAPPING TABLE OF FI TYPES 126
TABLE 7: COMPARISON IN TERMS OF NUMBER OF FEATURES INTERACTIONS 128

Chapter 1

Introduction:
Motivation and Background

1.1 Introduction

A hallmark of the introduction of digital technologies into the telephone network is the extension of the basic call service through switch-based software. Instead of only being able to carry out a basic conversation, where one party calls another and carries on a conversation until one party hangs up, a variety of different behaviours can now be supported during the course of a call.

Instead of only allowing a conversation between two parties, new behaviours such as forwarding calls, placing callers on hold, and blocking calls are realized through the use of telephony *features*. A single feature is defined as an extension/a modification of the basic service [10]. These features operate according to protocol rules defined between the users of the system and the network.

Due to market demands and high competition, the rapid development and deployment of features has become an important goal for telecom companies.

Different features of a system may be designed by different designers at different times. Because features are designed independently and validated in isolation, it is possible that their behaviour changes in certain feature combinations. When this occurs, it is said that there is "Feature Interaction" [67]. Extensive system testing is used during development to help ensure that features will function together properly when combined.

Feature interactions impact all phases of the software lifecycle. Timely mechanisms for resolving untoward feature interactions at all stages of the software lifecycle must be part of any new service development process.

Over the past several years, a large number of techniques to avoid, detect and solve feature interactions have been proposed in order to reduce the need for testing and therefore the time it takes to get new features to market [48][50][23]. These techniques act at all stages of the feature life cycle going from the requirement stage to the implementation stage.

1.2 The Feature Interaction problem

It is difficult to give a precise and complete definition of the "Feature Interaction" because of the huge variety of problems that could be classified under this heading. Meanwhile many formulations of the problem have been proposed and much research has been done in this area [63] [61] [53].

As a non-formal definition, we use the definition of [53] stating that there is "feature interaction":

1. When a feature inhibits or subverts the expected behaviour of another feature (or another instance of the same feature).
2. When the joint execution of two features provokes a supplementary phenomenon that cannot occur during the processing of each of the features considered separately.

Note: This definition is not formal enough. We will give a more formal definition in Chapter 6.

Zygan-Maus [63] distinguishes two levels of challenges:

1. The service level challenges: Such challenges are independent of how the involved services are implemented. This means that they will persist independent of whether services are implemented in a switch, on an IN platform or in TINA (Telecommunication Information Networking Architecture) [49]. Service level Interactions are purely logical interactions.

2. The technical level challenges: Such challenges are dependent on implementation. When a new feature is being designed, feature functionality has to be mapped on network architecture and linked to the basic call processing. The feature implementation has to provide the functionality required of each particular network element in support of the new feature and has to minimize interaction prone impacts on existing feature implementations. At this level, feature designers should take into consideration the real network constraints such as network signalling restrictions, charging restrictions, network timing conflict, concurrent resources usage attempts...etc.

In this thesis we focus on the service level interactions. We wish to detect interactions when features are specified (at the design and integration level) before the implementation.

1.2.1 Example of feature interaction

Figure 1 illustrates an instance of such problem between two common features, namely Originating Call Screening (OCS) and Call Forwarding Busy Line (CFBL). OCS forbids the establishment of a call to phone numbers on a screening list, while CF forwards incoming calls to another phone number. The two features interact inappropriately when A, whose OCS screening list includes C, calls B. Since B, who subscribes to CFBL to C, is busy, the incoming call is forwarded to C.

However, C was on A's screening list, and therefore the connection should not have happened, as it violates assumptions related to OCS. So CFBL has inhibited the expected behaviour of OCS.

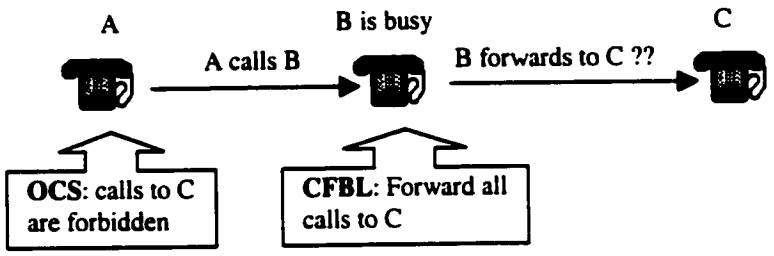


Figure 1: Illustration of a telephony feature Interaction

Note: Feature interaction is necessary and inevitable in a feature-oriented specification, because so little can be accomplished by features that are completely independent. In fact, some interactions are expected by designers. However, which are and which are not is a matter of designer's judgment and outside of the scope of this thesis.

1.2.2 Addressing feature Interaction

The feature interaction problem can be addressed according to three approaches: Avoidance, Detection and Resolution [43].

- Avoidance:

  The objective of an avoidance approach is to define additional guidelines (constraints, service platform and service environments) to prevent the manifestation of unwanted interactions. This assumes that the causes of the interactions are known, which is not always the case.

  This approach can be adopted starting from the early phases of specification and design of features. An example of use of avoidance approach is the Wireless Intelligent Network (WIN) system, where the feature interaction problem is solved, at least in part, by giving pre-defined priorities to different features [10].

- Detection:

  Approaches for detection aim to determine whether or not a set of independently specified features can cause conflicts when they are composed. The detection analysis can be applied through the whole lifecycle of a feature [23][48][50], since the cause of interaction can be related to any phase of the feature lifecycle.

- Resolution:

The objective of a resolution approach is to find solutions to interactions when they occur [11][50][60][68]. But before trying to solve the interactions that may occur between features, an accurate analysis of the undesired behaviour should be performed. Proposed techniques solve the interaction by:

- Replacing the undesired behaviour by a reasonable one [11].

  Example: Consider the example of detecting the interactions between features CW (Call Waiting) and 3WC (Three Way Calling). We suppose that users A and B are in a phone conversation.

  Feature CW: If C calls A, the latter is informed by a CW-tone. A may generate a flash-hook signal to put B on hold and to be connected to C. A may switch between B and C with flash-hook signal.

  Feature 3WC: A can flash the hook to put B on hold, and then A can call C. While A and C are in a phone conversation and B is on hold A can flash the hook a second time to add B in the conversation. After that, if C hangs up or if A flashes the hook, then A and B are in a normal conversation.

  Interaction: Let us assume that C calls A and then the latter flashes the hook just before being informed by the CW-tone. The flash-hook signal must be interpreted by CW or by 3WC? In both cases, B is put on hold but A is connected to C only in the first case. We obtain undesirable state in the joint behaviour of CW and 3WC. This undesirable state consist on a non-deterministic state from which event "flash A" may lead to different states.

  Resolution: By removing this undesirable state and generating a behaviour corresponding to a mutual exclusion between these two features the interaction is resolved.

- Build a negotiation protocol between the features involved in the interaction. This protocol consists of an exchange of necessary information to avoid the interaction [50][60][68].

Detection/Resolution during the service creation process are known as off-line detection/resolution while detection/resolution at run time are known as on-line detection/resolution.

In this thesis, the method proposed in Chapter 6 to detect feature interaction is an off-line process.

1.3 Contribution of the Thesis

The major contributions of this thesis is the development of a model for describing telephony features and of a method for detecting feature interactions at two stages: the requirements stage and the specification stage.

1.3.1 Contribution 1: Feature Interaction Filtering at the Requirements Stage

In Chapter 4, we present a model for describing Telephony features at the requirements stage. This model is built using the Use Case Maps Notation and allows us to describe many features. Based on this model, we propose a method to filter feature interactions at the requirements stage. This method offers a quick and rough evaluation of possible feature interactions before the Feature Interaction Detection Process (Proposed in Chapter 6). This preliminary evaluation allows the detection process to focus on feature combinations where interactions are possible and therefore reduces the cost of the detection process. Concrete examples are given in Chapter 4.

1.3.2 Contribution 2: Feature Interaction Detection Method

In Chapter 6, a Feature Interaction Detection System is developed for detecting feature interactions. This method aims to detect interactions occurring at the abstract specification level and resulting in violation of system integrity. This technique in based on the Formal Description technique LOTOS and uses Abstract Data Types to detect those violations. Combining features manually in the basic system could avoid some existing interactions and create new virtual interactions. Because of this, we chose to describe Features independently in our Feature Interaction Detection System. The Feature Interaction Detection System consists of three parts:

two feature specifications and a process called Feature Interaction Detector (FI Detector) for detecting the interactions.

Our method detects feature interaction by executing the feature Interaction detection system specification. The designer can reach those interaction points either by a step by step execution or using the goal oriented execution technique.

1.4 Organization of the Thesis

The six remaining chapters will cover the following issues:

Chapter 2: Related work: Feature Interaction Detection Requirement and Specification stages

We present a survey of related work on Feature Interaction Detection at both requirement and specification stages. We also introduce the techniques used to do such Feature Interaction Detection.

Chapter 3: Describing Requirements with Use Case Maps

In this chapter we present some existing requirement description techniques and we focus on the Use Case Maps notation that we are going to use in our work.

Chapter 4: Feature Interaction Filtering at Requirement Stage

First we introduce a model for describing Telephony features at the requirements stage. Based on this model, we propose a method to filter feature interactions at the same stage. Finally we present the results of the filtering on switch and IN based features.

Chapter 5: Specifying Features using LOTOS

In this chapter, we give an overview of the LOTOS specification language and its main operators. Our main objective in specifying the system model and features in LOTOS is to provide a specification that can be used for validating and detecting feature interactions.

Chapter 6: Feature Interaction Detection Method

We describe an improved formal definition of Feature Interaction and a Feature Interaction Detection System. Our technique aims to detect violation of system properties at the design stage, based on the Formal description Techniques (FDT) LOTOS and uses Abstract Data Types (ADT) to detect these violations.

Chapter 7: Conclusion and Future Work

Conclusion and future work are presented in this chapter.

Chapter 2

Related work: Feature Interaction Detection at Requirement and Specification Stages

We present a survey of related work on Feature Interaction Detection on telephony systems at both requirement and specification stages.

Feature Interaction is a research area of some importance, and a number of papers are published every year on the subject. Six International Workshops have been held so far [1][2][3][4][5][6]. In this survey, we present only work that is in some way related to our approach.

2.1 Addressing Feature Interactions in the Requirements Stage

Requirement description plays an important role in the development of telecommunication systems. Early conflict detection can help prevent costly and time-consuming problem fixes during implementation [14][60].

- A. Gammelgaard and J. E. Kristensen [10] propose to let feature specifications be restrictions to the class of deterministic labelled transition system. A features specification consists of two parts:
  1. Network properties: they are formulas expressing static constraints. All states in a system must satisfy all given network properties.
  2. Declarative transition rules: Dynamic constraints are formalized by declarative transition rules. Such rules consist of a precondition, a trigger event, and a post-condition. If a state satisfies the precondition, then this state must be origin of a transition labelled by the trigger event, and the resulting state of the transition must satisfy the postcondition.

Network properties and the pre- and postconditions of declarative transition rules are formulated in a simple logic, which is a restriction of ordinary first order logic. A new feature can be added by replacing some rules of the core service by new rules and introducing new formulas.

Executing those rules by matching the postconditions with the preconditions allows finding inconsistencies. An interaction is detected when the system arrives to a state that doesn't satisfy the defined network properties.

By adding new features to the system, new rules and new predicates are defined. The matching process of postconditions and preconditions fails if there are missing rules. In our experience, we found that the detection process becomes very difficult if the set of rules describing the features is not complete.

- In [60] Buhr and al use the UCM (Use Case Maps) notation (see Section 3.2.2 for detailed introduction to UCM) to describe Telephony features. The proposed method generates tables from UCM behaviors and provides a framework for humans to add information that will enable executable prototypes to be generated. Features are modeled as competing rule engines and interactions are detected and resolved at run time by coordinating through a blackboard.

- Heisel and Souquieres [48] propose a method of *requirements elicitation* in order to detect feature interactions. The proposed approach is inspired by object oriented methods and gives guidance on how to identify, express, and systematically transform requirements into a formal specification. The proposed method uses *agendas* (a list of steps to be performed when carrying out some tasks in the context of software engineering) and first order logic to express and incorporate constraints into the system requirements. Two constraints are said "*Interaction Candidates for one another*" if they have common preconditions but incompatible postconditions. After expressing those constraints, an analysis of preconditions and postconditions determine which candidates could lead to an interaction. The incompatibility of postconditions takes place either in the state immediately following the state that is referred to by the postcondition or in a later state. Although this approach uses a simple logic like the one used by the previous approach, the interactions are expressed differently.

- Jonsson et al [17] propose a technique for hierarchically structuring requirements. Requirements are described as a structured hierarchy of predicates, capturing system properties, and shielded from concrete details of system implementation by one or several levels of abstract logical concepts. The set of abstract concepts, which represents a collection of abstract predicates, can be seen as forming a *vocabulary* of the application domain. This method uses a simple linear-time temporal logic using the defined predicates. Each requirement can be represented as an observer. These observers are added to the system, which is then subject to exhaustive state-space exploration. The exploration will detect when a requirement is violated, and as a side effect shows the sequence of events leading to the violation. Our method (Described in Chapter 4) employs Use Case Maps notation to describe requirements. UCM's elements are used to structure these requirements at different levels of abstraction. Then a filtering procedure is conducted to detect interaction.

- In a paper co-authored by the author of this thesis and Nakamura et al [50] propose a feature interaction filtering method at the requirement level. The method extensively utilizes the requirement notation method Use Case Maps (UCMs), which helps designers to visualize a

global picture of call scenarios. In this framework, the addition of a feature is achieved by using the stub plug-in concept of UCMs. That is, a set of sub UCMs describing the feature's functionality are plugged into stubs of the basic call scenario in a *"root"* UCM. Thus, each feature is characterized by the stub configuration. The method proposes a pair wise composition of the features and gives one of the following verdicts: (a) FI occurs, (b) FI never occurs, (c) FI-prone.

In our work we use a similar model to describe features and to conduct the filtering. The relationship between our method and the method of [50] will be discussed in section 4.7.

- Aho et al [11] describe a language called Chisel for defining requirements. The authors claim that this language is unambiguous, applies to a variety of network technologies, and it has a sound basis for translation to commonly used formal software specification languages. Telephony features are then described in the Chisel language using the editing tool SCF3/Sculptor. Finally feature interaction detection is conducted based on an analysis of sequences of events for each feature. In our work we use features described in the Chisel notation to derive feature's Use Case Maps.

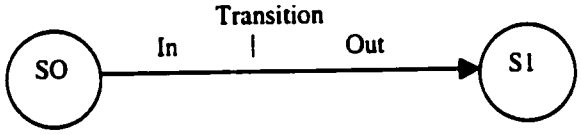
2.2 Formal Specification Methodologies for Telephony Systems

A formal description is a symbolic representation of a certain object in a given language. The language may use various kinds of symbols such as textual or graphical symbols. The description language uses strict rules for the construction of language expressions, the "formal syntax", and strict rules for the interpretation of well-formed language expressions, the "formal semantics". The main purpose of a formal description is to have unambiguous, clear and precise specification [64].

Various techniques have been developed for specifying telephony systems in a formal way. The main ones are: Finite State Machine Model (FSM), Petri Nets, LOTOS (Process Algebra) and SDL (Extended Finite State Machine EFSM).

2.2.1 Finite State Machines

Probably the earliest formal methods have been the ones based on the use of Finite State Machines "FSM" [30]. FSM is a transition model, where the behaviour of a given system is represented in terms of states and transitions. Each FSM is normally represented by a directed graph as outlined in figure 2 where a directed path represents an occurrence of an event which changes the state of the represented machine. The machine is in state S0 and when the event "In" happens, it changes state to S1 and the event "Out" happens.



Where S0 and S1 are states; "In" is an incoming event; and "Out" is an outgoing event

Figure 2: Finite State Machine

A state describes the current situation of the system, resulting from previous transitions, and at the same time it describes which transitions are possible in the future of the system. For example, in a telephony system, a user could be in *Dialing* state, in *Ringing* state, or in *Talking* state.

Telephony features described using FSM can be found in the Feature Interaction Contest 2000 [6][37].

The main shortcomings of the FSM model are:

1. The lack of a data model.
2. The lack of an architectural model
3. The lack of explicit representation of concurrency

In order to address these shortcomings, other models were defined, as described below.

2.2.2 Petri Nets

Petri Nets is a formal and graphically appealing language that is appropriate for modeling systems with concurrency. The Petri Net notation has been under development since the beginning of the sixties, where Carl Adam Petri defined the language [64]. It was the first time

that a general theory for discrete parallel systems was formulated. The language is a generalization of automata theory, making it possible to express the concept of concurrently occurring events.

Petri Nets are abstract machines that are used to describe the behaviour of systems. They are represented by directed graphs containing two types of elements: places and transitions. Places, which contain tokens, are represented by circles; transitions, which allow tokens to move between places are represented by lines. An example of a simple Petri Net is shown in figure 3.

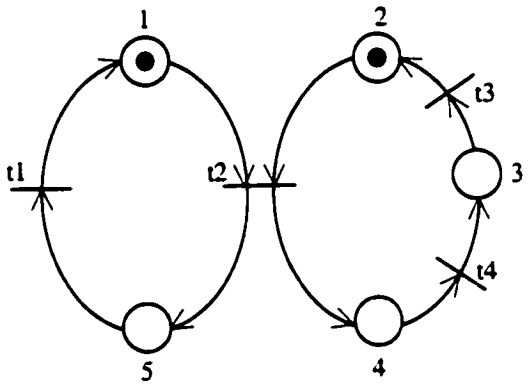


Figure 3: A simple Petri Net

When all the input places to a transition contain at least one token, the transition is enabled and may fire. When the transition fires, one token is removed from each input place and one token is added to each output place.

Yoeli and Barzilai [51] introduce the concept of extended Petri Nets (EPN) and use it to model the call processing operations in an automatic telephone exchange.

Two common problems with the FSM and Petri Nets are [38]: 1) the limited role assigned to data. Many features rely on data values and data structures for essential aspects of their functionalities, and so they are difficult to represent by Petri Nets. 2) The lack of process structure, which is very useful for design. Extended Finite State Machine (EFSM) methods such as SDL, remedy this situation.

2.2.3 LOTOS

We give an overview of the LOTOS specification language and its main operators in Chapter 5.

[44] Describes the Plain Old Telephone System (POTS) using LOTOS. Four different structural approaches could be adopted for specifying a telephone system:

- The resource-oriented style: In the resource-oriented style, the specification structure shows the architectural components of the design [45]. In [35], a formal specification of an IN network model was developed using the resource-oriented style.
- The state-oriented style: In the state oriented style there is an explicit reference to system states.

The resource-oriented style and the state-oriented style are implementation-oriented and suggest an implementation architecture.

- The constraint-oriented style: In the constraint-oriented style, one focuses on the composition of the requirements, expressed as behaviours [45].
- The monolithic style: In the Monolithic style, the specification is described as a tree of alternatives, i.e. expanded execution sequences are explicitly enumerated

The constraint-oriented style and the monolithic style are requirement-oriented.

The work presented in [21] and in [38] describes a new approach for specifying telephone systems using a mixture of the constraint-oriented style and the state-oriented style.

2.2.4 SDL

The Specification and Description Language (SDL) is an object-oriented, formal language defined by The International Telecommunications Union–Telecommunications Standardization Sector (ITU–T) (formerly Comité Consultatif International Telegraphique et Telephonique [CCITT]) as recommendation Z.100 [57]. The language is intended for the

specification of complex, event-driven, real-time, and interactive applications involving many concurrent activities that communicate using discrete signals.

Just as the other formal languages, SDL covers different levels of abstraction, from a broad overview down to detailed design.

The basic theoretical model of an SDL system consists of a set of extended finite state machines (EFSMs) that run concurrently. These machines are independent of each other and communicate by means of discrete signals. SDL does not use any global data. It has two basic communication mechanisms: asynchronous signals (and optional signal parameters) and synchronous remote procedure calls. Both mechanisms can carry parameters to interchange and synchronize information between SDL processes and their environment.

Examples of specifying telephone system using SDL are presented in [53][34].

2.3 Addressing Feature Interaction at Specification Level using LOTOS

In this section we limit ourselves to reviewing work closely related to ours. We give a brief overview of the Feature Interaction detection methods using LOTOS.

- Boumezbeur and Logrippo [56] applied *the step by step execution* on the LOTOS specification of a telephone system. At each step of the step by step execution, the user chooses the next step to be taken among all possible actions that are offered at that point. This is useful for checking the conformance of a system defined informally to its formal description in LOTOS. In practice, this can be done by checking if test sequences that should be allowed according to the informal definition are also accepted by the formal specification; or checking if the test sequences obtained by executing the specification are included in the formal definition of the system; or by checking if test sequences that are not specified informally are not accepted by the formal specification. It is, however, a slow method and nowadays it is not used in practice.

- Stepien and Logrippo [19] proposed a method called "Backward Reasoning" to explore all the potential alternatives leading to feature interaction. The method involves specification of telephony features in LOTOS. Interactions to be detected are caused by ambiguity of actions. An observable action in a LOTOS specification is ambiguous if in the behaviour tree of the specification there is a branching point where the action is the first observable one in at least two branches. Ambiguity represents non-deterministic behaviour of the system being specified. To prove that an action is ambiguous, backward reasoning for LOTOS is applied. It consists of a combination of forward and backward execution. Forward execution of the specification is applied to reach the action, then, using the resulting behaviour expression, backward execution is performed to find a different trace leading to the same action.

- In [20] a method for representing and verifying intentions in telephony features using abstract data types is presented. Feature intentions describe the intended behaviour of telephony features. The first step of the method is to specify a feature's intentions using abstract data types. An intention is represented as an operation of Boolean result indicating whether a given combination of the basic data involved in an operation is allowed:

  Intention: Fid, partyRole, operation, Restriction_set -> Bool

  For example the origination call screening (OCS) prohibits any connection with a number that is in the screening list. This feature intention can be formulated as:

  Intention (Focs, called (N), connect, L) = N NotIn L;

  Where **Focs** identifies the feature originating call screening, **N** the phone number involved in a called role in operation **connect** and **L** is the restriction set that in this case is a screening list, **N NotIn L** is a Boolean expression that verifies whether the number N is in the screening list or not.

  Intentions of a feature are described independently of other features without consideration of potential interactions at this stage. They are described for every operation that exists in the system regardless of which feature is actually used, and

are expressed as Abstract Data Types operations which specify the intention's violation. The specification language considered is LOTOS. The second step consists in executing the formal specification of the system with features. The abstract data types descriptions of feature intentions are included in the specification, and a monitor for verifying intentions of features described as LOTOS processes is introduced to verify the intentions as described in the abstract data types every time an action of the specification is executed.

Our FI Detection method presented in Chapter 6 is inspired by this idea of representing intentions in Telephony Features using Abstract Data Types.

- In [45], Faci and Logrippo developed a methodology for detecting feature interactions using LOTOS testing theory. First, they defined two notions of composition and integration of features. Composition expresses the synchronization of features on their common actions with POTS and their interleaving on their independent actions. Integration expresses the extension of POTS with features, such that each feature is able to execute all of its actions that are allowed in the context of POTS. Then, they reason about interactions in terms of the conformance relation studied in LOTOS testing theory, in the following way: an interaction exists between features if their integration does not conform to their composition.

- In [35], Kamoun and Logrippo developed a method for detecting feature interactions between IN services using the Goal Oriented method (Goal Oriented method is described in Chapter 6). The method detects interactions resulting in violation of features properties. It is based on formalization of feature's properties, derivation of goals satisfying the negation of these properties, and use of Goal Oriented Execution to detect traces satisfying these goals. A trace satisfying a goal shows that an interaction exists between the specified features by describing a scenario violating one of the properties of the introduced features.

Our Method uses also the Goal Oriented Execution method. We simplified the goal to be just a "VR" event (VR: Violation Report). We also use a global observer process called Feature Interaction Detector (FIDetector) to capture interactions regarding the following four issues: Connections, Billing, Signals and Display. This will be described in Chapter 6.

Chapter 3

Describing Requirements With Use Case Maps

In this chapter we present the different existing requirement description techniques and we focus on the Use Case Maps notation that we are going to use in our work.

3.1 Introduction

Emerging telecommunications services and features require standardization bodies (ANSI, ETSI, ISO, ITU, TIA, IETF, etc.) to describe and design increasingly complex functionalities, architectures, and protocols [24]. In the early stages of the design process, many features, services, and functionalities are described using informal operational descriptions, tables and visual notations such as Message Sequence Charts (MSCs) [33]. As these descriptions evolve, they quickly become error prone and difficult to manage. The need of precisely documenting all stages of the design process, which is very important in the industrial environment, becomes critical in the standardization process [25].

Following the practice in several standard groups, the development of each phase of a telecommunication standard is divided in three stages [25] (shown in figure 4): 1) service descriptions, 2) message sequence information, and 3) protocol and procedure specification.

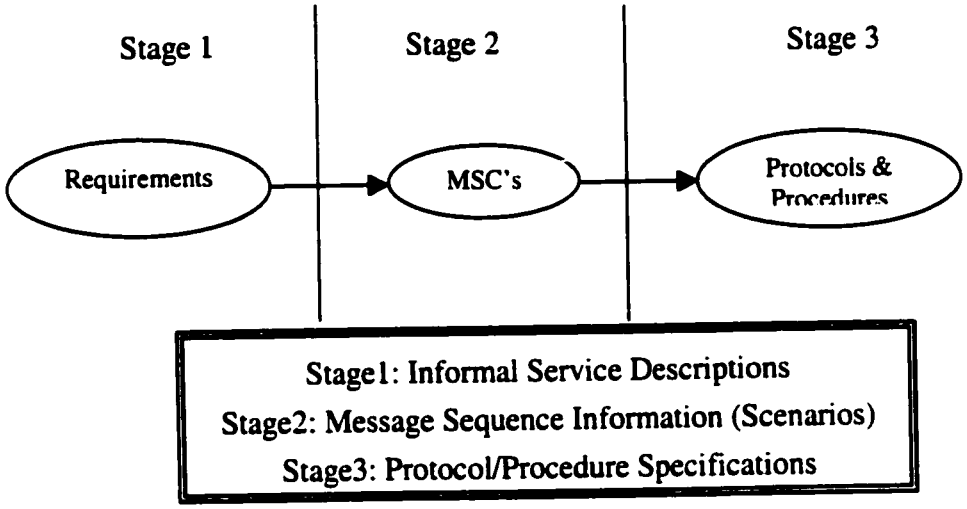


Figure 4: Three Stages Methodology

- In stage one, a description of a service should be given from the user's perspective. This stage describes what the service is supposed to do, not how it will do it.
- Stage two describes the capabilities and processes within the network. This is achieved by using sequences of messages between the different involved entities.
- The final stage produces the protocol specification.

3.2 Stage1: Requirement Description

During the past few years, a lot of research has been done in the area of Requirement description. Many models (Prototype model, Use Case Model, Organized by Roles Model, Organized by Classification) have been proposed for capturing the user requirements. The most commonly used model is use cases. The objective of this model is to capture the functional requirements from the user point of view. There are several reasons why use cases have become popular and universally adopted. According to [32] the two major reasons are: 1) they offer

systematic and intuitive means of capturing functional requirements. 2) They drive the whole development process since most activities such as analysis, design, and test are performed starting from use cases.

Scenario-based approaches are now widely used in industry for the design of distributed systems. One of the main reasons is that scenarios describe top-level critical requirements that need to be fulfilled by the detailed design, and thereafter by implementation.

The following introduces the readers to the requirement description techniques that are used in this thesis: Chisel Diagrams and Use Case Maps. However the bulk of this chapter will consist of an introduction to Use Case Maps, which are extensively used in this thesis.

3.2.1 Chisel Diagram Notation

Chisel diagrams are a scenario-based approach that is used to describe requirements for communications services and features. The language Chisel is intended to reflect current practice for writing these requirements. This language could be applied to a variety of network technologies, and it has a sound basis for translation to commonly used formal software specification languages [7].

For illustration, a basic two-party POTS (Plain Old Telephony System) Chisel diagram is given in figure 5.

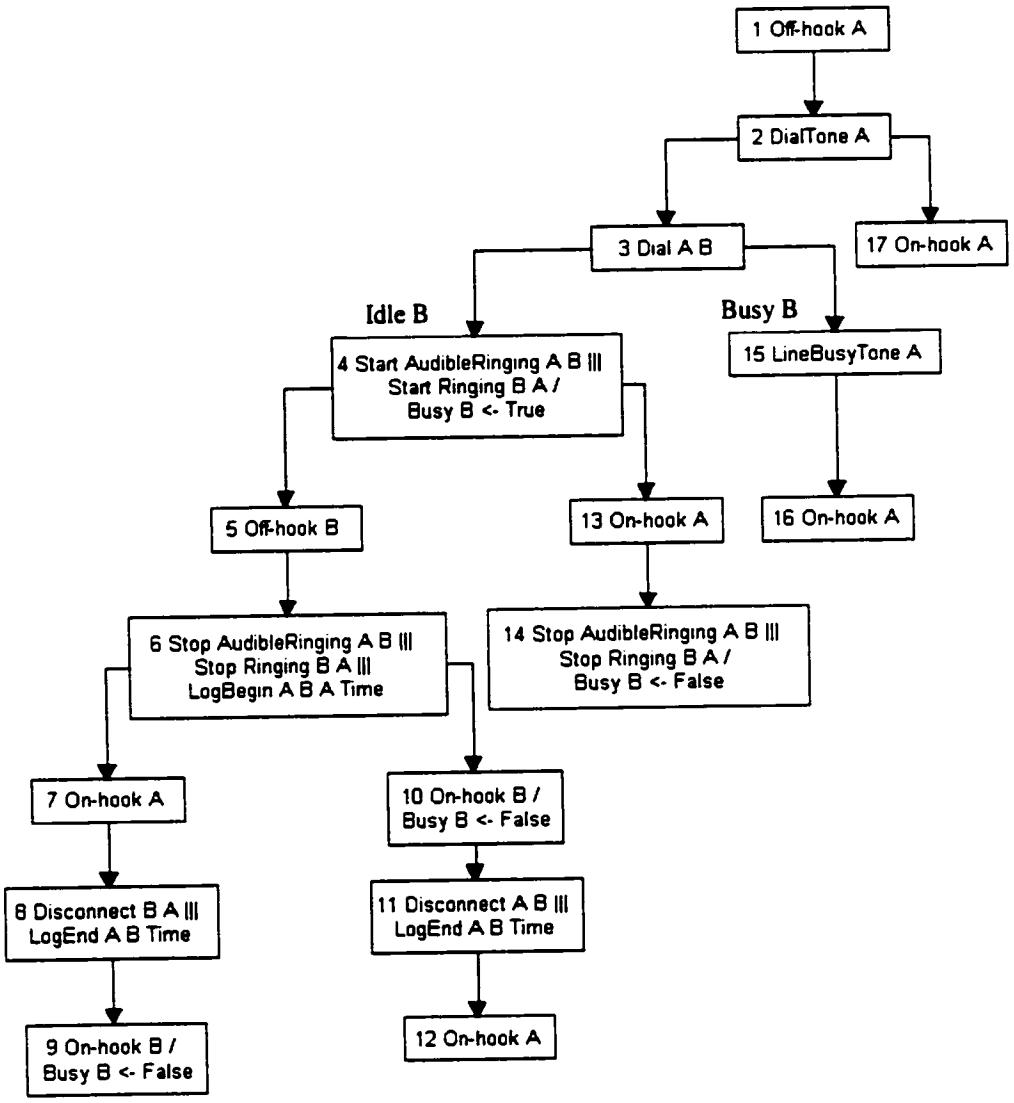


Figure 5: Chisel Diagram for POTS

This Chisel diagram includes both telephones in a two-party call, and also some messages for the billing system.

A node (one of the boxes) in a chisel diagram contains a number, which uniquely identifies the node within the feature, and one or more events and variable assignments. The

nodes are connected by directed edges (arrows in the diagrams). Multiple events in a node are separated by vertical bars (|||). A node containing multiple such events is equivalent to any possible sequence of those same events (i.e., A ||| B means {AB or BA}; A ||| B ||| C means {ABC or ACB or BAC or BCA or CAB or CBA}; and so forth).

This is a description of the main events involved in POTS:

- *Dial A B* means that the subscriber at address A dials the address B.
- *DialTone A* means that dial tone occurs at address A.
- *Start Ringing A B* means that alerting starts at address A for a call originated at address B.
- *Start AudibleRinging A B* means that the ring back tone is provided at address A while waiting for the user at address B to answer the call.
- *Stop Ringing A B* and *Stop AudibleRinging A B* mean to stop the ringing or tone occurring at address A in relation to a call to or from B.
- *LineBusyTone A* means that the telephone to which A is attempting a connection is busy.
- *Disconnect A B* informs A that B has disconnected a connection with A. (It is a signal from the switch to a user, signalling the user that a connected party has gone *On-hook*. The *On-hook* event is the signal from the user to the switch that the user is disconnecting)

Variables are used in conditions on edges, to define when an edge can be followed in constructing an event sequence from the diagram and to restrict possible interleavings of event sequences. A variable defines one or more sequences of events. For instance *Busy B* (Busy B <- True in figure 5, node 4) defines the set of event sequences having one of the following properties:

- An event sequence containing an *Off-hook B* not followed by *On-hook B*.
- An event sequence containing *Ringing B* not followed by *Disconnect B A*.

Note: All of the POTS event sequences start and end with Busy A = False (Idle A = True).

To define the value of a variable after an event an assignment statement can be included with the event to say that the variable takes on a new value after the event. The format of this is: <event> / <var> ◄——— <value>.

Note: Value changes are shown in nodes (nodes 4,9,10 and 14 in figure 5).

A condition next to an edge means that to continue an event sequence by following that edge, the condition must be true at the end of the event sequence. C syntax is used in the conditions (~ for not, && for and, || for or).

3.2.2 Use Case Maps

3.2.2.1 Introduction

Like Chisel Diagrams, Use Case Maps (UCM) is a Scenario-based approach. It is a visual notation for representing use cases. It has been proposed by Buhr and Casselman [58].

The UCM notation is used to describe scenario paths in terms of causal relationships between responsibilities. UCM paths are wiggly lines that enable a person to visualize scenarios threading through a system without the scenarios actually being specified in any detailed way.

The notation is intended to be useful for requirement specification, design, testing, maintenance, adaptation, and evolution [69]. Already, UCMs have been used in a number of areas [69]:

- Requirements engineering and design of:
  - Real-time systems
  - Object-oriented systems
  - Telecommunication systems
  - Distributed systems
- Detection and avoidance of undesirable feature interactions
- Evaluation of architectural alternatives
- Functional testing

- Documentation of standards

UCMs have raised a lot of interest in the software community, which led to the creation of a user group at the beginning of 1999, with more than one hundred members from all continents [69]. Currently, UCM standardization is underway within ITU-T.

3.2.2.2 UCM Notation Elements

In this section, we introduce the UCM notation. We limit ourselves here to the UCM elements that we are going to use in this thesis. Use Case Maps for Object-Oriented Systems [58] and Use Case Maps as Architectural Entities for complex systems [59 represent more complete tutorials on the UCM notation.

The core notation consists of only scenario paths and responsibilities along the paths. The basic path notation addresses simple operators for causally linking responsibilities in sequence, as alternatives, and in parallel.

A UCM path may have any shape as long as it is continuous. It starts at a starting point (depicted by a filled circle) and ends at an end point (shown as a bar). Between the start and end points, the scenario path may perform some responsibilities along the path, which are depicted by crosses × with labels. Responsibilities are abstract activities that can be refined in terms of functions, tasks, procedures, events, and are identified only by their labels. Tracing a path from start to end is to represent a scenario as a causal sequence of events.
Note: Start points may have preconditions or triggering events attached, while responsibilities and end points can have post-conditions.

Figure 6 illustrates the basic elements of UCMs.

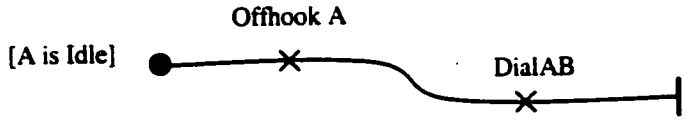


Figure 6: UCM Basic path

The responsibilities can be bound to components, which are the entities or objects composing the system. Figure 7 illustrates an UCM with three components: phone1, phone2, and a switch. We use the connection phase of a simplified telephone system in this figure because it is easy to understand. An initiator (phone A) tries to establish a connection with a Responder (phone B) via a simple switch. The scenario starts with a responsibility "OffHook A" where user A picks up the phone. This is the first activity that initiates the connection. Then "DialAB" is performed where user A dials the phone number of user B. We have now two alternatives (the OR-Fork is describes below) each one of which is associated with a pre-condition. For example user A receives a busy tone when the precondition is [B is busy].

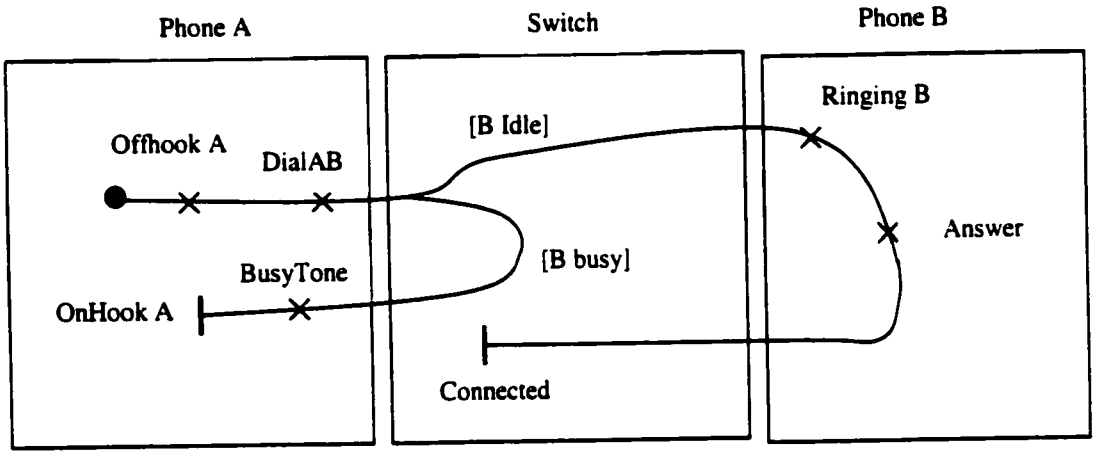


Figure 7: Bound UCM

Several paths can be composed by superimposing common parts and introducing forks and joins. There are two kinds of forks and Joins.

1. OR-Fork/Join: Depicted by branches on paths. They describe alternative scenario paths, which mean that one of the paths is selected to proceed at each branch.

- OR-Fork (Figure 8a): Splits a path into two (or more) alternatives. Alternatives may be guarded by conditions represented as labels between square brackets.
- OR-Join (Figure 8b): merges two (or more) overlapping paths.

2. AND-Fork/Join: Depicted by branches with bars, which describe concurrent scenario paths.

- AND-Fork (Figure 8c): Splits a path into two (or more) concurrent segments.
- AND-Join (Figure 8d): Synchronizes two (or more) paths together.

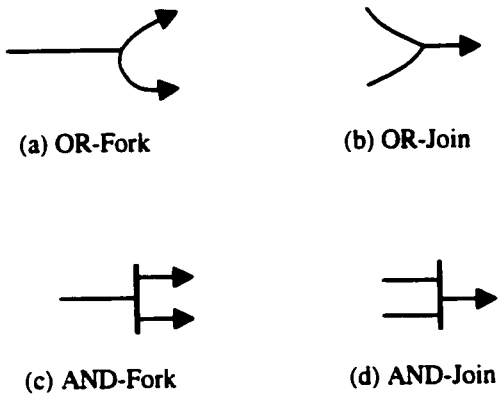


Figure 8: OR-Forks/Joins and AND-Forks/Joins

3.2.2.3 Advanced Notation Elements

More advanced operators can be used for structuring UCMs hierarchically and for representing exceptional scenarios and dynamic behaviour. When maps become too complex to be represented as one single UCM, a mechanism for defining and structuring sub-maps becomes necessary. A top-level UCM, referred to as root map, can include containers (called stubs) for sub-maps (called plug-ins). The stub plug-in concept allows UCMs to have a hierarchical path structure, to defer details, and to reuse the existing scenarios.
Stubs are of two kinds: Static Stubs and Dynamic stubs.

- **Static Stubs** (Figure 9): Represented as plain diamonds, they contain only one plug-in, hence enabling hierarchical decomposition of complex maps.

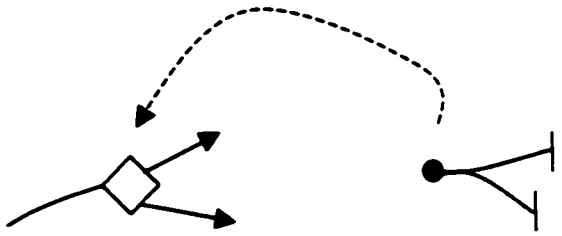


Figure 9: Static stubs have only one plug-in (sub-UCM)

- **Dynamic stubs** (Figure 10): represented as dashed diamonds, they may contain several plug-ins, whose selection can be determined at run-time according to a selection policy (often described with pre-conditions).

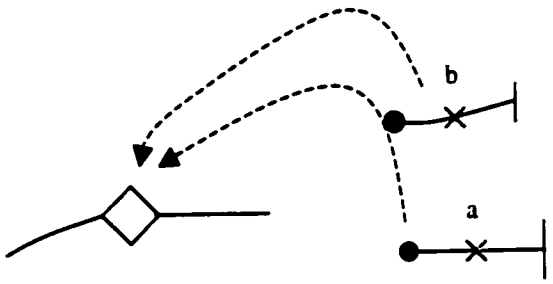


Figure 10: Dynamic stubs may have multiple plug-in

3.2.2.4 Philosophy of UCMs

The Use Case Maps notation aims to link behaviour and structure in an explicit and visual way. According to [59] UCM paths are first-class architectural entities that describe causal relationships between responsibilities, which are bound to underlying organizational structures of abstract components.

UCMs can be derived from informal requirements or from use cases if they are available. Responsibilities need to be stated or be inferred from these requirements. For illustration purpose, separate UCMs can be created for individual system functionalities, or even for individual scenarios. However, the strength of this notation mainly resides in the integration of scenarios.

It is important to clearly define the interface between the environment and the system under description. This interface will lead to the start points and end points of the UCMs paths, and it also corresponds to the messages exchanged between the system and its environment. These messages are further refined in models for detailed design (e.g. with Message sequence Charts, see Section 3.3).

3.2.2.5 UCM Tools

There currently exists only one tool that supports the UCM notation: The UCM Navigator (UCMNav). This tool is used for creation and maintenance of UCMs. UCMNav

ensures the syntactical correctness of the UCMs manipulated, generates XML descriptions, exports UCMs in Encapsulated Postscript or Maker Interchange Format (for Adobe Framemaker) formats, and generates reports in PostScript.

3.3 Stage2: Message Sequence Chart

Message Sequence Charts (MSC) [33] is a graphical and textual language for the description and specification of the interactions between system components. The main area of application for Message Sequence Charts is the specification of the communication behaviour of distributed systems, like telecommunication switching systems. Message Sequence Charts may be used for requirement specification, simulation and validation, test-case specification and documentation of real-time systems. MSCs are often used in combination with SDL (Specification Description Language)[57].

Figure 11 describes the interaction between 3 components: C1, C2 and C3 via exchanging messages a, b, c and d.

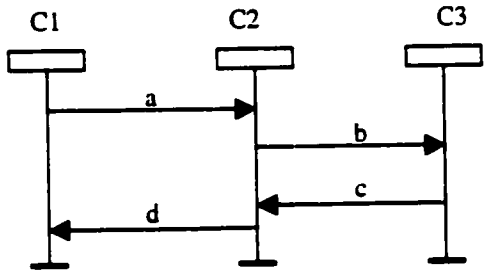


Figure 11: MSC Example

3.4 Benefits of UCMs and their relation with MSCs

These are some of Use Case Maps benefits:

- During early stages, UCM can be composed of paths where responsibilities are not allocated to any component. However, designers are likely to include architectural elements such as internal components. In this case the description of these components, their nature, and some relationships (e.g., components that include sub-

components) are required. Communication links between components are usually not required, but they can be added.

- UCMs do not specify anything about details such as data transfers along paths, local data values at points along paths, and local decisions based on local data values.
- Use Case Maps are used to describe and integrate use case representing the requirements. UCMs give us the big picture at a high level of abstraction with using hiding mechanism.
- UCMs are intended to bridge the gap between requirements (use cases) and detailed design (MSCs for example), since they are expressed above the level of messages exchanged between components. More than one MSC may be derived for a single UCM. In figure 12, the paths in the UCM show the causal sequence abc in an abstract manner. Two possible implementations of this UCM are shown in the form of two MSCs.

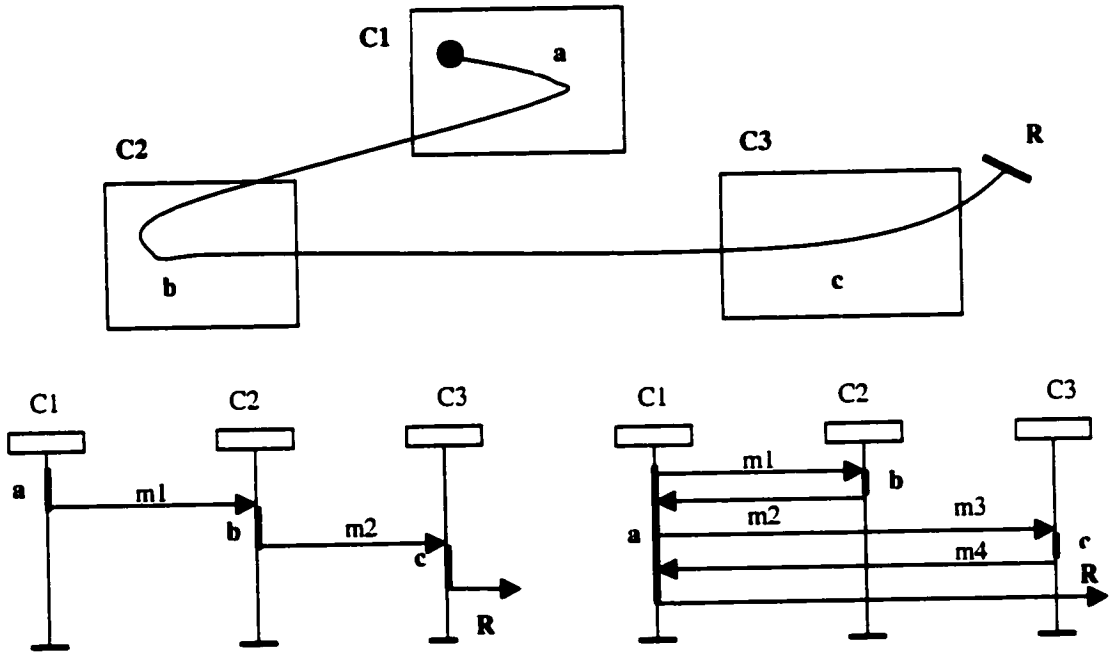


Figure 12: Causal Sequence of a UCM

- UCMs are not executable, but they can be manually translated to models that allow fast prototyping and validation. LOTOS, which will be introduced in the next chapter,

is well suited for representing UCMs. Translation and execution of UCMs will be discussed in detail in Chapter 5.

- Test suites can be generated directly from UCM. The test cases generated from UCMs can be executed against the specification in order to prove consistency.

Chapter 4

Feature Interaction Filtering at Requirement Stage

4.1 Motivation

Communication Protocols are rules that are followed for orderly communication between two or more communicating parties. They are needed in order to ensure that the total system formed by the individual parties and their interaction is meaningful to all the parties concerned and performs the functions required. The basic functions of a protocol usually include: Connection, Disconnection, Access control (for security purposes), Addressing, Error control, Flow control and Synchronization. Figure 13 provides an abstract view of the relationship between network and users.

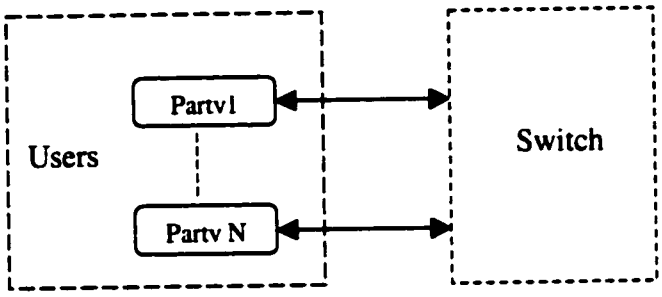


Figure 13: A reference model of a protocol system

The total communication protocol system is often divided into smaller ones, depending upon the stage that has been reached in the communication. For example, one protocol could be used to set up, prepare or establish communication, another could be used to ensure effective interchange of information after the connection has been established and a third protocol could be used to ensure the proper termination or closing down of the connection between the parties.

In practice this type of partitioning makes the total communication process easier to understand and enables modifications to various parts of the protocols to be made more simply and reliably.

4.2 Description of Services at Requirement stage

4.2.1 Service Decomposition

Intuitively, at least four main steps are needed to provide a service to an end user:

- Request the service
- Check for service availability and user authorization
- Provide the service
- Update the corresponding data and release the allocated resources

Each step may involve procedures to request resources, to set up the communication between network components, to access and update data...etc.

Starting from this simple intuitive decomposition we can decompose the Basic system service, the called Plain Old Telephone System (POTS), into four steps:

- ***Service request***: When a user wants to be served (e.g., wishes to make a call), the user indicates this desire to the network and receives an indication that this service can be provided. In the most familiar voice world this is accomplished through a user's action of taking a telephone off-hook and through a network's action of issuing a dial tone. The dial tone signals that the network is ready to provide a voice service, that is, that the access to the network has been granted.

- ***Checking the information***: The number dialed by the user could be invalid or out of service. In such a case an announcement will be played to the caller and a "busy-tone" signal is sent to him, otherwise the call procedure will continue.

- ***Provide the service***: This step is reached when the number dialled is valid and in service. Depending on the state of the called party, busy or idle, appropriate signals are sent to either parties and the service is provided. We consider that the busy tone signal sent to the caller, when the called party is busy, as a part of the service. During this step the billing process starts.

- ***Disconnection***: During this step the allocated network resources are released and the billing information is updated. Depending on who disconnects first appropriate signals are sent to either parties.

4.2.2 UCM Call Model

4.2.2.1 Introduction

Telephony features are usually complex and difficult to design and to implement. The specification of features written in a natural language (e.g. English) can be unclear or ambiguous and may be subject to different interpretation. As a result, independent implementations of the

same feature may be incompatible. There is therefore, a need for a notation to help designers to understand and analyze these features.

Our main examples in this thesis are based on the feature requirements defined in the International Feature Interaction Detection contest held on the occasion of the Fifth International Workshop on Feature Interactions in Telecommunications and Software Systems (FIW98)[5]. The contest defined POTS and 12 switch–based and IN features. The requirements are described using the state-based language "Chisel" (see Section 3.2.1). Each feature is described with an end-to-end point of view and the different actions are not bound to network entities. Due to the nature of the state-based method, it is difficult to represent concurrent behaviours and feature addition is achieved by "gluing nodes" in the Chisel diagram, which results in difficulty to achieve global visualization in one picture. To cope with this problem we employ the requirement notation UCM.

4.2.2.2 UCM service description

To model a service (transactional, telephony...etc.) we use the service decomposition idea defined in 4.2.1. We build a "root UCM" (or simply root map) that specifies the scenario path structure commonly used by all services.

Figure 14 illustrates the service model described using UCM where each step is modeled as a "stub" containing a UCM sub map.

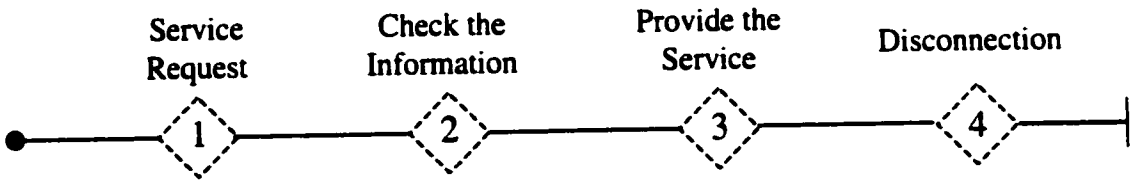


Figure 14: UCM service Model

4.2.2.3 UCM Call Model

The phone call model is derived from the UCM service model by defining the four UCM stubs corresponding to the four service stages.

The end points are not shown in the UCM service model. We choose to hide this information in the high level description and describe it at lower levels.

Figure 15 describes these four Stages for a phone call. This UCM is also called "Root Map".

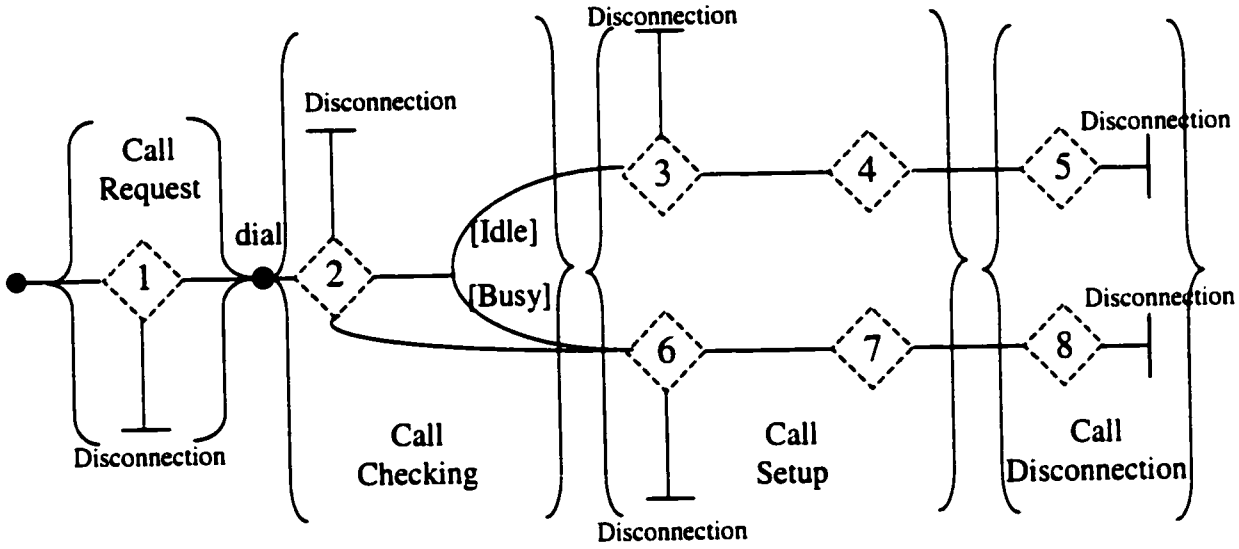


Figure 15: UCM Call Model (Root map)

Stub 1: Pre-Dial Stub
Stub 2: Post-Dial Stub
Stub 3: Idle Stub
Stub 4: Idle Setup Stub
Stub 5: Idle Disconnection Stub
Stub 6: Busy Stub
Stub 7: Busy Setup Stub
Stub 8: Busy Disconnection Stub

Note: A user can on-hook at any time before the disconnection stubs. In order to detect feature interactions we want to go as far as possible in the scenario, so we assume that no on-hook occurs before the disconnection stubs.

4.2.2.4 Stub Description

In our UCM model (Root map) we suppose that the user A is the caller, the user B is the called party (A will try to call B), C and D are parties introduced by some features.

❶ ***Call Request***

The Call request contains one stub:

- **Pre-Dial stub (Stub 1 in figure 15)**: During this step user A requests to be connected to another user. This stub contains the actions performed between "Off-hook A" (A is supposed Idle) and "dial AB" (If we suppose that A dials B's number).

❷ ***Checking the call information***

All the checking is performed within one stub: the Post dial stub.

- **Post-Dial stub (Stub 2 in figure 15)**: This stub contains the actions performed after "dial AB" action. The checking consists of looking for the relevant user information, which could be related to a subscribed feature.

  The call could be blocked if the subscriber doesn't meet the authorization rules defined by specific features. For example: When caller A is an OCS (Originating Call Screening) subscriber and B is in A's screening list, the caller will receive an announcement message telling him that the call is denied. So the call is blocked.

  Note: The path between stub 2 and stub 6 is followed when the destination state (B in this case) is not relevant (we don't care whether B is idle or busy).

❸ ***Call Setup***

In this step the service is provided to the involved users. These four stubs cover all the possible scenarios based on whether the destination is idle or busy and on the features the users are subscribed to.

- **Idle Stub (Stub 3 in figure 15):** Contains the actions performed when the destination (B) is Idle. These actions occur before the establishment of the communication. The attempt to establish the call is successful but the call is not set up yet.

- **Idle Setup Stub (Stub 4 in figure 15):** Contains the actions performed after the destination (B) goes "Off-hook" and before the disconnection process (when one of the two parties decides to end the communication).

- **Busy Stub (Stub 6 in figure 15):** Contains the actions performed when the destination (B) is Busy. These actions occur before the establishment of the communication if any.

- **Busy Setup Stub (Stub 7 in figure 15):** Contains the actions performed after the destination (example: a third party C) goes "Off-hook" and before the disconnection process (when one of the two parties decides to end the communication).

❹ ***Disconnection***

- **Idle Disconnection Stub (Stub 5 in figure 15):** Contains the actions performed when one of the two parties involved decides to end the call.

Note: The caller A can hang up before the establishment of the call. In this case the disconnection procedure is described within the previous stubs.

- **Busy Disconnection Stub (Stub 8 in figure 15):** Contains the actions performed when one of the two parties (example: A or the third party C) involved decides to end the call.

4.2.2.5 Plug-ins of POTS

Figure 16 represents UCMs plug-ins for the basic call model, (or POTS- Plain Old Telephony System), described based on the first FI detection contest specifications.
There are six UCMs plug-ins in figure 16. Each one is identified by a name, e.g., Pre-Dial Plug-in. They are considered as default plug-ins.
1: Pre-Dial Plug-in (Default)
2: Post-Dial Plug-in (Default)
3: Idle Plug-in (Default)

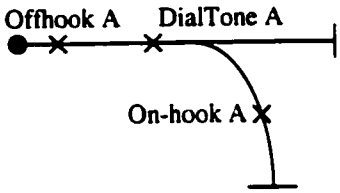
4: Idle Setup Plug-in (Default)

5: Disconnection Idle Plug-in (Default)

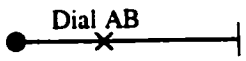
6: Busy Plug-in (Default)

7: None

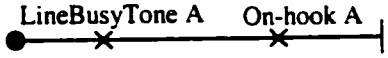
8: None



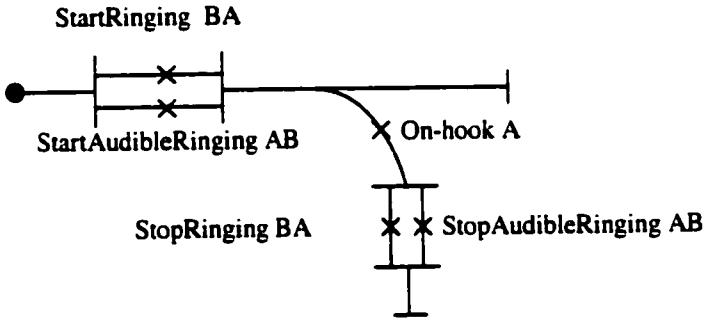
Stub1 :Pre-Dial plug-in



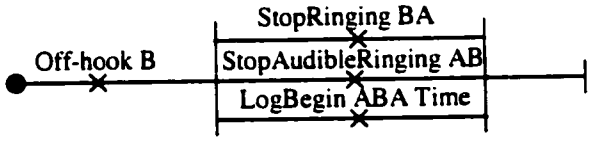
Stub 2: Post-Dial plug-in



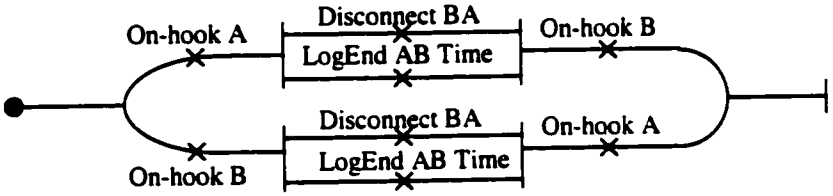
Stub 6: Busy plug-in



Stub 3: Idle plug-in



Stub 4: Idle Setup plug-in



Stub 5: DisconnectionIdle plug-in

Figure 16: Use Case Maps for Basic Call Model

The basic call model describes the basic actions for the establishment of a communication between two users A and B. First the user A goes pick up the phone (off-hook) then dials B's number. If B is busy then a "LineBusyTone" signal is sent to A (Busy plug-in),

otherwise the actions in the Idle plug-in are executed. The Idle plug-in starts with two actions taking place in either order: the phone at B's side starts ringing (StartRinging BA) and A receives an audible ringing (StartAudibleRinging AB) telling him that the phone on the called party is ringing. When the user B pick up the phone (off-hook B) to answer three actions took place at either order: the phone at B's side stops ringing (StopRinging BA), A stops receiving the audible ringing (StopAudibleRinging AB) and the billing procedure starts (LogBegin ABA Time) with respectively the caller, the called party, the charged party and the time of start billing as parameters.

4.3 Feature description using the UCM Model

In the 1998 Feature Interaction Contest [5], the feature requirements were described using the Chisel diagram notation.

We classify the features acting during the call establishment into two classes: *Originating features* and *terminating features*.

Features like TWC (Three Way Calling) and CW (Call Waiting), which need an already established communication between two users before they can perform their specific actions, cannot be classified into these two categories.

- **Originating features**

This class of features contains the ones that can be activated when the feature's subscriber tries to establish a call.

Among originating features we can mention OCS (see Section 1.2.1) and INTL (see Section 4.4.1).

Note: When describing an originating feature we suppose that the caller A is the feature's subscriber.

- **Terminating features**

This class of features contains the ones that can be activated when the feature's subscriber receives a call.

Among terminating features we can mention TCS (see Section 4.4.3), CFBL (see Section 4.4.5), INFB (see Section 4.4.4) and CND (see Section 4.4.2).

Note: When describing a terminating feature we suppose that the called party B is the feature's subscriber.

4.4 Feature Addition

Feature requirements described in the Chisel notation are represented by means of UCMs. The mapping from Chisel notation to UCM notation is straightforward because like UCMs the "Chisel diagrams" use an end-to-end model.

Table 1 gives some translation guidelines to translate Chisel notation to UCM notation:

Chisel Notation	UCM Notation
Event	Responsibility
Directed edges (arrows)	Scenario paths
Conditions	Guards
Interleaving operator |||	AND Fork

Table 1: Translation guidelines from Chisel to UCM

We consider in this thesis that the features are an Extension/Modification of the POTS because they use the stubs defined within the model. Adding features extends scenarios in the basic call model. In our framework, this is achieved in a simple way by using the stub plug-in concept of UCMs. Intuitively we only replace some default submaps with specific ones. The

UCMs obtained will be sets of submaps describing scenarios specific to the feature combined with the basic call model (Feature + BCM).

4.4.1 Originating feature: INTL (IN Teen Line)

Teen Line restricts outgoing calls based on the time of day (i.e., hours when homework should be the primary activity). This can be overridden on a per-call basis by anyone with the proper identity code. This is an IN feature.

Let us add INTL to the basic call. This is done by plugging INTL submap "INTL Pre-dial plug-in"(figure 17) into the Pre-dial stub of the root map in figure 15. This replaces the plug-in shown in figure 16.

Plug-ins for IN Teen Line:

1:INTL Pre-Dial plug-in

2:Default

3:Default

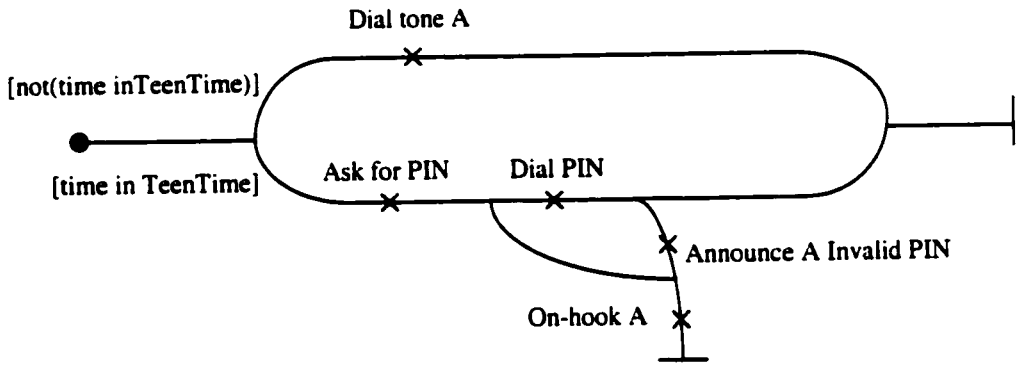
4:Default

5:Default

6:Default

7:None

8:None



Stub 1:INTL Pre-Dial plug-in

Figure 17: INTL plug-ins

When an INTL subscriber tries to place a call, two possible outcomes for his attempt are considered based on the time of the day and the correctness of his PIN.

1) If the call is initiated outside the Teen time the call should continue normally
2) If the call is initiated during the Teen time then a message asking for a PIN is sent to the caller (Ask for PIN). If the caller dials the valid PIN then the call continues normally

otherwise an announcement is sent to the user (Announce A invalid PIN) telling him that the PIN entered was invalid. In this case the Call is blocked.

4.4.2 Terminating feature: CND (Calling Number Delivery)

CND is a feature that allows the called telephone to receive a calling party's Directory Number (DN) and the date and time. In the on-hook state, in a real network, the delivery of this information occurs during the long silence between the first and second power ringing cycles. For the purpose of the thesis, we assume the capability of delivering the number, and deliver it whenever an idle called party receives the Ringing event.

The addition of CND is done by plugging CND submap "Calling Number delivery Idle plug-in" (figure 18) into the Idle stub of the root map in figure 15. This replaces the plug-in shown in figure 16.

Plug-ins for Calling Number Delivery:

1:Default

2:Default

3:Calling Number delivery Idle plug-in

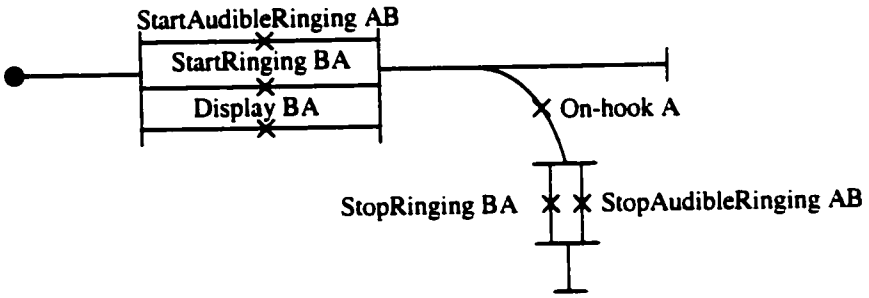
4:Default

5:Default

6:Default

7:None

8:None



Stub 3: Calling Number delivery Idle plug-in

Figure 18: CND plug-ins

4.4.3 Terminating feature: TCS (Terminating Call Screening)

Terminating Call Screening restricts incoming calls. Calls from lines that appear on a screening list are redirected to a vague but polite message.

The addition of TCS is done by plugging the TCS submap "TCS Post-Dial plug-in" (figure 19) into the Post-Dial stub of the root map.

Plug-ins for Terminating Call Screening:

1:Default

2:TCS Post-Dial Plug-in

3:Default

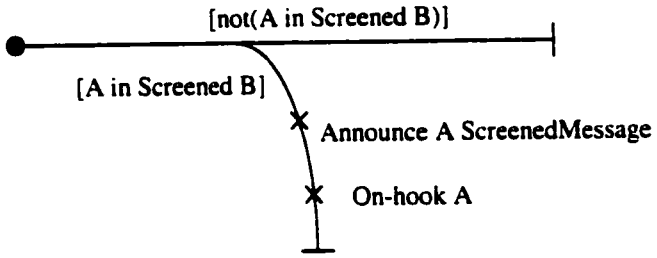
4:Default

5:Default

6:Default

7:None

8:None



Stub 2: TCS Post-Dial Plug-in

Figure 19: TCS plug-ins

Two possible outcomes are considered based on the presence or the absence of A (the caller party) in the B's screened list:

If A is in the B's screened list (condition: [A in Screened B]) then A receives an announcement telling him that is not allowed to call B, otherwise (condition: [not (A in Screened B)]) the call continues normally.

4.4.4 Terminating feature: INFB (IN Free Phone Billing)

The IN Freephone feature allows the subscriber to pay for incoming calls.

The addition of INFB is done by plugging INFB submap "INFB Idle setup plug-in" (figure 20) into the Idle Setup stub of the root map.

Plug-ins for IN Free phone Billing:

1:Default

2:Default

3:Default

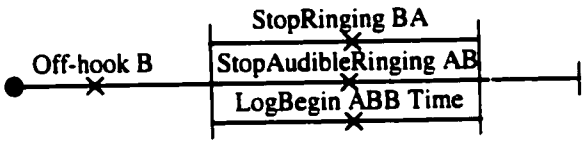
4:INFB Idle setup plug-in

5:Default

6:Default

7:None

8:None



Stub 4: INFB Idle Setup plug-in

Figure 20: INFB plug-ins

B is the INFB subscriber. B is charged when answering incoming calls. This is defined by the third parameter, which is the charged party in the billing action (LogBegin ABB Time instead of the normal ABA).

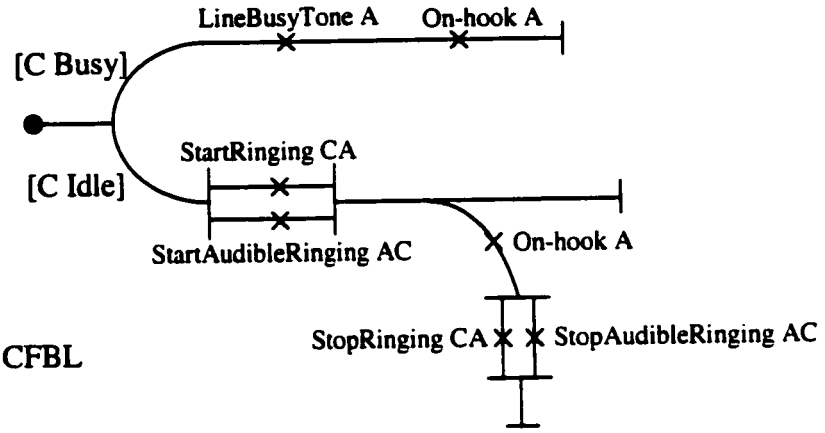
4.4.5 Terminating feature: CFBL (Call Forwarding Busy Line)

All calls to the subscribing line are redirected to a predetermined number when the line is busy. The subscriber pays any charges for the forwarded call from his station to the new destination. The subscriber's originating service is not affected.

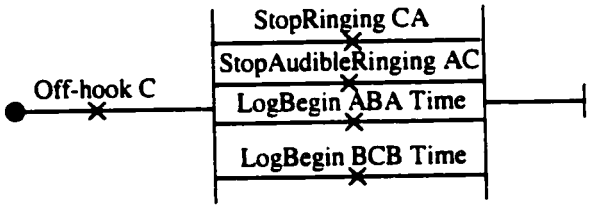
The addition of CFBL is done by plugging CFBL submaps "Busy CFBL" into the Busy stub, "Busy Setup CFBL" into the Busy Setup stub and the "Busy Disconnection CFBL" into the Busy disconnection stub.

Plug-ins for Call forwarding busy line:

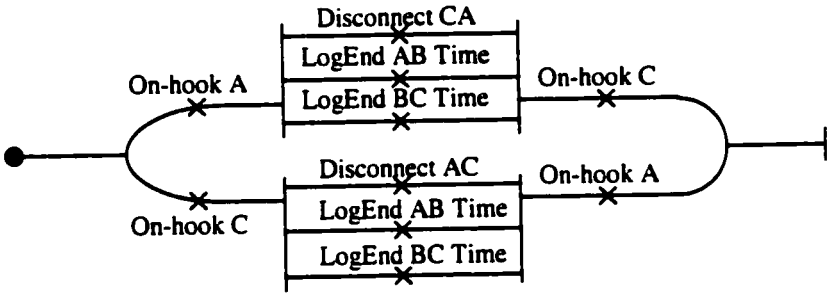
1:Default
2:Default
3:Default
4:Default
5:Default
6:Busy CFBL
7:Busy Setup CFBL
8:Busy Disconnection CFBL



Stub 6: Busy CFBL plug-in



Stub 7: Busy Setup plug-in



Stub 8: Busy Disconnection plug-in

Figure 21: CFBL plug-ins

In this scenario, we suppose that the originator is A, CFBL subscriber is B and the forward party is C. When the third party C is busy, the originator A gets a line busy tone signal. When C is idle the originator A gets an Audible Ringing signal while the destination C gets a

Ringing signal. Once the connection is established the originator A pays for "AB" leg while B pays for the forwarded leg of the call "BC".

4.5 Feature Interaction Filtering Method

Feature Interaction detection algorithms need a significant amount of work due to the need of considering all possible combinations of behaviours. Thus, Feature Interaction detection can be expensive and even infeasible task [50]. Therefore, it would be helpful to have a method that can be used before feature interaction detection to estimate which feature combinations have a possibility of feature interaction.

The goal of feature interaction filtering is:

- To localize where interactions could take place (e.g. in which stub)
- To take out the interaction free scenarios from further analysis

4.5.1 Stub Configuration Vector

We can characterize features in terms of stub configuration vector, that is, information regarding which feature submap is plugged into which stub of the root map. In this section, we propose a vector representation called stub configuration vector (or simply SC-vector), to characterize features.

General Definition: A stub configuration vector (or simply SC-vector) is a vector of length n $F = [f_1,\ldots,f_n]$, where f_i is the name of the plug-in of the i-th stub.

Note: For our UCM model n is equal to 8.

Example:
With A an INTL subscriber we have:
INTL=[INTL Pre-Dial plug-in, default, default, default, default, default, none, none]

Similarly, suppose that B is subscribes to TCS. Then,
TCS=[default, TCS Post-Dial plug-in, default, default, default, default, none, none]

4.5.2 Feature Composition

Once each individual feature is characterized by an SC-vector, we compose different configurations, in order to examine FI-Filtering between multiple features.

Composition Operators:

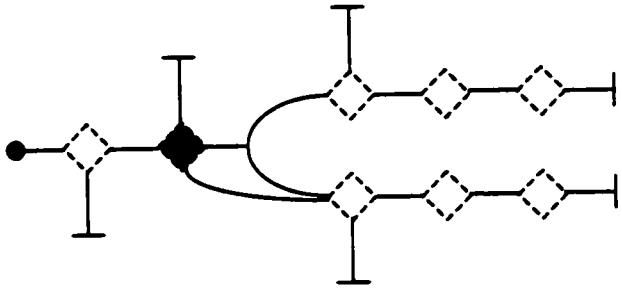
Suppose that f and g are two plug-ins plugged into the same stub in the root map (the proposed UCM model). Let *default* denote any default plug-in describing basic call scenarios. Let *ng* (stands for "no good") denote a special result not contained in the given plug-ins. Then, composition of f and g, denoted by $f \bullet g$, is defined as follows:

$$f \bullet g = g \bullet f = \begin{cases} f & (\text{if } f = g) \quad \text{(A1)} \\ f & (\text{if } g = default) \quad \text{(A2)} \\ f & (\text{if } g = none) \quad \text{(A3)} \\ ng & (\text{if } f \neq g) \text{ and } f,g \neq default \quad \text{(A4)} \end{cases}$$

Figure 22: Composition Operator "."

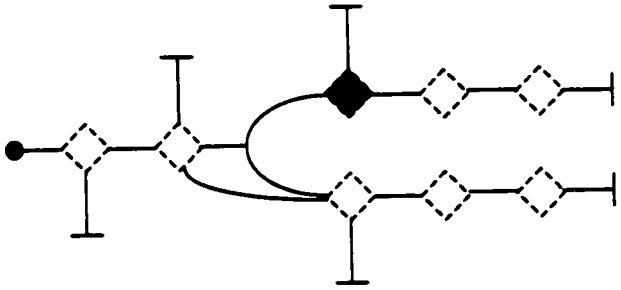
The intuitive semantics of the composition is explained as follows: (A1) composition of the same submaps yields the same submap, (A2) a feature submap f can override a default map of basic call scenario, (A3) a feature submap f can override a missing submap ("none" in the SC-stub), (A4) two different feature submaps cannot be plugged into the same stub, since a non deterministic behaviour arises between f and g.

Now, we define the composition of SC-vectors:
Let $F = [f_1,\ldots,f_n]$ and $G = [g_1,\ldots,g_n]$ be given SC-vectors. Then the composition of F and G, denotes by $F \oplus G$, is defined as $H = F \oplus G = [h_1,\ldots, h_n]$ where $h_i = f_i \bullet g_i$ for all i.
The composition of two SC-vectors is carried out by applying the "$\bullet$" operator to each pair of corresponding vector elements. Figure 23 illustrates an example of feature composition.



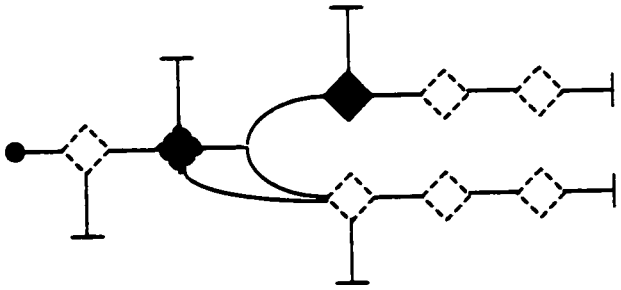
UCM for Feature 1(F1)

Feature1 = [default, F1 Post-Dial plug-in, default, default, default, default, none, none]



UCM for Feature 2(F2)

Feature2 = [default, default, F2 Idle plug-in, default, default, default, none, none]

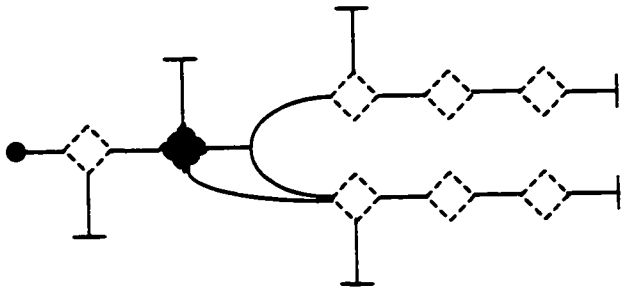


UCM for Feature 1 ⊕ Feature 2

Feature1 ⊕ Feature 2 = [default, F1 Post-Dial plug-in, F2 Idle plug-in, default, default, default, none, none]

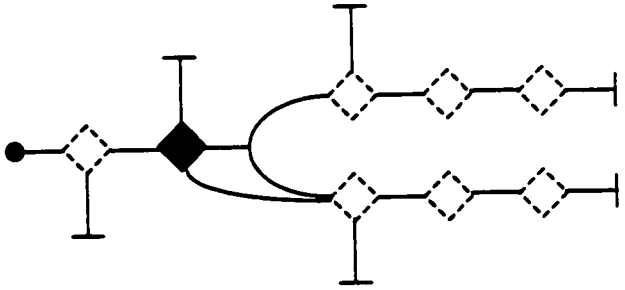
Figure 23: Feature Composition (1)

However the two following features, described in figure 24, cannot be combined because of a conflict in the Post-dial stub so they need further investigation to detect possible interactions.



UCM for Feature 1

Feature1 = [default, F1 Post-Dial plug-in, default, default, default, default, none, none]



UCM for Feature 2

Feature2 = [default, F2 Post-Dial plug-in, default, default, default, default, none, none]

Feature1 ⊕ Feature 2 = [default, **ng**, default, default, default, default, none, none]

Figure 24: Feature Composition (2)

Example:

Let us compose INTL with TCS:

INTL = [INTL Pre-Dial plug-in, default, default, default, default, default, none, none]

TCS = [default, TCS Post-Dial plug-in, default, default, default, default, none, none]

INTL ⊕ TCS = [INTL Pre-Dial plug-in, TCS Post-Dial plug-in, default, default, default, default, none, none]

So the two features can be combined without problems.

4.5.3 Feature Interaction targeted

We use the definition given in Chapter 1, which states that there is an interaction between features when the combined specification is inconsistent in some way, either due to specifying inconsistent state changes or inconsistent observable actions. The inconsistencies that we have addressed will be formalised by the two rules we give below.

4.5.4 Filtering rules

Let H = $F \oplus G$.

Filtering Rule 1:

There exists *ng* in H $\Rightarrow$ FI occurs (non determinism)

A *ng* entry appears in H iff the combination of the two features requires that a submap *fi* in F and a submap *gi* in G be plugged into a stub *i* simultaneously. If this is done, different scenarios are possible at the entry of the same stub, which causes non-determinism.

Figure 25 describes the inconsistency introduced by trying to put two different plug-ins into the same stub. The entry point represents state S1 from which we have two possibilities : execute X1 to get to state S2 or execute X2 to get to state S4.

- Inconsistent observable actions: Observing action X1 while we are expecting action X2 to occur and vice versa.
- Inconsistent state changes: Expecting to get to state S2 by executing X1 while we get to the state S4 by executing X2 and vice versa.

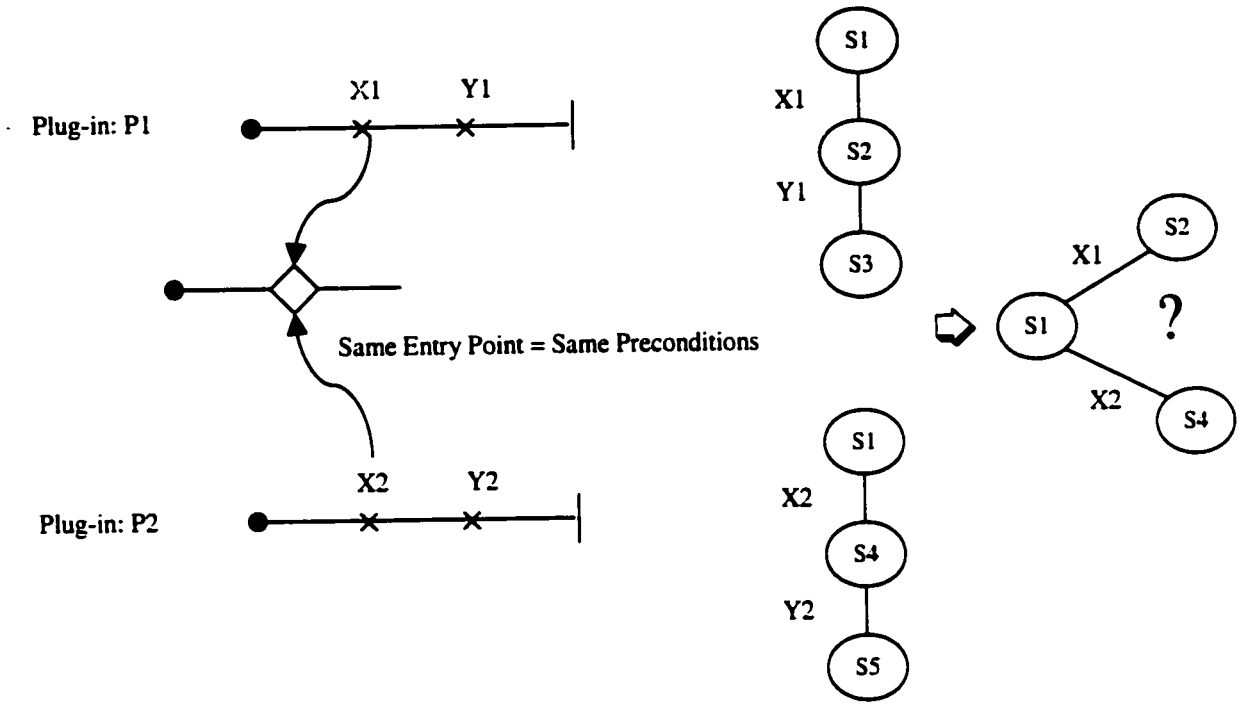


Figure 25: Non determinism (Interaction)

Filtering Rule 2:

(There is no *ng* in H) and (*f2* ≠ *default*) and (∃ *gi* ≠ *default where* $6 \leq i \leq 8$) ⇒ FI could occur (Inconsistent state changes)

The condition: "There is no *ng* in H" is introduced not to include cases already treated in the Filtering rule1.

This rule derives from the way we have decomposed our services. As described in section 4.2.2.2 in stage 2, the network checks the information (authorization, user data...etc.) to determine whether the service should be provided or not (Stubs 1 and 2 represent the originating part of the call). This means that the service could be denied or blocked in this stage if the user doesn't meet the required conditions. Stages 3 and 4 deal essentially with calls processing (Stubs 3 to 8 represents the terminating part of the call).

Let's analyse the preconditions of the filtering rule 2:

- $f_2 \neq default$: The feature *F* has a specific plug-in (f_2) for the Post-Dial stub (which treats the checking part), this means that there is a possibility for the call not to take place.
- $\exists gi \neq default \text{ for } 6 \leq i \leq 8$: The feature *G* has a specific behaviour in stubs 6 or 7 or 8, this means that the call may be forwarded/routed to a third party. The third party has only the terminating part of the call described in stubs 6, 7 and 8.

Interaction: Since the call could be denied (blocked) in the originating part of the call (feature F, $f2 \neq default$) and this part of the behaviour is missing in feature G for the third party, an interaction could occur.

Example: User A is an OCS subscriber (OCS has a specific Post-dial plug-in in stub 2. $OCS_2 \neq default$). User B is a CFBL subscriber (CFBL has a specific plug-ins in stubs 6, 7 and 8. ($CFBL_i \neq default$ for $6 \leq i \leq 8$). When A calls B the call is forwarded to C and not blocked, which is a feature interaction.

4.5.5 Completeness

From the previous section we have seen that when one of the two filtering rules is applied there is a possible interaction. For completeness purpose let's analyse all remaining cases.

We distinguish three cases:

❶ Feature F has specific plug-ins (different from default) for at least one of the stubs 3, 4 and 5, feature G has also specific behaviour (different from default) for at least one of the stubs 3, 4 and 5, and $F \oplus G$ doesn't contain any "ng" (see figure 26): These two features act at distinct call levels of the same leg of the call (before, during or after the call establishment). Therefore their behaviour could be executed sequentially. So F and G are said to be "Interaction Free".

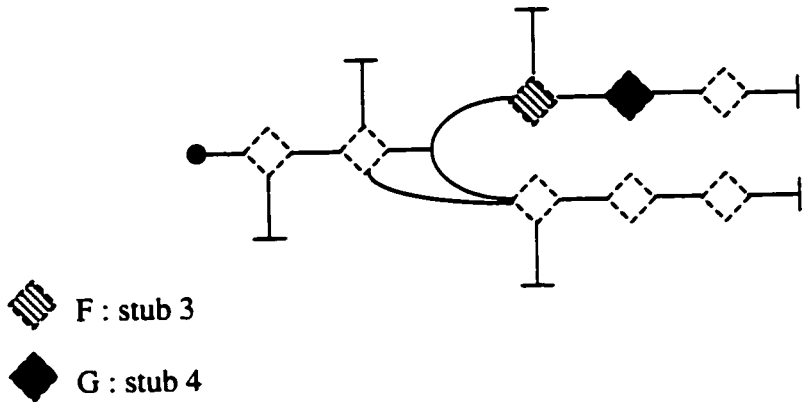


Figure 26: $F \oplus G$:Interaction Free Case 1

❷ Feature F has specific plug-ins (different from default) for at least one of the stubs 3, 4 and 5, feature G has specific behaviour (different from default) for at least one of the stubs 6, 7 and 8, and $F \oplus G$ doesn't contain any "ng" (see figure 27): These two features have different preconditions: for feature F, the called party should be Idle whereas for feature G the called party should be in a busy state (or we don't care about the called state) and the call is forwarded to a third party. Therefore the two features are mutually exclusive (Only one of them is active at a time). So F and G are Interaction free.

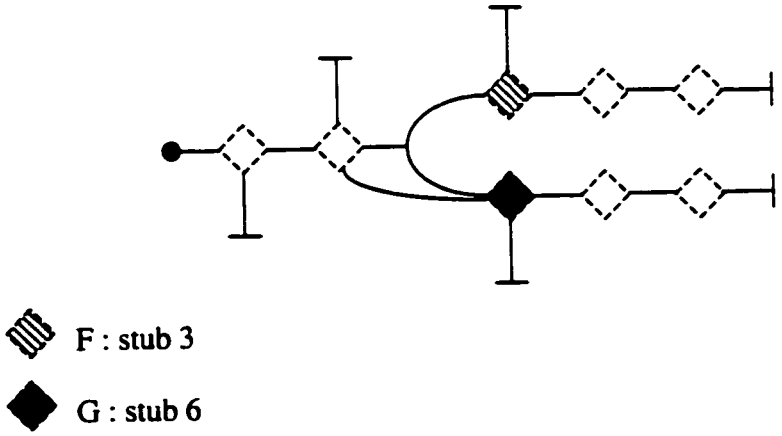


Figure 27: $F \oplus G$ Interaction Free Case 2

❸ Feature F has a specific plug-in in stub 2 ($f_2 \neq$ *default*), feature G has a specific behaviour (different from default) for at least one of the stubs 3, 4 and 5, and $F \oplus G$ doesn't

contain any "ng" (see figure 28): The model is designed such that the stub 2 precedes stubs 3, 4 and 5. The priority is given to feature G over F and we assume that in this case F and G are Interaction free.

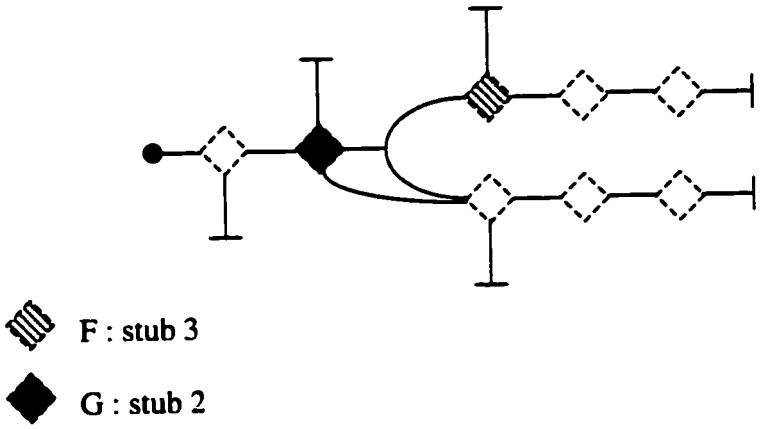


Figure 28: $F \oplus G$ Interaction Free Case 3

Notes:

- All the combinations involving the stub 1 where rules 1 and 2 are not applied are Interaction free because the call is not initiated yet at that level.
- The detection method presented in Chapter 6 doesn't consider any precedence assumptions.

4.5.6 Feature Interaction Filtering Method

Interaction between features depends on the way these features are assigned to subscribers. For example, an interaction may exist between feature F1 and feature F2 if F1 and F2 are assigned to the same subscriber. It is possible that there would be no interaction if they are assigned to different subscribers. Therefore, to detect interaction, we should look at all possible assignments of features to subscribers. In the case of two features, we should look at the system when both features are assigned to the same subscriber, and when the two features are assigned to different subscribers.

In our system, the assignment of features to subscribers is done statically, i.e. a feature is specified independently and assigned to a single subscriber.

Before looking for FI scenarios we choose to discard the irrelevant scenarios from the filtering process. This reduces the number of possible scenarios to be investigated.
So scenarios that should be discarded are those of the following types:

- A user subscribes to two features, one originating and one terminating. These scenarios are useless since a user cannot be a call originator and a call terminator at the same time.
- B or C subscribes to an Originating feature. These scenarios are useless since the originating feature is inactive for B and C during the scenario.

Filtering Method Input:

Feature Stub configuration Vectors: F = [$f1$,...,fm], G = [$g1$,...,gm]...etc.

Filtering Method Output:

Proceeding by a pair wise filtering the Output will be one of these three verdicts:

(1) Feature Interaction occurs

(2) Feature Interaction could occur

(3) F and G are Interaction Free

We provide a filtering routine to be used in the general filtering procedure presented below.

Filtering Routine:

❶ Make a composed vector $F \oplus G$

❷ If some "ng" elements exist in $F \oplus G$ then conclude that FI occurs (verdict 1) (From the Filtering Rule 1)

❸ If Filtering rule 2 holds then conclude that FI could occur (verdict 2) otherwise F and G are Interaction Free

Filtering procedure:

The procedure consists of three steps:

Step 1:

This step aims to treat scenarios where the same user subscribes to two features. The step 1 is divided into 2 sub steps:

- The feature SC vectors are provided
- Apply the Filtering Routine and get the verdict.

Step 2:

This step aims to treat scenarios where the two features are distributed between the caller and the called parties only (i.e. A and B).

The step 2 is also divided into 2 sub steps:

- The feature SC vectors are provided
- Apply the Filtering Routine and get the verdict.

Step 3:

This step treats scenarios where a third party C is the feature subscriber and B arranges to forward/route the call to C.

The step 3 is divided into 2 sub steps:

Step 3.1: SC Refinement:

C is a terminating party. So only the terminating part of the SC vector interests us.

To be able to detect the feature interactions we should look more in depth to the 6^{th}, 7^{th} and 8^{th} stubs for the feature that introduces the third party (i.e. CFBL, INCF, INFR...etc.). Figure 29 illustrates the refinement process.

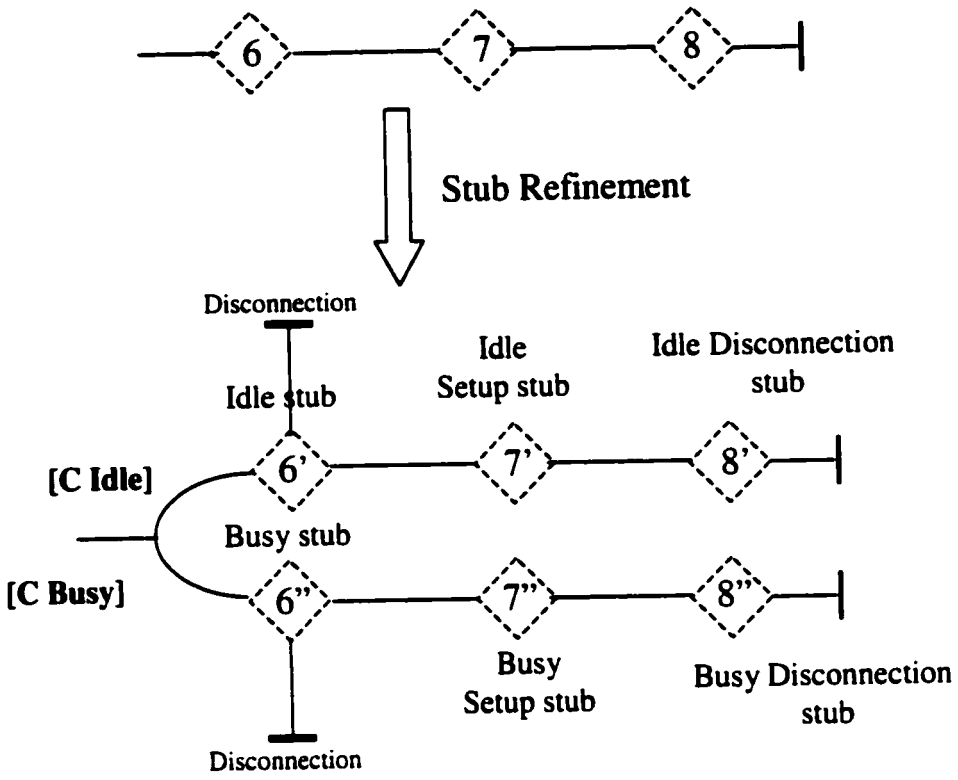


Figure 29: UCM Refinement

Stub Configuration sub-vector: A stub configuration sub-vector (or simply SC sub-vector) is a vector F' = [f'_1,...,f'_n], where f'_i is the name of the plug-in of the i-th stub.

Note: For our UCM model n is equal to 6 and an SC sub-vector corresponds to a vector:
F'= [6' plug-in, 7' plug-in, 8' plug-in, 6" plug-in, 7" plug-in, 8" plug-in]

Step 3.2: Define the Stub Configuration sub-Vector (F' and G') for each feature

Note: The SC sub-vector will be the terminating part for the feature that introduces the third party. The feature that introduces the third party will be refined as follows:

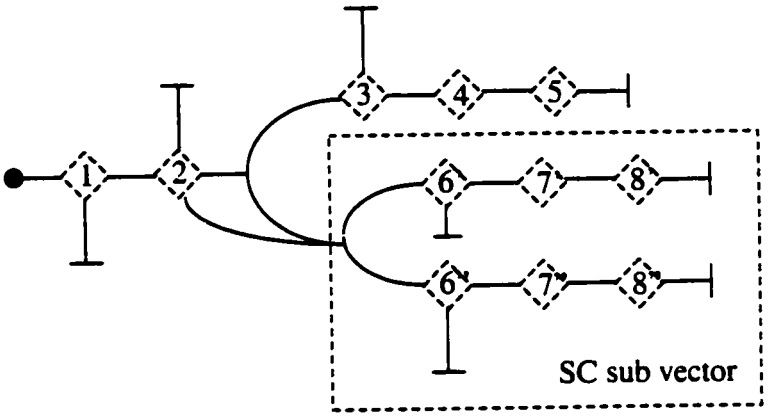


Figure 30: Feature causing the forward/Routing of the call (B)

The feature, to which the third party is subscribed, is be described as:

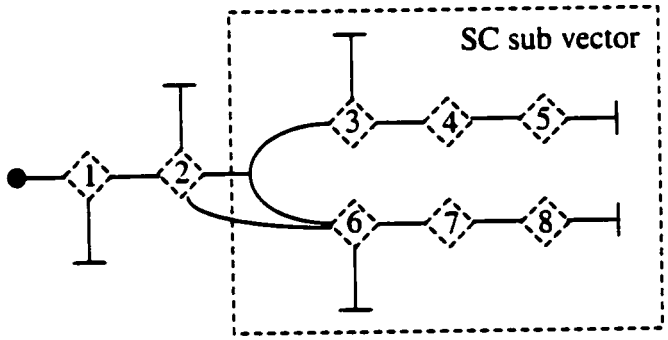


Figure 31: Feature to which C subscribes

- Make a composed SC sub-vector H'= [h'1,...,h'm] = $F' \oplus G'$ (that contains only 6 stubs)
- If some "ng" elements exist in H', conclude that (1) FI occurs (Filtering Rule 1)
- If the feature to which C subscribes has a specific behaviour for stub 2 (which is not described in the SC sub-vector) then conclude that (2) FI could occur (The filtering rule 2)

  Else F and G are (3) Interaction Free (In the cases where C is a feature subscriber)

4.6 Application

As an application of the proposed method we will illustrate the interactions happening between OCS (Originating feature), CND (Terminating Feature) and CFBL (Terminating

Feature). A is the caller; B the called and C is the third party. Table 2 illustrates some of the different possible distributions of the features between the users. Those that are not listed are of no interest.

	User A	User B	User C
1	OCS	CND	-
2	CND	OCS	-
3	OCS, CND	-	-
4	-	OCS, CND	-
5	OCS	CFBL	-
6	-	OCS, CFBL	-
7	-	CFBL	OCS
8	CND	CFBL	-
9	-	CND, CFBL	-
10	-	CFBL	CND
11	CND	CFBL	-
12	OCS	OCS	-
13	CND	CND	-
14	-	CFBL	CFBL

Table 2: Distribution of the features between users

Scenarios 3, 4 and 6 are useless because one of the users subscribes to both an originating and a terminating feature. Scenarios 2, 8, 11 and 13 are also useless since user A subscribes to a terminating feature. Scenarios 7, 12 are useless since B and C subscribe to an originating feature. This leaves scenarios 1, 5, 9, 10, 14, which are examined in detail below.

The two features are distributed between the caller and the called parties only:

- **Scenario 1:** OCS (A) and CND (B)

Stub #	1	2	3	4	5	6	7	8
CND	default	default	Calling Number delivery Idle plug-in	default	default	default	None	None
OCS	default	OCS Post-Dial Plug-in	Default	default	default	default	None	None
CND ⊕ OCS	default	OCS Post-Dial Plug-in	Calling Number delivery Idle plug-in	default	default	default	None	None

There is no *ng* in the resulting behaviour and filtering rule 2 does not hold.

No Interactions Detected.

- **Scenario 5:** OCS (A) and CFBL (B)

Stub #	1	2	3	4	5	6	7	8
OCS	default	OCS Post-Dial Plug-in	Calling Number delivery Idle plug-in	default	default	default	None	None
CFBL	default	default	default	default	default	Busy CFBL	Busy Setup CFBL	Busy Disconnection CFBL
OCS ⊕ CFBL	Default	OCS Post-Dial Plug-in	Calling Number delivery Idle plug-in	default	default	Busy CFBL	Busy Setup CFBL	Busy Disconnection CFBL

There is no *ng* in the resulting behaviour. However OCS affects the post-dial stub (stub 2) whereas CFBL affects the stubs 6, 7 and 8.

According to the filtering rule 2 a **FI could occur.** The interaction is that the call is not blocked when the call is forwarded to third party C, which is in the screening list of A.

The same user subscribes to two features:

- **Scenario 9:** CND (B) and CFBL (B)

Stub #	1	2	3	4	5	6	7	8
CND	default	default	Calling Number delivery Idle plug-in	default	default	default	None	None
CFBL	default	default	default	default	default	Busy CFBL	Busy Setup CFBL	Busy Disconnection CFBL
CND ⊕ CFBL	default	default	Calling Number delivery Idle plug-in	default	default	Busy CFBL	Busy Setup CFBL	Busy Disconnection CFBL

No *ng* in the resulting behaviour and filtering rule 2 does not hold.

No Interactions Detected.

A third party is the feature subscriber and B arranges to forward/route the call to C

- **Scenario 10:** CFBL (B) & CND(C)

CFBL = [default, default, default, default, default, Busy CFBL, Busy Setup CFBL, Busy Disconnection CFBL]

CFBL SC sub-vector:

CFBL_Refined = [Idle CFBL_Ref, Idle Setup CFBL_Ref, Idle Disconnection CFBL_Ref, default, none, none]

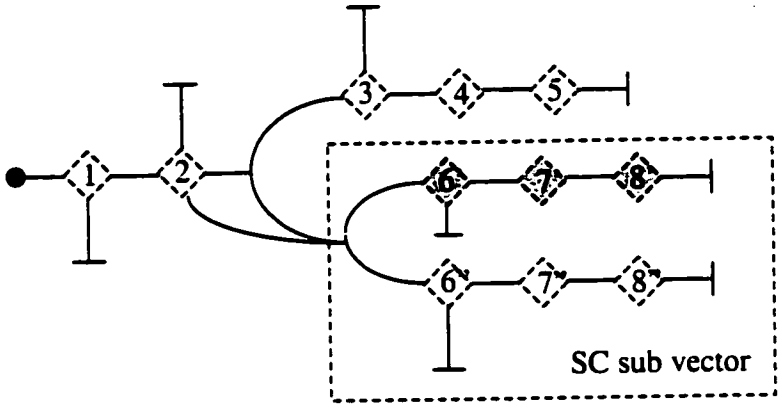


Figure 32: CFBL (B) SC sub vector

CND_Refined = [Calling Number delivery Idle, default, default, default, None, None]

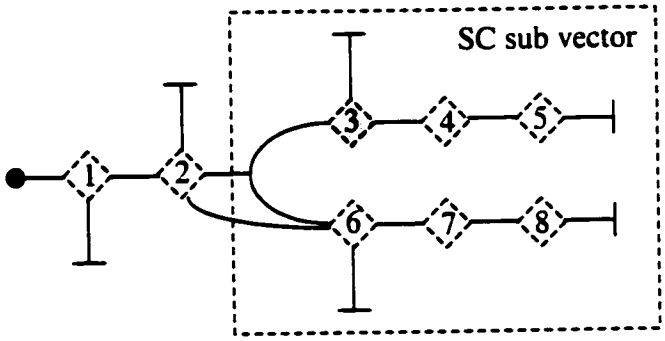


Figure 33: CND (C) SC sub vector

CFBL_Refined ⊕ CND_Refined = [**ng**, Idle Setup CFBL_Ref, Idle Disconnection CFBL_Ref, default, default, none, none]

There is *ng* in the resulting behaviour. **FI occurs.** Display Conflict: The number is not displayed at C.

- **Scenario 14:** CFBL (B) & CFBL(C)

The way CFBL is defined in the contest prevents the forwarding loop because CFBL as designed tests the forwarded-to line for busy and returns LineBusyTone if it is busy. CFBL is deactivated at C.

4.6.1 Results

We have prepared UCMs for the following nine features: Originating Call Screening (OCS), Terminating Call Screening (TCS), IN Free Routing (INFR), Call Forwarding Busy Line(CFBL), IN Teen Line (INTL), Call Number Delivery (CND), IN Freephone Billing (INFB), IN Call Forwarding (INCF) and IN Charge Call (INCC).

Table 3 shows the filtering results.

	OCS		**TCS**		**INFB**		**INCC**		**INTL**		**CND**		**CFBL**		**INCF**		**INFR**	
	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
OCS	-	-	-	1	-	3	1	-	3	-	-	3	-	2	-	2	-	2
TCS			-	-	3	-	-	1	-	3	3	-	3	2	3	2	3	2
INFB					-	-	-	1	-	3	3	-	3	1	3	1	3	1
INCC							-	-	1	-	-	3	-	2	-	2	-	2
INTL									-	-	-	3	-	3	-	3	-	3
CND											-	-	3	1	3	1	3	1
CFBL													-	3	1	1	1	1
INCF															-	1	1	1
INFR																	-	1

(a) : Same user

(b) : Different users

(1) FI occurs

(2) FI could occur

(3) FI never occur

(-) Useless scenarios: only one feature is active

Table 3: Filtering Results

Statistics & Discussion:

39 Scenarios (from 90 possible scenarios. These 90 scenarios represent all possible combinations of feature pairs among 2 and/or 3 different users) are discarded before the filtering process because they are useless scenarios. 51 scenarios remain to be investigated.

19 Scenarios (about 37 %) lead to a Feature interaction.

9 Scenarios (about 18%) need more investigation. (FI could occur)

23 Scenarios (about 45%) are safe scenarios (Interaction free).

The number of combinations with verdict (1) FI occurs and (3) FI never occur is 42, which is almost 82 % of all the combinations. That is, 82% of all scenarios can be filtered by the proposed method in an inexpensive way.

4.7 Comparison between our method and the method presented in [29]

The filtering methods presented in [50] and in this chapter have similarities and differences:

Similarities

- Both methods use the same UCM call model to describe features.
- The feature composition mechanism is slightly different: In [50] the feature composition operator operates on matrices while it operates on vectors in our method.
- They use the same filtering rule 1 to detect non-determinism.
- Neither method covers all possible features because they are based on the same UCM model.

Differences

- The filtering method described in this thesis is a "feature oriented method" since each feature is described as a UCM, which contains all involved users. However the filtering proposed in [50] is user oriented in the sense that the composition takes care of every single user scenario.
- Our filtering rule 2 derives from the way the service is decomposed in UCM and detects inconsistent state changes. While [50] detects the scenario changes after the feature composition. One of the theorems introduced in [50] states that if no user's scenario is changed after feature composition there is no feature interaction.

4.8 Conclusion and Limitations

We have developed a method for feature interaction filtering that uses information that is available at the design stage of a telephone system.

This method allows the designer to localize where interactions could occur during a call and facilitates the further detection by taking out the interaction-free scenarios.

The method leads to certain results concerning existence or nonexistence of interactions, however in order to perform the filtering our structural model should be followed.

Another limitation of our method is that the model used for filtering does not cover all possible features. Features that handle more than one communication leg are not covered. Features like TWC (Three Way Calling) and CW (Call Waiting) deals with more than one communication leg. To be able to describe these features, a new model that describes more than one communication leg should be considered. As a sketch of solution, we propose a UCM model where each call leg (for instance AB between A and B, AC between AC) is described in a separate stub. The triggering event, which attempts to introduce a new call leg, would be described out of these call stubs representing involved call legs. We should be able to go from one call leg to another when the feature-triggering event is triggered. Figure 34 gives a sketch of what such model looks like, where stubs 1 and 2...etc describe the actions performed within each single call leg.

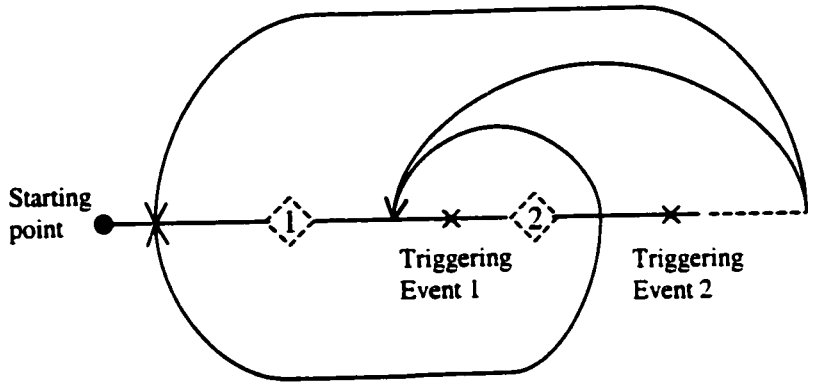


Figure 34: Sketch of multi-leg call Model

Chapter 5

Specifying Features Using LOTOS

In this chapter, we give an overview of the LOTOS specification language and of its main operators.

Our main objective in specifying the system model and features in LOTOS is to provide a specification that can be used for detecting feature interactions.

5.1 Introduction

LOTOS (Language of Temporal Ordering Specification) (ISO 8807,1989) was developed by the FDT experts of the working group ISO/TC97/SC21/WG1 during 80's. It is a specification language developed for the formal description of the various elements of the OSI (Open System Interconnection) architecture such as services and protocols. Nowadays, the LOTOS application area has been extended to cover some other domains such as hardware [42] and telephony [43][18].

The basic idea of LOTOS is that systems can be specified by defining the temporal relations among the interactions that constitute their externally observable behaviour (ISO 8807, 1989). LOTOS is made up of two components:

(i) A data type component, which is based on the formal theory of algebraic abstract data types ACT ONE (Ehrig and Mahr, 1985) [15]. It deals with the description of data structures and value expressions.

(ii) A control component, in which the external observable behaviour of the system is described. It is based on Milner's Calculus of Communicating Systems (CCS) [61], which includes the concepts of parallel processes that communicate through a synchronization mechanism. However the concept of multi-way synchronization is derived from Hoare's CSP [66].

A number of excellent LOTOS tutorials exist in the literature [41][11][66]; therefore, we limit ourselves to a very brief overview of the language and of its use in the context of our research.

All the LOTOS reserved words used in this thesis are written in **Bold**.

5.2 LOTOS Abstract data types

In LOTOS, the representation of values, value expressions and data structures are derived from the algebraic specification method ACT ONE. The properties and operations of data are defined without any indication about how these data are represented and manipulated in memory. In addition, LOTOS provides features such as the use of a library of predefined data type, extensions and combinations of already existing specifications, parameterization and actualization of specifications, and renaming of specifications, in order to facilitate the specification of systems with a large number of operations, equations and complex data types.

A data type definition in LOTOS consists of a signature and possibly of a list of **eqns** (equations). A signature of a type is a definition of its **sorts** and **opns** (operations). Sorts define the domain name of the data. **opns** define the formats of operations on the data. **eqns** provide a means to define the semantics of operations.

Consider the following type definition of the users directory numbers:

```
type Number  (* define the type name *)
is Boolean, NaturalNumber (* list other sorts used to construct this data type *)
(* Signature *)
sorts number (* define the sort name *)
opns (* specify the format of operations *)
        null, A, B, C, D          :→ number (* Constants *)
        _eq_                      : number, number → Bool
        _ne_                      : number, number → Bool
        to_nat                    : number → Nat

        (* List of equations *)
        eqns    forall n1, n2:number

                ofsort nat
                        to_nat(A) = 0;
                        to_nat(B) = Succ(0);
                        to_nat(C) = Succ(Succ(0));
                        to_nat(D) = Succ(Succ(Succ(0)));
                        to_nat(none) = Succ(Succ(Succ(Succ(0))));

                ofsort Bool
                        n1 eq n2 = to_nat (n1) eq to_nat(n2);
                        n1 ne n2 = not (n1 eq n2) ;
endtype
```

The signature of the type Number, identified by the keyword **sorts**, includes the sort number and the operations null, A, B, C, D, eq, ne and to_nat. The operations null, A, B, C, D result in five elements of the sort number, eq and ne define respectively the equality and inequality between variables of this type, and to_nat is the operation which maps a variable into a natural number (the type defining natural numbers is NaturalNumber and is already defined in the abstract data type library). This mapping is used to define the semantics of eq and ne operations, as described below.

The first five equations of sort nat define the images of the defined variables, null, A, B, C and D by the operation to_nat, then, to each variable of sort number corresponds a value of sort Nat (defined in the type NaturalNumber). The equation of sort Bool then defines the equality and inequality between two values of sort number. They are equal, respectively not equal, if their images by the operation to_nat are equal, respectively not equal. The operations eq and ne are already defined in the library type NaturalNumber.

5.3 The Control component

The behaviour component is the part of a specification that deals with the description of the system behaviour. In this part, a system is modeled as a collection of processes interacting with each other.

5.3.1 LOTOS Process

A process is viewed as a black box interacting with other processes or with the system environment via synchronization on its observable gates. It is basically defined by a set of observable gates, on which synchronization occurs, and by a behaviour expression. A behaviour expression is built by combining LOTOS actions by means of operators and possibly instantiations of other processes.

The syntax of a process definition is of the form:

Process process_name [gate_list] (formal_parameter_list) : functionality

< Behaviour expression>

endproc

In addition to the set of observable gates and to the behaviour expression, a process can also have a set of parameters, denoted in the definition above by parameter_list. This set represents the set of parameters through which values can be passed to the process from outside. The parameterization of a process also enables reusability.

5.3.2 LOTOS Action

An *action* is the basic element of a behaviour expression. It consists of a gate name, a list (possibly empty) of events, and possibly a predicate, which defines the conditions that should hold for the event to be offered. An event can either offer (!) or accept (?) a value. Predicates establish a condition on the values that can be accepted or offered.

An action has the following syntax:

Gate	Event1	Event2	Optional Predicate
G	? Get:Type	!Put	[Get <> 0]

As an example, consider the following two actions:

1. OffHook ?caller: number

The action occurs at gate "OffHook", which expects from the environment a value of sort number for the caller number.

2. DialTone !caller

The action occurs at gate "DialTone" and offers the value of the caller number (already assigned in the previous action) to the environment.

Actions are considered to be atomic in the sense that they occur instantaneously, without consuming time. Two types of actions exist in LOTOS. There are internal actions that a process can execute independently, are unobservable to the environment and are represented by the internal action i; and there are actions that need to synchronize with the environment in order to be executed. The environment of a process consists of other processes, or some external world that can be a human observer.

5.3.3 LOTOS Behaviour Expressions

The following are the basic behaviour expressions:

- *Inaction*: **stop**

  It represents a deadlock, i.e. No more actions can be executed.

- *Successful Termination*: **exit**

  It indicates a normal termination of a behaviour, i.e. a process has successfully performed all its actions. The keyword **exit** is also used in process definitions to express the process functionality (denoted in the syntax given above by functionality). In fact, a process has functionality **exit** if it can terminate successfully, i.e. it is able to perform an exit at the end. If a process cannot perform an **exit**, its the functionality is **noexit**.

- *Process Instantiation*: Process_Name[gate_list](actual parameter_list)

  The instantiation of a LOTOS process is equivalent to the invocation of a procedure in a programming language (such as Pascal). It can occur in the behaviour expression of other processes or in the behaviour expression of the process itself.

5.3.4 LOTOS Operators

It is possible to construct more complex expressions from those mentioned in section 5.3.3 by using LOTOS operators:

- *Action prefix operator:* a ; B

  The action prefix operator, written as a semi-colon (;), expresses sequential composition of an action a with a behaviour expression B.

For example, when a user (the caller) picks up the phone to make a call, she/he will get a tone. This can be expressed by a behaviour expression composed of two actions:

OffHook ?caller: number ;
DialTone !caller ;...

- *Choice Operator:* B1 [] B2

  The choice operator [] denotes the choice between two or more alternative behaviours.

  For example, when a user picks up the phone to make a call and gets a tone he/she can either dial a number (called number) and continue the processing of a call or hang up (OnHook). This can be expressed by the behaviour expression:

```
OffHook ? caller:number ; DialTone !caller;
        (Dial !caller ? called: number ; ...
        [ ]
        OnHook !caller ; stop )
```

- *Enabling:* B1>> B2

  The enable operator >> has a similar function as the action prefix operator but is used to express sequential composition of two behaviour expressions. B1 has to terminate successfully (**exit**, see section 5.3.3) in order for B2 to be executed.

- *Disabling*: B1[>B2

  The disable operator [> is used to express situations where B1 can be interrupted by B2 during normal functioning. For example, a normal processing of a call could be interrupted at any point if the caller hangs up. This could be expressed by the behaviour expression:

```
(OffHook ? caller:number ;
 DialTone !caller ;
 Dial !caller ? called: number ;
 ...
) [>   OnHook !caller ; ...
```

- Guarded Behaviour: [P] → B

  The behaviour expression B can be executed if and only if the formula P is true, it becomes **stop** otherwise. For example, a telephone can ring at a called side only if the called is not busy. This could be expressed by the behaviour expression:

```
[called NotIn Busy] → StartRinging !called! caller ; ...
```

- *Interleaving operator:* **B1 ||| B2**

  If B1 and B2 are in interleaving, they can perform their actions independently of each other. This operator expresses the concept of parallelism between behaviours where no synchronization is required. For example, a user dials a number, which is idle, then two

actions in either order can take place: the phone rings at the called party (StartRinging) and the caller gets a Ringback tone (StartAudibleRinging).

```
OffHook ? caller:number ;
DialTone !caller ;
Dial !caller ? called: number ;
    (
    StartRinging !called! caller ;
    exit
            |||
    StartAudibleRinging !caller !called ;
    exit
    ) >> ( Offhook !called ;...
```

- *Parallel Composition:* **B1|[g_1,...,g_n]| B2**

The parallel composition of B1 and B2 on the gate list $g_1,\ldots,g_n$ expresses the fact that B1 and B2 behave independently, with the exception that they must synchronize on the gates $g_1,\ldots,g_n$, which means that processes B1 and B2 must participate in the execution of every action defined with a gate name g_i, $i \in \{1,\ldots,n\}$. Then the interleaving operator, explained above, can be defined as parallel composition on an empty gate list. Therefore ||| and || (see below) are special cases of this operator.

Synchronization of processes on a gate g_i, $i \in \{1,\ldots,n\}$ occurs, if each process provides an action with a gate name g_i, the list of events offered with the actions match, and the predicates (if any) are satisfied. The list of events of two actions "match" if the following conditions are satisfied:

1) The numbers of events in the two actions match.
2) An event in one action offers (!) the same value or accepts (?) a value of the same sort.

We will give now an example of use of this operator:
Consider the following two processes: CFBL_feature and INTL_feature, which represent the partial specification of the features: Call Forwarding Busy Line feature and IN Teen Line.

In order to detect the interactions between these features, we synchronize them on their common gates: OffHook, DialTone, Dial, OnHook,...etc. (the feature interaction detection method is discussed in Chapter 6)

```
Process CFBL_feature[Offhook,..] (B_State, ...) :noexit:=
OffHook !A;
DialTone !A;
(
 Onhook !A; stop
  [ ]
 (
  Dial !A!B;...
 )
)
...
endproc


Process INTL_feature[Offhook,..] (B_State, T, ...) :noexit:=
OffHook !A;
(
     Ask_For_PIN !A;
     Dial_PIN !A ? x:PIN;
 (
    [x eq Invalid_PIN] →      PlayAnnoucement!A;
                              Onhook! A; stop
                    [ ]
    [x eq Valid_PIN] →        DialTone !A;
                              Dial !A!B; ...
                    )
)
...
endproc
```

Let us compose these two processes as follows:

The synchronization between these two processes is described as:

```
CFBL_feature[Offhook,..] (B_State,...)
|[OffHook,DialTone,OnHook,Dial]|
INTL_feature[Offhook,..] (B_State, ...)
```

The processes will execute as follows. They start by synchronizing on gate "OffHook" offering the same value, which is A. The process CFBL_feature is now blocked waiting for the "DialTone" action from INTL_feature. However the actions "Ask_For_PIN !A", "Dial_PIN !A ?x:PIN", which are not in the synchronization list, are executed. Then depending on whether the predicate [x eq Valid_PIN] is true (in the case where the user enters a valid PIN:Personal Identification Number or not) the "DialTone" or the "OnHook" action are executed.

- *Full Synchronization:* B1 || B2

  The full synchronization of B1 and B2 is a parallel composition in which B1 and B2 must synchronize on all their gates. This is also a special case of the interleaving operator, where the set $\{g_1, ..., g_n\}$ is the set of all the gates of the two processes.

- *Hiding operator* : **hide** $g_1, ..., g_n$ **in** B

  Used to hide actions synchronizing on gates ($g_1, ..., g_n$), which become internal (i.e. they become i) for the environment. Thus, hidden actions cannot synchronize with the environment.

5.4 Mapping From UCM to LOTOS

We have chosen LOTOS in order to narrow the gap between the requirement notation Use Case Maps and the executable model. We are going to use the UCM description of the features discussed in Chapter 3 to obtain the corresponding LOTOS specification.

We propose a mapping between the UCM notation and the corresponding LOTOS operators. We don't cover all the UCM components but only the ones we use in our UCMs. A method for translating UCM into LOTOS was outlined in [22]. However, a complete translation

is difficult and is the subject of ongoing work. In this research, we used a manual translation based on the guidelines that follow, which are limited to a subset of the UCM notation.

Translation Guidelines:

- "Start points" and "end points" are usually represented by LOTOS gates.
- UCM responsibilities are also represented as gates, sometimes with additional message exchanges.
- LOTOS gates representing UCM responsibilities and channels that are not observable by users are hidden through the **hide** operator.
- **Sequence**: The sequence is a very common pattern that can be found in all UCMs. In following example, we consider a partial UCM containing three consecutive responsibilities. We obviously see that actions are directly mapped onto LOTOS gates and the sequence is mapped into the prefix operator ";".

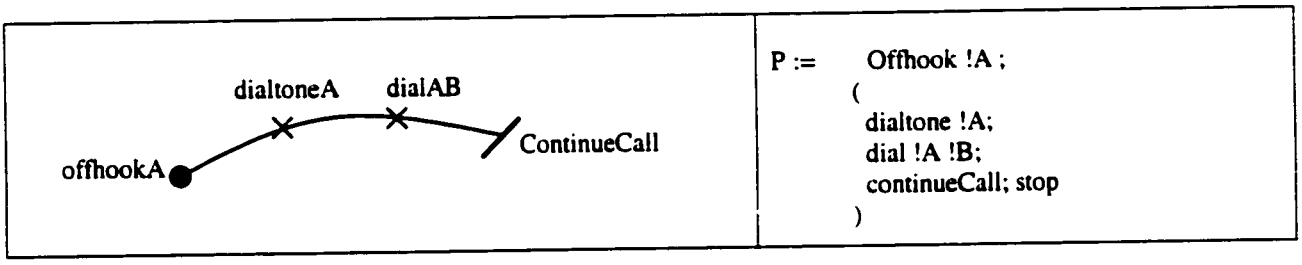


Figure 35: Sequence

- **OR-Fork**

Suppose the user A has a choice between dialing a number and hanging up.

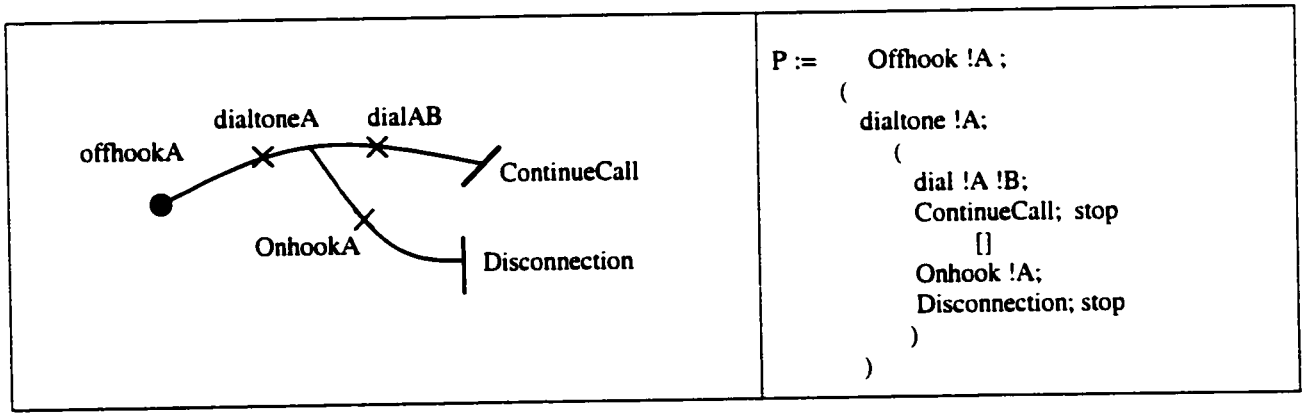


Figure 36: OR-Fork

Figure 36 presents such a (simplified) UCM. It shows two exclusive paths that will never join. The LOTOS choice operator ([]) is used in the interpretation of the OR-Fork.
The choice construct allows multiple alternatives (more than two options) and this is reflected in the LOTOS code accordingly. When an OR-Fork occurs, we choose between the continuation of the sub UCM (ContinueCall in figure 36) and the path segment (Disconnection).

We also have the possibility to add guards to the all segments. Figure 37 illustrates the case where the subsequent actions depends on the Called party's (B) state.

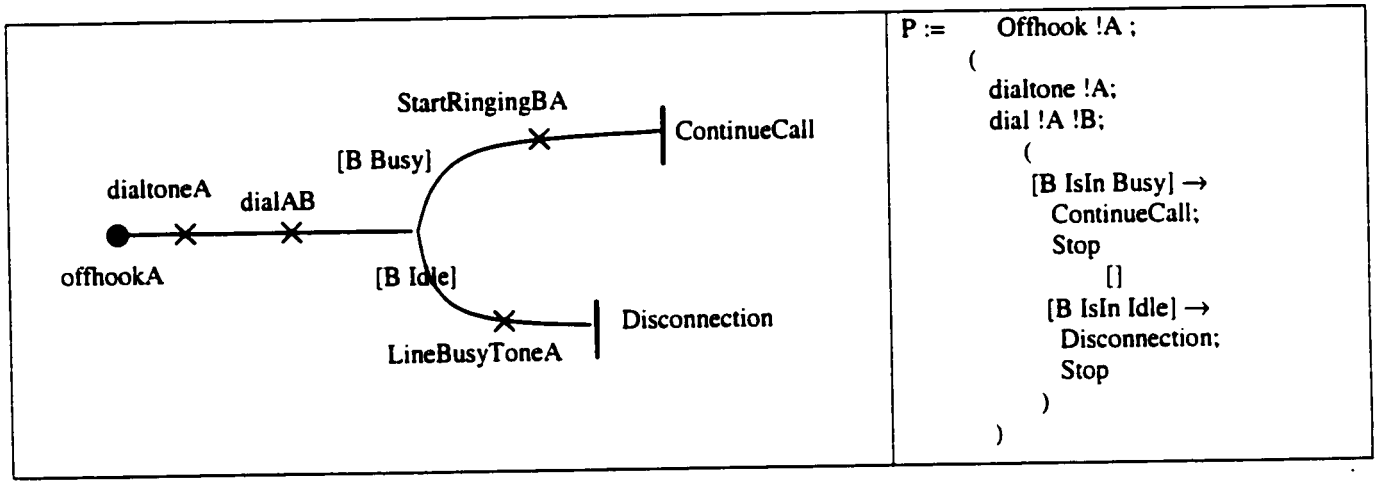


Figure 37: Guarded Behaviour

- **AND-Fork:**

Many actions could take place concurrently. In the next example (figure 38), when user A dials B's number, while B is Idle, two actions take place concurrently: "StartAudibleRinging AB" and "StartRinging BA". This is represented with two concurrent paths after an AND-Fork. The LOTOS interleaving operator is used here to represent that two tokens follow the two paths concurrently. The AND-Fork adds new concurrent path segments, and we may have more than two exiting paths (thus at least one new Path segment), without guards.

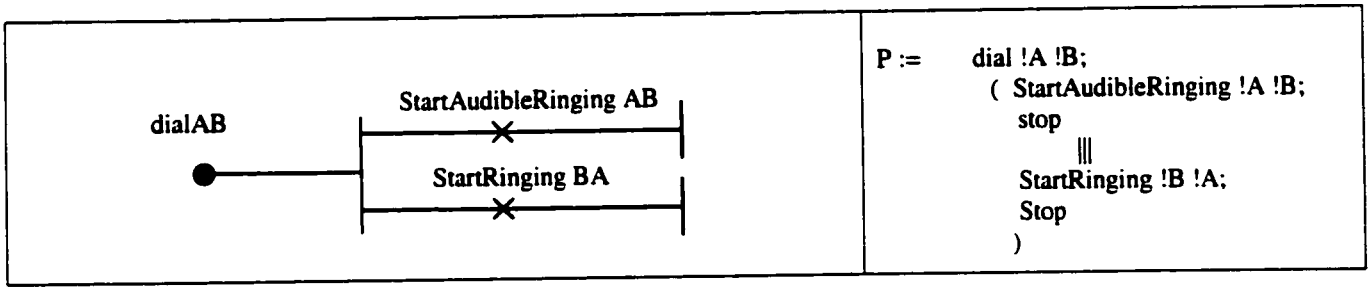


Figure 38: AND-Fork

- **OR-Join**

An OR-Join merges two (or more) overlapping paths. These are two exclusive paths that will join. Figure 39 presents such a (simplified) UCM. Again the LOTOS choice operator ([]) is used in the interpretation of the OR-Join. The choice construct allows multiple alternatives (more than two options) and this is reflected in the LOTOS code accordingly.

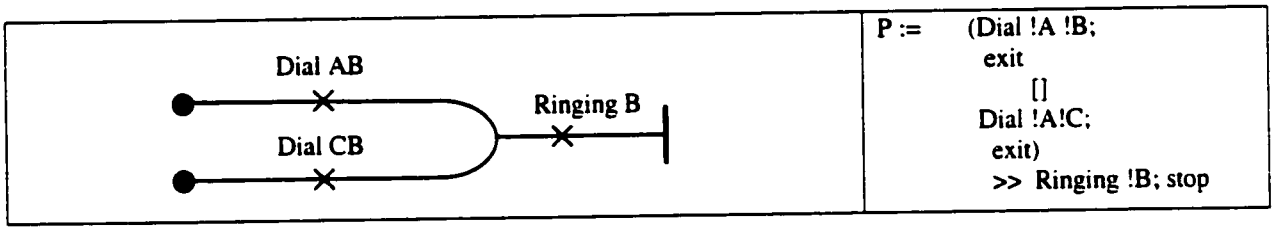


Figure 39: OR-Join

- **AND-Join**

Many actions could take place concurrently. For example, two actions take place concurrently: "StartAudibleRinging AB" and "StartRinging BA".

Again the LOTOS interleaving operator is used to represent that two tokens follow the two paths concurrently and synchronize on a further responsibility. The AND-join gives the possibility to add another path after concurrent responsibilities.

Figure 40 illustrates a case where the two events: "StartAudibleRinging AB" and "StartRinging BA" take place in either order and they are followed by the action "offhook B" (where the user picks the phone to answer the incoming call).

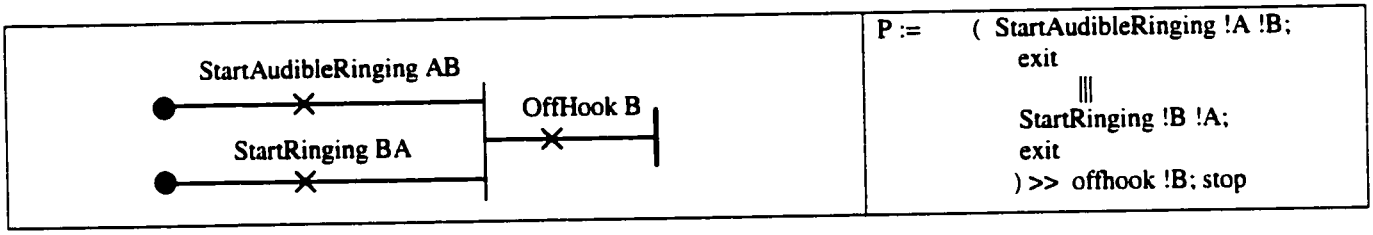


Figure 40: AND-Join

- Abstract data types are used to represent databases, operations, and conditions (LOTOS guard expressions)
- UCM components are represented as processes synchronized on their shared channels/gates.
- Components with stubs have sub-processes, one for each stub. The plug-in is mapped to LOTOS according to the rules defined above.

Note: This mapping is done manually.

5.5 The LOTOS Specification

In this section, we describe the LOTOS specification of services that were given in the FI contest by describing their main abstract data types and the structure of their specification.

5.5.1 Design of the Abstract Data Types

In order to perform the detection and verification mechanisms, we need first to determine the categories of data and the rules associated with them. Telephone systems operate with a limited set of data and rules: phone numbers, signals, features names, and databases of various types (Messages, Signals, PIN Number, Billing, Display).

Our specification is at a high level of abstraction since we are only interested in the observable behaviour of the system and not in the implementation details. However, these could always be refined when the implementation details become relevant.

Table 4 describes the main data types that have been specified; some of the specified data were not used for feature specification but are still required to complete the feature interaction detection procedure. Those serving for the detection procedure will be discussed in detail in Chapter 6.

Data Type	Meaning
Number	Defines the telephone numbers of the users
Feature	Defines the list of the features described in the specification
PIN	Defines the possible Personal Identification Numbers required for some features
Message	Defines the messages that can be played to the users such as AskForPIN
State	Defines the state of the numbers: Ringing, linebusy...etc.
ConnectionRecord	Defines the allowed and the forbidden connections
ConnectionRecord_set	Defines the set of connections
BillingRecord	Defines the billing data
BillingRecord_set	Defines the set of billing data
SignalRecord	Defines signals received by users
SignalRecord_set	Defines the set of Signals
Violation	Defines the different kinds of violation (e.g. violation of connections, Inconsistency of signals, violation In Billing...etc.

Table 4: The main Abstract data Types

- **Type Feature**

The type Feature describes the operations OCS, TCS, CFBL...etc, eq, ne and h. The operations OCS, TCS, CFBL...etc result in nine elements of the sort feature, "eq" and "ne" define respectively the equality and inequality between two features, and "*h*" is the operation that maps a feature into a natural number.

```
type Feature is NaturalNumber
sorts Feature
opns
        OCS,        (*! constructor *)  (* Originating Call Screening *)
        TCS,        (*! constructor *)  (* Terminating Call Screening *)
        CFBL,       (*! constructor *)  (* Call Forwarding Busy Line  *)
        CND,        (*! constructor *)  (* Call Number Delivery *)
        INFB,       (*! constructor *)  (* IN Free Billing *)
        INTL,       (*! constructor *)  (* IN Teen Line *)
        INFR,       (*! constructor *)  (* IN Free Routing *)
        INCF,       (*! constructor *)  (* IN Call Forwarding *)
        INCC,       (*! constructor *)  (* IN Call Charging *)  :→ Feature

        _eq_, _ne_              : Feature,Feature → Bool
        h                       : Feature → Nat

eqns
        forall F1, F2: Feature
        ofsort Bool
                F1 eq F2 = h(F1) eq h(F2);
                F1 ne F2 = h(F1) ne h(F2);

        ofsort Nat
                h(OCS)  = 0;
                h(TCS)  = Succ(0);
                h(CFBL)= Succ(Succ(0));
                h(CND) = Succ(Succ(Succ(0)));
                h(INFB) = Succ(Succ(Succ(Succ(0))));
                h(INTL) = Succ(Succ(Succ(Succ(Succ(0)))));
                h(INFR) = Succ(Succ(Succ(Succ(Succ(Succ(0))))));
                h(INCF) = Succ(Succ(Succ(Succ(Succ(Succ(Succ(0)))))));
                h(INCC)= Succ(Succ(Succ(Succ(Succ(Succ(Succ(Succ(0))))))));
endtype (* Feature *)
```

Note: Operations h, _eq_ and _ne_ are similar to operations to_nat, _eq_ and _ne_ described in detail in section 5.2.

The LOTOS specification of the remaining data types is provided in the Appendix.

5.5.2 Feature Specification

The stubs are represented as processes in LOTOS and sequences of them are represented by using the LOTOS operator enable ">>" between them (with "accept" if data has to be transferred to the following process)

```
Process Feature1[...]:=
P[...] >> Q[...] >> Z[...]
```

Process P, respectively Q, should contain at least one "exit" to fire the process Q, respectively Z. In addition to the fact that the features synchronize on the "exit" operator they

have to synchronize on their common gates and these gates could belong to different stubs. Considering this fact, we chose to flatten all the stubs of the UCM model. By doing this we solve the synchronization constraints and offer to the designer more flexibility to specify the features.

This is the LOTOS specification of Call Number Delivery obtained by a direct translation from the CND UCM defined in Chapter 4 section 4.4.2. Other LOTOS specifications of some telephony features are presented in the Appendix.

LOTOS specification of CND (Call Number Delivery):

Scenario where B subscribes to CND and A calls B:

```
process CND_feature[Offhook, DialTone, Onhook, Dial, StartRinging, StartAudibleRinging, StopRinging,
StopAudibleRinging, LineBusyTone, LogBegin, LogEnd, Display_number, SetLastDisplay, Disconnect,
Connection, Billing, Signal, Display](B_State: State) :noexit:=

Offhook ! A ;
Dialtone ! A ;
( onhook !A; stop
        []
Dial !A!B;
        (
        [B_State eq busy] →     LineBusyTone!A ;
                                Onhook!A ;stop
                []
        [B_State eq idle] →     StartRinging !B!A ;
                                Display_number!B!A;
                                SetLastDisplay !B!A;
                                StartAudibleRinging !A !B;

                                ( Onhook !A;
                                StopRinging !B!A;
                                StopAudibleRinging !A!B;stop

                                        [ ]
                                Offhook !B;
                                StopRinging !B !A;
                                StopAudibleRinging !A !B;
                                LogBegin !A!B!A;
                                (
                                Onhook !A;
                                Disconnect !B!A;
                                LogEnd !A!B;
                                Onhook !B;
                                stop
                                        [ ]

                                Onhook!B;
                                Disconnect !A!B;
                                LogEnd !A!B;
```

```
                                    Onhook !A;
                                    stop
                                    )
                            )
                    )
)
endproc
```

Chapter 6

Feature Interaction Detection Method

In this section, we propose a Feature Interaction Detection Method at design stage, based on the use of LOTOS, that uses extensively Abstract Data Types (ADT) to detect system inconsistencies.

6.1 Formal Definition of Feature Interaction

Logical interaction between two or more features occurs when one or some of the requirements or assumptions, that must be satisfied in the network, is violated.

Therefore, we develop a method based on expressing the feature requirements as properties and we say that interactions occur when these properties becomes contradictory, introducing inconsistencies in the system description, which is its formal specification.

We use a modified formal definition of "Feature Interaction" introduced in [53]:

Let S be an executable specification of a basic telephony system (POTS), and let F_1, F_2, ..., F_n, be specifications of n features.

We use $S \oplus F_1 \oplus F_2 \ldots \oplus F_n$ to denote the system obtained by integrating i features, $1 \leq i \leq n$, to the basic telephony system (POTS).

Let FP_1, FP_2, ..., FP_n (Feature Properties) be n formulas expressing respectively the feature properties of F_1, F_2, ..., F_n and let $N \models P$ denote that a system specification N satisfies formula P, i.e N is a model of P [53].

We say that there is an interaction between features F_1, F_2,..., F_n if:

$\forall i,\ 1 \leq i \leq n,\ S \oplus F_i \models FP_i$, but

$\neg (S \oplus F_1 \oplus F_2 \ldots \oplus F_i \models FP_1 \wedge FP_2 \wedge .. \wedge FP_n)$

We will usually consider the case where n=2 since most interactions reveal themselves in contexts where two features only are active [56].

Our feature interaction detection method doesn't address all possible kinds of interactions between features. We limit ourselves to the following types of interaction:

- Connection violations
- Inconsistency of signals given to users
- Incorrectness of Billing
- Inconsistency in the display function

Example of interaction: Connection violation between OCS and CFBL

Consider the following scenario: User A is an OCS subscriber and user C is in his screening list. User B is a CFBL subscriber and C is the number to which the incoming calls are forwarded when busy.

OCS property is : PFocs : $\neg$Connection(A,C)

CFBL property is: PFcfbl : Connection(A,C)

These two properties could not coexist in the same system, which is a feature interaction. How feature properties are derived is treated in detail further in this chapter.

6.2 Feature Interaction Detection Method

Our detection method will use feature descriptions from the user's point of view. This implies some very important advantages: Firstly, it can be applied at a very early stage of the design of new features. Secondly, the descriptions can be quickly created. Potentially, the feature specification will be created using UCMs. Each UCM will represent a single feature.

Combining the features and detecting possible interactions between them will be carried out automatically. Hence, the more complex task is not carried out by the designer but by some automatic detection mechanism and tools.
Figure 41 describes the feature interaction detection process.

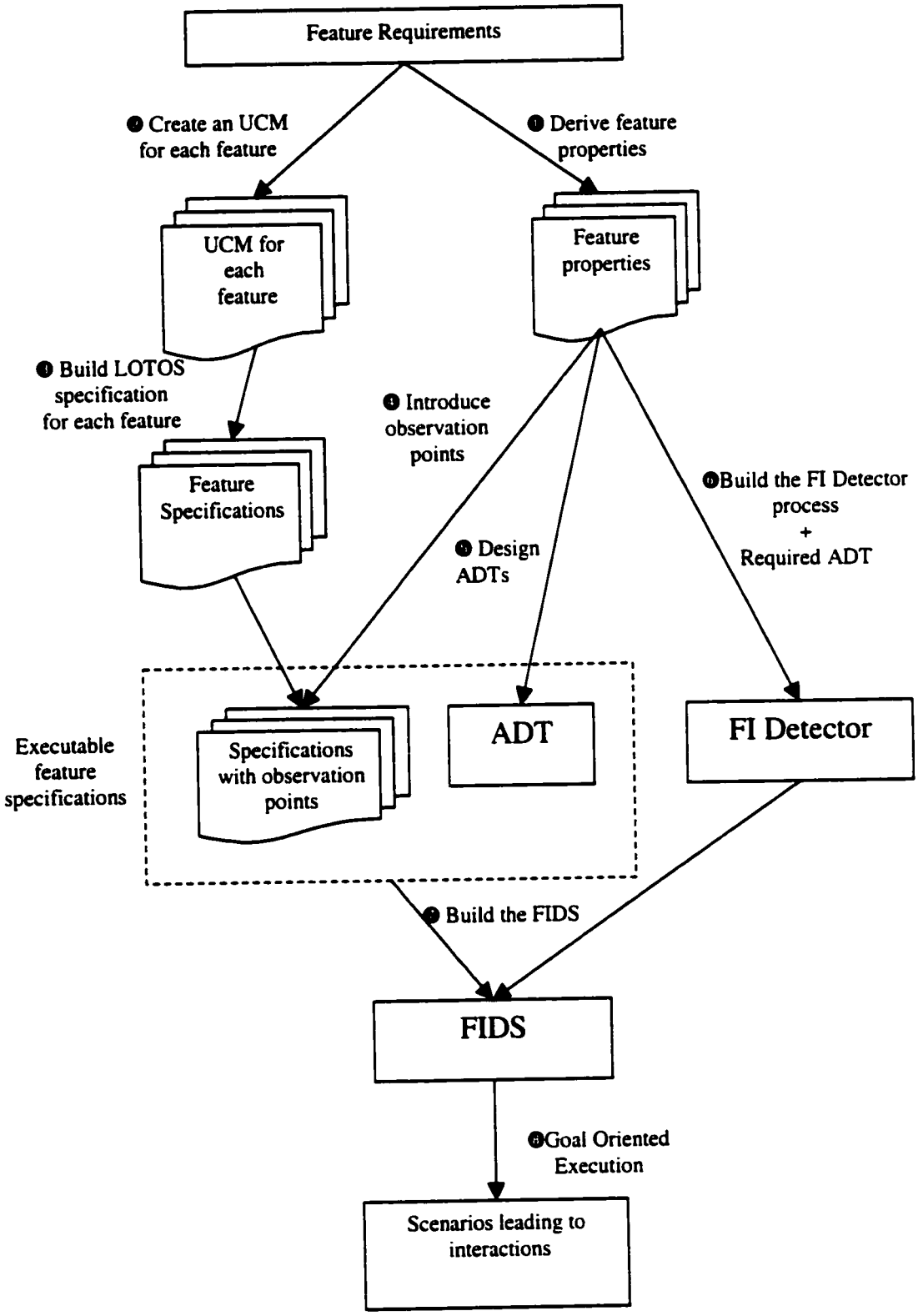


Figure 41: Feature Interaction Detection cycle

Given a set of feature requirements, described using natural language or visual notations such as "Chisel Diagrams", feature properties are derived (step 1) and a UCM for each feature is created (step 2); These UCMs are then mapped to LOTOS to obtain a LOTOS specification for each feature (step 3). Step 4 consists of introducing observation points into the specification for detection purposes. In order to run the LOTOS specification we need to design abstract data types (step 5). The Feature Interaction Detector (FID) process is built in step 6.
By combining together these three pieces: Feature Specifications + ADT + FID (step 7), we obtain the Feature Interaction Detection System (FIDS).
The last step (step 8) consists in running the FIDS and results in generating scenarios leading to interactions using the Goal Oriented Execution Tool.

These steps will be explained in detail in the following sections.

6.2.1 Deriving Features Properties

We believe that how to derive the properties of features and how to represent them are the key issues of a good feature interaction detection method.

Nowadays, communication services and features are commonly described using an amalgam of informal operational and declarative descriptions, tables, and visual notations such as Use Case Maps and Message Sequence Charts (MSCs) [33]. However using an informal representation can lead to different understandings of a given feature. For example, the informal requirements of feature INFB is "the IN Freephone Billing (INFB) feature allows the subscriber to pay for incoming calls." When deriving the properties of a feature from such definitions, divergences could occur in understanding the exact scope of "incoming calls". Is a forwarded call an "incoming" call? If it is, should the subscriber of INFB pay for the whole call or only for the forwarded part of the call?

We believe that the most clear and unambiguous manner to express feature properties is to use logic.

We choose to formalize a feature using declarative transition rules. Such rules consist of a precondition, a trigger event, and a post condition. The *pre* and *post* conditions of declarative transition rules are formulated in simple logic, which is a restriction of ordinary first order logic. The trigger event is also expressed in simple logic.

Declarative transition rule: Precondition $\xrightarrow{TriggerEvent}$ PostCondition

Features are described independently of other features. This means that the specification of a specific transition is done solely in consideration of the given feature. There is no consideration of potential interactions with another feature at this stage. This is important especially in the context of a multi-vendor environment where characteristics of a feature of a vendor may not be known to another vendor. Also, from a design point of view, this characteristic allows the feature provider to add new features without having to understand their behaviour in combination with other features.

The intented interpretations of the predicates should be clear. For instance Ringing (B,A) means that the telephone is ringing at B when A is calling B. The feature descriptions always depend on the basic service, to which actions like offhook(x), dial(x) and ringing(x) belong.

When a feature is activated its corresponding rule is fired. That is, the preconditions are met and the trigger event takes place. The post condition represents the resulting behaviour. We consider these post conditions as "**feature properties**".

As mentioned in the previous section, we are going to investigate only interactions dealing with connections establishment, Billing Records, Signal and Display. The derived feature's properties listed here are related to these four issues and are based on the best of our knowledge and on our practical experience with FI detection.

These are examples of feature rules and properties:

1. Call Number Delivery (CND) rule:

CND (B) $\xrightarrow{Ringing(B,A)}$ Display(B,A)

Where CND (B) means that B is a CND subscriber.

CND property is: Display (B,A), which means that A's number is displayed at B's side.

2. IN Freephone Billing (INFB) rule:

INFB (B) $\wedge$ Ringing (B,A) $\xrightarrow{offhook(B)}$ Billing (B,AB)

Where INFR (B) means that B is an INFB subscriber.

INFB property is: Billing (B, AB), which means that B is charged for the call leg between A and B.

3. Originating Call Screening (OCS) rule:

OCS (A) $\wedge$ InOCS_list(A,C) $\xrightarrow{dial(A,B)}$ $\neg$ Connection(A,C)

Where OCS (A) means that C is an OCS subscriber and InOCS_list(A,C) means that C is in the OCS screening list of A.

OCS property is: $\neg$Connection(A,C) which means that A and C should not be connected.

4. Call Forwarding Busy Line (CFBL) rule:

CFBL (B,C) $\wedge$ Busy (B) $\wedge$ $\neg$Busy (C) $\xrightarrow{dial(A,B)}$ Connection(A,C)

Where CFBL (B,C) means that B is a CFBL subscriber and C is the number to which the incoming calls are forwarded when he is busy. Busy(B) means that user B is busy and $\neg$Busy (C) means that user C is not busy.

CFBL property is: Connection(A,C) which means that A and C are connected.

6.2.2 UCM creation and Mapping to LOTOS

A UCM for each feature is created and translated to LOTOS using the rules introduced in Chapter 5. (Step 2 and 3 in figure 41)

6.2.3 Introducing observation points in the specification

In order to detect possible interactions we should provide the process FID (Feature Interaction Detector) with the relevant data during the execution of the specification. This data should be sent from the features to the FID. In order to do it, we introduce "**observation points**"

in the specification of each feature. These observation points are LOTOS gates with parameters (data). Each gate is responsible of carrying data leading to a different kind of Feature Interaction.

We distinguish four observation points:

- ***Connection***: Used to communicate information about the allowed and forbidden connections between users. This data is used to detect connection violation.

- ***Signal***: Used to communicate the signals received by users. This data is used to detect signal conflicts. For example, a user should not receive two different signals at the same time.

- ***Billing***: Used to communicate billing information. This data is used to detect incorrectness of billing information. For example, we could not have two different users charged for the same leg: Billing (A, AB) and Billing (B, AB).

- ***Display***: Used to communicate the display information. This data is used to control display violation. Example: an absence of display on a CND subscriber.

The FID and the features should synchronize on these gates. These detection points are introduced according to the feature properties.

In the following we discuss how to introduce the observation points in the two cases:

1) When the feature behaviour is part of the basic service behaviour (POTS: described as default plug-ins in Chapter 4)
2) And when the feature behaviour modifies the basic service behaviour (Not Default plug-ins).

6.2.3.1 Information related to connection

Depending on the two cases mentioned above, we provide two different rules to insert observation points dealing with connection.

- Feature behaviour is part of the basic service (default plug-ins):

When the phone starts ringing at the called party it means that the connection attempt succeeded. So the rule will be: Insert in the specification a LOTOS action called "Connection" immediately after a "StartRinging" gate with the parameters: the name of the feature, the two users involved in the connection and a boolean value indicating whether the connection is allowed or not.

Example: Consider the following LOTOS sketch of a specification of a feature F that has default behaviour.

```
Offhook !A;
Dial !A !B;
StartRinging! A !B;
Connection ! F !A !B !True;
...
```

- Feature behaviour modifies the basic service behaviour:

The insertion of the connection gate depends on the presence of the predicate connection in the feature rule post condition. When the feature rule has the predicate connection as post condition, the action "connection" is inserted after a state where the precondition is satisfied and the trigger event is fired.

For example, consider the TCS rule:

TCS (B), InTCS_list(A,B) $\xrightarrow{dial(A,B)}$ $\neg$ Connection(A,B)

This is a sketch of TCS specification:

```
...
Dial !A!B;
[A IsIn TCS_list] →     Connection !TCS !A !B !false;
                        PlayAnnouncement !A !ScreenedMessageTCS;...
        [ ]
[A NotIn TCS_list] → ...(* Default Behaviour *)
```

The connection gate is inserted after the "dial" event and at a state where A is in the screening list of B.

6.2.3.2 Information related to Billing

Depending on the two cases mentioned in section 6.2.3, we provide two different rules to insert observation points dealing with billing.

- Feature behaviour is part of the basic service (default plug-ins)

The billing starts after the LogBegin action. The rule is to insert in the specification the LOTOS action: "Billing" immediately after a "LogBegin" gate with the name of the feature, the charged party and the two users involved in the connection leg.

Example: Consider the following LOTOS sketch of a specification of a feature F which has a "default idle setup" plug-in.

```
...
StartRinging! A !B;
Offhook !B;
LogBegin !A !B !A;
Billing !F !A !B !A;
...
```

- Feature behaviour modifies the basic service behaviour:

The feature rule should have the predicate Billing as post condition. The rule is to insert the action "Billing" according to the feature rule i.e. we should look for a state where the precondition is satisfied and the trigger event is fired then we introduce the Billing gate.

For example consider INFB rule:

INFB (B) $\wedge$ Ringing (B,A) $\xrightarrow{offhook(B)}$ Billing (B,AB)

This is a sketch of the INFB specification:

```
...
StartRinging! A !B;
Offhook !B;
LogBegin !A !B !B;
Billing !A !B !B;
...
```

Note: Only the Billing parameters are modified with respect to the The INFB property.

6.2.3.3 Information related to Signals

For signals there is no distinction between default and specific behaviours.

Within the specification insert a LOTOS action: "Signal" immediately after the reception of a signal by a user. The "Signal" gate has the following parameters: The feature name, the signal, user who receives the signal and the other party involved in the signal.

Example: Consider the following LOTOS code from a specification of a feature F:

```
...
StartRinging! A !B;
Signal !F !StartRinging !A !B
...
```

6.2.3.4 Information related to Display

Depending on the two cases mentioned in section 6.2.3, we provide two different rules to insert observation points dealing with display.

- Feature behaviour is part of The basic service (default plug-ins):

Insert in the specification the LOTOS action: "Display" after a "StartRinging" gate with the following parameters: the name of the feature, the called party, the user initiating the call and a boolean value indicating whether there is a display or not. For the basic service, this boolean value is set up to false.

Example Consider the following LOTOS code from a specification of a feature F:

```
Offhook !A;
Dial !A !B;
StartRinging! A !B;
Display ! F !B !A !False;
...
```

- Feature behaviour modifies the basic service behaviour:

The feature rule should have the predicate "Display" as post condition. The rule is to insert the action "Display" according to the feature rule i.e. we should look for a state where the precondition is satisfied and the trigger event is fired then we introduce the Display gate.

For example consider CND rule:

CND (B) $\xrightarrow{Ringing(B,A)}$ Display(B,A)

This is a sketch of the CND specification:

```
...
StartRinging! A !B;
Display !CND !B !A !True;
...
```

6.2.4 Design of the Abstract Data Types

For implementation purposes we have chosen to use LOTOS ADTs to detect interactions. We thus define an operation "CausesViolation" having a Boolean value that indicates whether a violation occurs or not.

These are the abstract data types used to control the detection:

6.2.4.1 Storing Connection information

To store connection information, we defined two types: ConnectionRecord and ConnectionRecord_set.

- **ConnectionRecord**

A connection could take place between two users. The connection record is composed of: the feature name, the caller party number, the called party number and a boolean value to determine whether the connection is allowed or not.

The following is the LOTOS ADT description of ConnectionRecord type with interleaving comments:

```
type ConnectionRecord is Number,Boolean
sorts ConnectionRecord
opns (* specify the format of operations *)
```

"ct" is the constructor of the connectionRecord.

```
ct: Feature, Number, Number, Bool → ConnectionRecord
```

"num1" and "num2" are two operations used to extract the user numbers involved in the connection.

num1 : *ConnectionRecord* → *Number*
num2 : *ConnectionRecord* → *Number*

Bvalue is an operation for extracting the boolean value of a connectionRecord: if the connection is allowed then Bvalue returns true otherwise it returns false.

Bvalue : *ConnectionRecord* → *Bool*

eq, and *ne* are two operations used to compare two connectionRecord.

*_eq_, _ne_* : *ConnectionRecord, ConnectionRecord* → *Bool*

```
eqns (* List of equations *)

        forall   N1, N2, N3, N4   :number,
                 t1, t2           :ConnectionRecord,
                 B1, B2           :Bool
                 F, F1, F2        :Feature

        ofsort Number
                 num1(ct(F,N1,N2,B1)) = N1;
                 num2(ct(F,N1,N2,B1)) = N2;

        ofsort Bool
                 Bvalue(ct(F,N1,N2,B1)) = B1;
        ct(F1,N1,N2,B1) eq ct(F2,N3,N4,B2) = ((F1 eq F2) and (N1 eq N3) and (N2 eq N4) and (B1 eq B2))
                                                        or
                                                ((F1 eq F2) and (N1 eq N4) and (N3 eq N2) and (B1 eq B2));

        ct(F1,N1,N2,B1) ne ct(F2,N3,N4,B2) = not( ct(F1,N1,N2,B1) eq ct(F2,N3,N4,B2));

endtype (* ConnectionRecord *)
```

- **ConnectionRecord_set**

  This set contains all the connections involved in the communication during the execution of the specification. Connections records could be inserted into the connectionRecord_set either during the initialization part, or during the execution of the specification.

The following is the LOTOS ADT description of ConnectionRecord_set type:

type *ConnectionRecord_set is ConnectionRecord*
sorts *ConnectionRecord_sets*
opns

This operation defines an empty set.

{} : → ConnectionRecord_sets

The operation "insert" inserts a new connection in the connection set

insert : *ConnectionRecord,ConnectionRecord_sets → ConnectionRecord_sets*

The operations *_eq_* and *_ne_* are used to compare two ConnectionRecord_set.

*_eq_, _ne_* : *ConnectionRecord_sets, ConnectionRecord_sets -> Bool*

The operation "empty" checks whether the connection set is empty or not.

empty : *ConnectionRecord_sets → Bool*

The operation "isin" checks whether the connectionRecord is present in the connectionRecord_set or not.

isin : *ConnectionRecord, ConnectionRecord_sets → Bool*

The operation "CausesViolation_Connection" checks whether a new connection record is inconsistent with the existing data in the ConnectionRecord set.

CausesViolation_Connection : *ConnectionRecord, ConnectionRecord_sets → Bool*

```
Eqns (* List of Equations *)

    forall  t1,t2          : ConnectionRecord,
            s              : ConnectionRecord_sets,
            n1,n2,n3,n4    : Number,
            b1,b2          : Bool
            F, F1, F2      : Feature

ofsort ConnectionRecord_sets
    tail (insert(t1,s)) = s;

ofsort Bool
    {} eq {} = true;
    {} eq insert(t1,s) = false;
    insert (t1,s) eq {} = false;
    empty (s) = s eq {};
    isin (t1,{}) = false;
    t1 eq t2 => isin(t1,insert(t2,s)) = true;
    t1 ne t2 => isin(t1,insert(t2,s)) = isin(t1,s);
```

A new connectionRecord is not in conflict with an empty set of connectionRecord.

CausesViolation_Connection(ct(F, n1,n2,b1),{}) = false;

The operation "CausesViolation_Connection" checks if the insertion of a new connection record conflicts with existing constraints. A connection conflict occurs in two cases:

1. When one feature allows a connection, while another feature deny it. This is described by the following two conditions:

CausesViolation_Connection(ct(F1, n1,n2,b1),insert(ct(F2, n3,n4,b2),s)) =
((F1 ne F2) and (n1 eq n3) and (n2 eq n4) and (b1 ne b2)) ,
or
((F1 ne F2) and (n1 eq n4) and (n2 eq n3) and (b1 ne b2))
or

For example: suppose that the user C is a terminating call screening subscriber and has A in his screening list. The connection set should contain the connection record (TCS, A, C, false). During the execution of the specification, if a connection is established between users A and C (for some reason) the system tries to add the connection record (F, A, C, true) to the connection set. The addition of such connection causes a conflict, which is reported to the user.

2. When two features cause the connection of one user to two different parties. This is described by the following two conditions:

((F1 ne F2) and (n1 eq n3) and (n2 ne n4) and (b1 eq true) and (b2 eq true))
or
((F1 ne F2) and (n1 ne n3) and (n2 eq n4) and (b1 eq true) and (b2 eq true))
or
CausesViolation_Connection(ct(F1, n1, n2, b1), s);

endtype

6.2.4.2 Controlling signals

We defined two types:SignalRecord and SignalRecord_set to store signals sent to the involved users.

- **SignalRecord**

The SignalRecord is composed of the name of the process (the feature) originating the signal, the user receiving the signal, the other party involved in the signal if any.

Example: (CND, StartAudibleRinging, A, B) and (OCS, LineBusyTone, A, none)

The following is a sketch of the LOTOS ADT description of SignalRecord type:

type SignalRecord is Feature, Number, State, Boolean
sorts SignalRecord
opns

"st" is the constructor of the signalRecord.

st : Feature, State, Number, Number → SignalRecord
....
endtype (* SignalRecord *)

- **SignalRecord_set**

This set controls the signals received by the involved users. It is updated during the execution of the specification.

The following is the LOTOS ADT description of SignalRecord_set type:

```
type SignalRecord_set is SignalRecord
sorts SignalRecord_sets
opns
        ....
```

The operation "CausesViolation_Signal" checks if the insertion of a new signalRecord conflicts with existing signals.

```
        CausesViolation_Signal          : SignalRecord, SignalRecord_sets → Bool
eqns
 forall t1,t2            : SignalRecord,
        F1,F2            : Feature,
        S1,S2            : State,
        S                : SignalRecord_sets,
        n1,n2,n3,n4      : Number
        b1,b2            :Bool

        ....
        ofsort Bool
```

A new signalRecord is not in conflict with an empty set of signalRecord

```
        CausesViolation_Signal(st(F1,S1,n1,n2),{ }) = false;


        CausesViolation_Signal (st(F1,S1,n1,n2),insert(st(F2,S2,n3,n4),s)) =
                ((F1 ne F2) and (S1 ne S2) and (n1 eq n3))
                        or
                ((F1 ne F2) and (S1 eq S2) and (n1 eq n3) and (n2 ne n4))
                        or
                CausesViolation_Signal(st(F1,S1,n1,n2),s);

endtype (* SignalRecord_set*)
```

The operation "CausesViolation_Signal" checks if the insertion of a new signalRecord conflicts with existing signals. We consider two possible inconsistencies between signals:

❶ A user receives two different signals

SignalRecord1: (F1, S1, A, B): Feature 1 generates a signal S1 to user A caused by user B.

SignalRecord2: (F2, S2, A, C): Feature 2 generates a signal S2 to user A caused by user C.

User A receives two different signals: S1 and S2 generated from two different features.

This inconsistency is detected in the condition: ((F1 ne F2) and (S1 ne S2) and (n1 eq n3))

Example:

SignalRecord 1: (F1, StartAudibleRinging, A, B)

SignalRecord 2: (F2, LineBusyTone, A, none)

The user A receives two different signals: "Start Audible Ringing" and "LineBusyTone" which denotes a feature interaction.

❷ A user receives the same signal but generated by two different users

SignalRecord 1: (F1, S, A, C)

SignalRecord 2: (F2, S, A, D)

The user A receives the same signal S but caused by different users.

This conflict is detected in the condition: ((F1 ne F2) and (S1 eq S2) and (n1 eq n3) and (n2 ne n4))

Example:

SignalRecord 1: (F1, StartAudibleRinging, A, C)

SignalRecord 2: (F2, StartAudibleRinging, A, D)

User A receives the same signal "Start Audible Ringing" but the involved users are different (C and D).

Note: The full description of types SignalRecord and SignalRecord_set is provided in Appendix.

6.2.4.3 Storing Billing information

To store Billing information, we defined two types: BillingRecord and BillingRecord_set.

- **BillingRecord**

For each connection leg there is a billing record containing the parties involved and the charged party. The BillingRecord is composed of the feature name, the caller party number, the called party number and the charged number.

The following is a sketch of the LOTOS ADT description of the BillingRecord type:

```
type BillingRecord is Feature, Number
sorts BillingRecord
opns
```

"bt" is the constructor of the BillingRecord.

```
        bt                          : Feature, Number, Number, Number → BillingRecord

....

endtype (* BillingRecord *)
```

- **BillingRecord_set**

This set contains the system billing records. The following is a sketch of the LOTOS ADT description of BillingRecord_set type.

```
type BillingRecord_set is BillingRecord
sorts BillingRecord_sets
opns
                ....
```

The operation "CausesViolation_Billing" checks if the insertion of a new Billing Record conflicts with existing Billing Records.

```
        CausesViolation_Billing  : BillingRecord, BillingRecord_sets → Bool

eqns

 forall  t1,t2 : BillingRecord,
         s: BillingRecord_sets,
         bts1,bts2: BillingRecord_sets,
         n1,n2,n3,n4,n5,n6:Number,
         b1,b2:Bool,
         F1,F2: Feature

         ....
         ofsort Bool
```

A new BillingRecord is not in conflict with an empty set of BillingRecord

```
                CausesViolation_Billing(bt(F1,n1,n2,n3),{ }) = false;
```

The operation "CausesViolation_Billing" checks if the insertion of a new Billing Record conflicts with existing Billing Records. The conflict occurs when two users are billed for the same connection leg. For example (F1,A, B, A) and (F2, A, B, B) are two conflicting Billing records where A and B pay for the same leg.

```
CausesViolation_Billing(bt(F1,n1,n2,n3),insert(bt(F2,n4,n5,n6),s)) =
((F1 ne F2) and (n1 eq n4) and (n2 eq n5) and (n3 ne n6)) or
((F1 ne F2) and (n1 eq n5) and (n2 eq n4) and (n3 ne n6)) or
CausesViolation_Billing( bt(F1,n1,n2,n3),s);
```

endtype

Note: The full description of types BillingRecord and BillingRecord_set is provided in Appendix.

6.2.4.4 Display information

We defined two types: DisplayRecord and DisplayRecord_set.

- **DisplayRecord**

The DisplayRecord is composed of the feature name, the called party number, and a boolean value indicating whether the display occurred or not.

The following is the LOTOS ADT description of DisplayRecord type:

type DisplayRecord is Number,Boolean
sorts DisplayRecord
opns (* specify the format of operations *)

"bt" is the constructor of the BillingRecord

dt: Feature, Number, Bool → DisplayRecord

...

endtype (* DisplayRecord *)

- **DisplayRecord_set**

This set contains the display records. Display records could be inserted into the DisplayRecord_set either during the initialization part of the data types, or during the execution of the specification.

The following is the LOTOS ADT description of DisplayRecord_set type:

type DisplayRecord_set is DisplayRecord

```
sorts DisplayRecord_sets
opns

(* Test if a connection can cause violation within the connection set *)
        CausesViolation_Display  : DisplayRecord, DisplayRecord_sets → Bool

Eqns  (* List of Equations *)

        forall   t1,t2              : DisplayRecord,
                 s                  : DisplayRecord_sets,
                 n1,n2,n3,n4        : Number,
                 b1,b2              : Bool
                 F1, F2             : Feature

ofsort Bool
        ...
A new DisplayRecord is not in conflict with an empty set of DisplayRecord
        CausesViolation_Display (dt(F1, n1, n2, b1),{ }) = false;
```

The operation "CausesViolation_Display" checks if the insertion of a new display record conflicts with existing constraints. This conflict happens when a user didn't get a display where he is supposed to get one and vice versa. This is an example of two conflicting display records: (F1, A, B, True) and (F2, A, B, False).

```
        CausesViolation_Display (dt(F1, n1,n2,b1),insert(dt(F2, n3,n4,b2),s)) =
        ((F1 ne F2) and (n1 eq n3) and (n2 eq n4) and (b1 ne b2))
                or
        ((F ne F2) and (n1 eq n4) and (n2 eq n3) and (b1 ne b2))
                or
        CausesViolation_Display (dt(F1, n1,n2,b1),s);

endtype
```

6.2.5 Feature Interaction Detection System Architecture

As mentioned in Section 5.3.4, when several processes are combined together by means of the parallel composition "|[gates]|", if an action is in the list of gates in the operator then in order for that action to execute, all processes must participate simultaneously (synchronize) on that action. Further, each process can provide its own conditions for the actions to execute, and all such conditions must be true simultaneously in order for this to happen. Thus the control we want to obtain can be achieved gracefully by using an independent control process in parallel with the system specification. This process is "FI Detector" and will synchronize with the feature specification on the four defined detection points.

The top level of the behaviour part of our specification (shown in figure 42) consists of three processes: Process Feature1, Process Feature2 and Process FI Detector.
Feature 1 and Feature 2 are the LOTOS specifications of the two features involved, obtained by direct mapping from the UCMs of these two features. These two processes are composed in parallel and synchronize through their common gates, except those representing signals.

Note: We discarded gates representing signals from the set of synchronization gates to be able to execute those actions independently so we can detect signal conflicts.

The process "FI Detector" is the one responsible of detecting the feature interactions between the two features.

Features1 and Feature2 communicate with the "FIDetector" via specific synchronization gates. The two processes use these gates in order to send relevant data to FIDetector. The FIDetector checks for system consistency, notifies the environment if such interaction occur and re-instantiates itself to continue the FI detection. Figure 42 describes the top level of the specification.

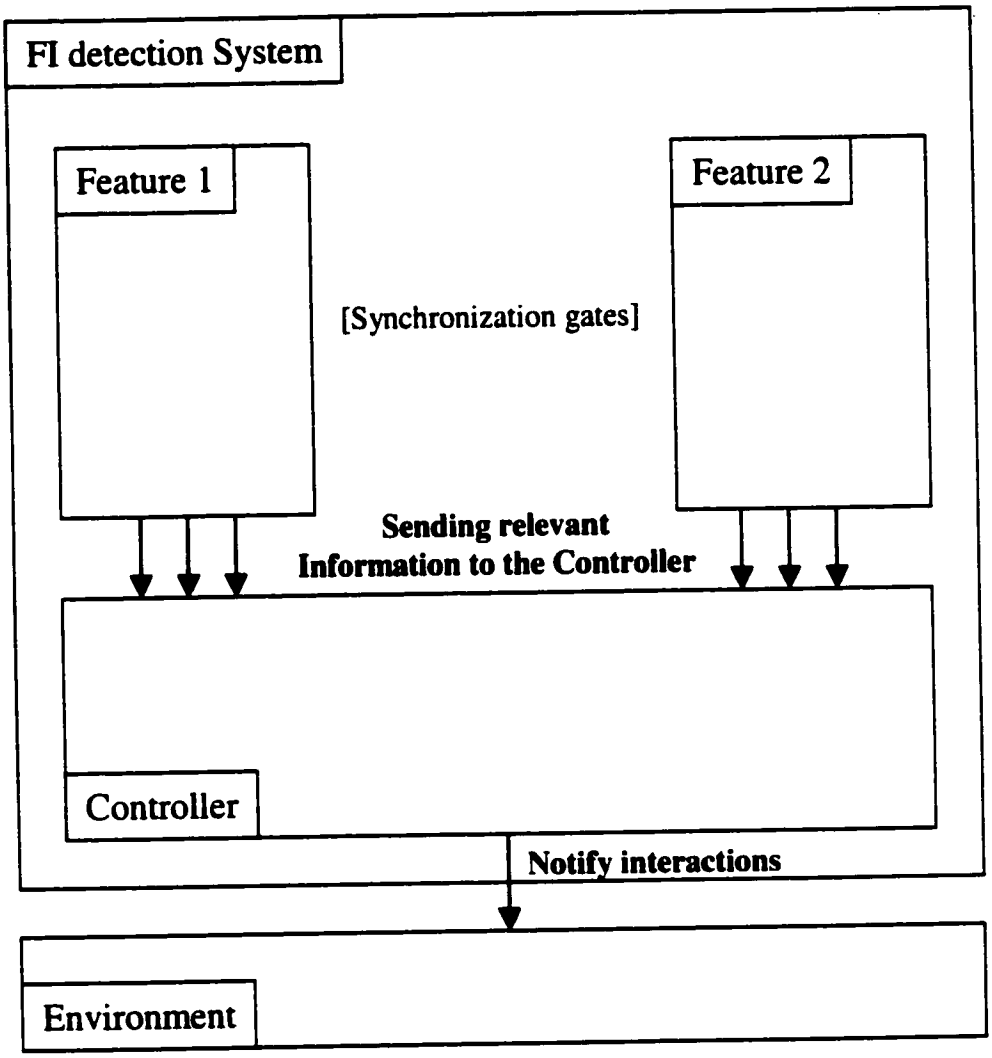


Figure 42: Graphical representation of the top-level of the specification

The specification below represents LOTOS top-level behaviour of the FI Detection System:

```
Specification FIDetection_System [offHook,dial,.....]: noexit:=
(* ...Abstract data type definitions... *)
behavior
(
        (
        feature1 [Offhook, DialTone, Onhook, Dial, Connection, Billing, Signal...](B_State...)

                |[Offhook, DialTone, Onhook, Dial,...]|

        feature2 [Offhook, DialTone, Onhook, Dial, Connection, Billing, Signal] (B_State,...)
        )
                |[Connection, Billing, Signal, Display]|

        FIDetector[Connection, Billing, Signal,VR,...](Connection_set, Billing_set, Signal_set...)
)
```

6.2.6 Design of the process FI Detector

The role of the process FI Detector is to gather the relevant information from the features via the observation points and detect the interactions that could occur during the execution of the specification of the two features.

The FI Detector waits for actions to appear on the four observation gates discussed in Section 6.2.3. If an identical action has already been detected, nothing is done. If it is a new action, it checks whether a conflict has been created according to the method discussed in Section 6.2.4.

This is a sketch of the LOTOS code for process FI Detector that corresponds to the algorithm in figure 43. The LOTOS code describes only the detection of connection conflicts. Conflicts in signals, billing and display are similar to the connection conflict. Comments are interleaved with the code.

process FIDetector [Connection, Billing, Signal, Display, VR]
(Connection_set:ConnectionRecord_sets, Billing_set:BillingRecord_sets,
Signal_set:SignalRecord_sets, Display_set:DisplayRecord_sets) :noexit:=

Step 1: FI Detector is waiting for actions on gates: Connection, Signal, Billing and Display to appear.

(Connection ? f: Feature ? x:Number ? y:Number ?b:Bool;

Step 2: Data is received on the connection gate. We call "data" the parameters of actions. In this example above, these are the values of f, x, y and b.

Step 3: The operation "IsIn" checks if the data received is already in the database.

- Data already exists in the database: FI Detector re-instantiates itself and goes back to step 1 to wait for new data.

[IsIn (ct (f, x, y, b), Connection_set)]-> FIDetector[Connection, Billing, Signal, Display, VR]
(Connection_set, Billing_set, Signal_set, Display_set)
[]

- Data doesn't exist in the database: Check if the new data conflicts with existing data in the database. The operation "CausesViolation_Connection" discussed in Section 6.2.4.1 detects such violations.

```
  [not(IsIn(ct(f, x, y, b), Connection_set))]->
(       [CausesViolation_Connection(ct(f, x, y, b), Connection_set)] ->
```

If the new data is inconsistent with existing data, the violation is notified to the environment by executing the specific action VR. Then FI Detector re-instantiates itself and goes back to step 1 to wait for new data.

```
        VR! ViolationOfIntentions;
        FIDetector[Connection, Billing, Signal, Display, VR] (Connection_set, Billing_set, Signal_set, Display_set)
                []
```

If the new data is consistent with existing data, the new data is inserted into the database and FI Detector is re-instantiated.

```
        [not(CausesViolation_Connection(ct(f, x, y, b), Connection_set))] ->
        FIDetector[Connection, Billing, Signal, Display, VR] (insert(ct(f, x, y, b), Connection_set), Billing_set,
Signal_set, Display_set)
    )
  )
)
[ ]
```

Same procedure for detecting Signal conflicts.

```
(Signal ? f: Feature ? x: signal ? y: Number ? z: signal;
...
[ ]
...
```

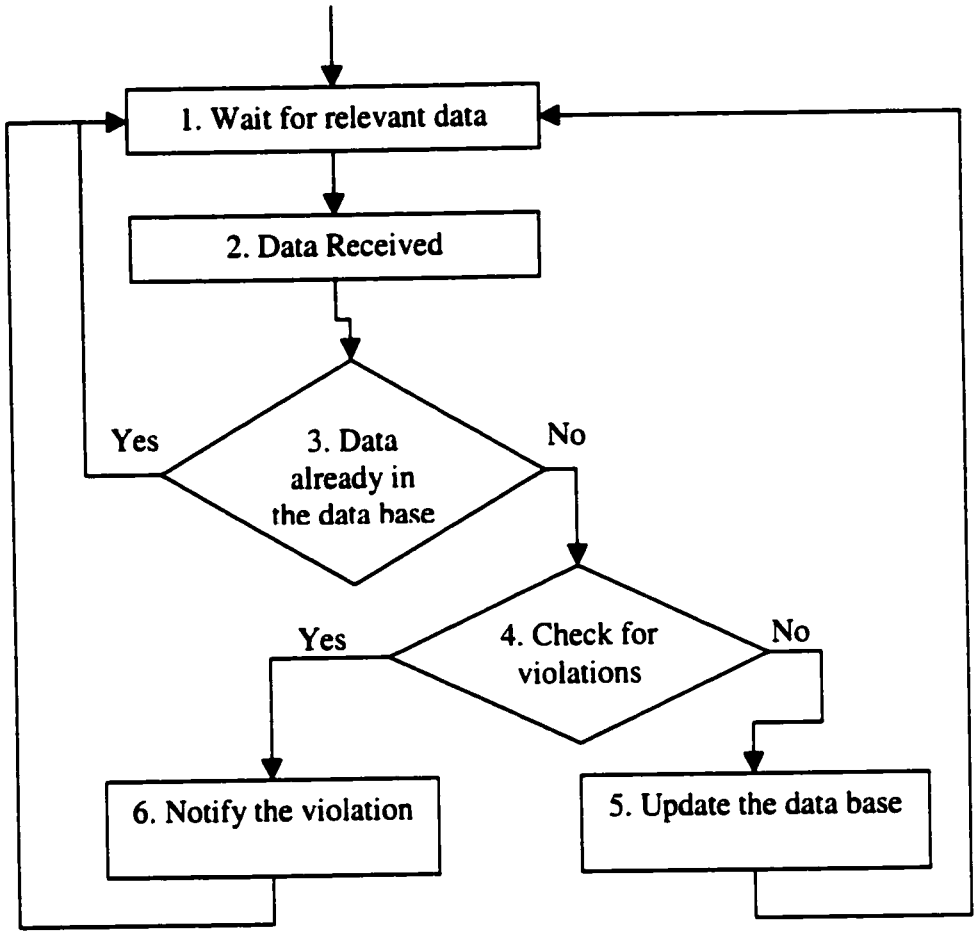


Figure 43: FI Detector procedure

6.3 Detection of Interactions

So far we have seen which interactions are considered in this work and how they emerge. Mechanisms for detection must now be considered. Our method works by executing a formal specification (which could be called also formal prototype or formal model) of the system with features. The ADTs included in the specification for this purpose will check execution to see whether a violation has occurred.

The ADT's represent a kind of a central call model monitor that captures every event associated with a call and verifies whether a violation has occurred or not.

In Chapter 4 we have filtered the interaction free scenarios from the relevant possible scenarios. However the verdict obtained was either "FI could occur" or "FI occurs".

The proposed mechanism is applied to investigate the cases having "FI could occur" as verdict and to characterize the interactions for those having "FI occurs" as verdict.

As seen in the filtering process two relevant cases should be considered:

- **Case 1**: Scenarios where the features are distributed between two users A and B.
- **Case 2**: Scenarios where a third party C is the feature subscriber and B arranges to forward/route the call to C.

For case 1 the UCM features are mapped directly to LOTOS. So we use the system as described above. However for case 2 only the terminating part (for user C) interests us. And to be able to detect the feature interactions we should map the UCM refinement proposed in the Chapter 4 into LOTOS. So we obtain a partial specification of the feature representing only the terminating part.

The system starts by executing the feature 1 (where A or B is subscribed to this feature) until the call is forwarded to C. At that time the specification describing the feature 2 (where C is subscribed) starts. This mechanism is shown on figure 44.

Note: Since the two features synchronize on their common gates this doesn't prevent the system from achieving its goal.

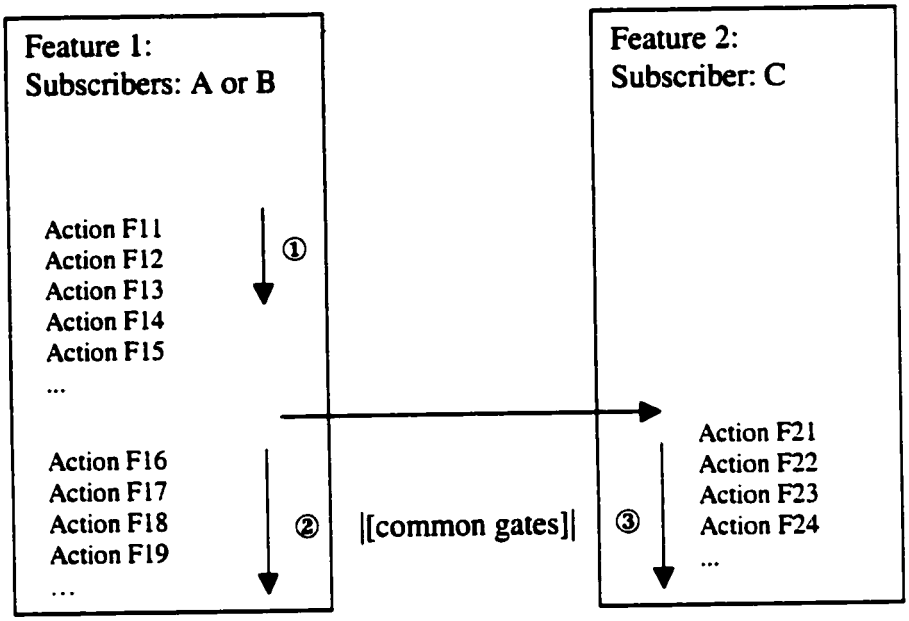


Figure 44: System with a third party as a feature subscriber

Example: Figure 45 illustrates such scenario between TCS and CFBL. B subscribes to CFBL and C subscribes to TCS. The CFBL process is executed first since TCS is blocked waiting for the action StartRinging_fwd to occur. A goes offhook and dials B's number. B is busy so the call is forwarded to C. Once the call is forwarded to C and the phone at C starts ringing the two processes start synchronizing on their common gates.

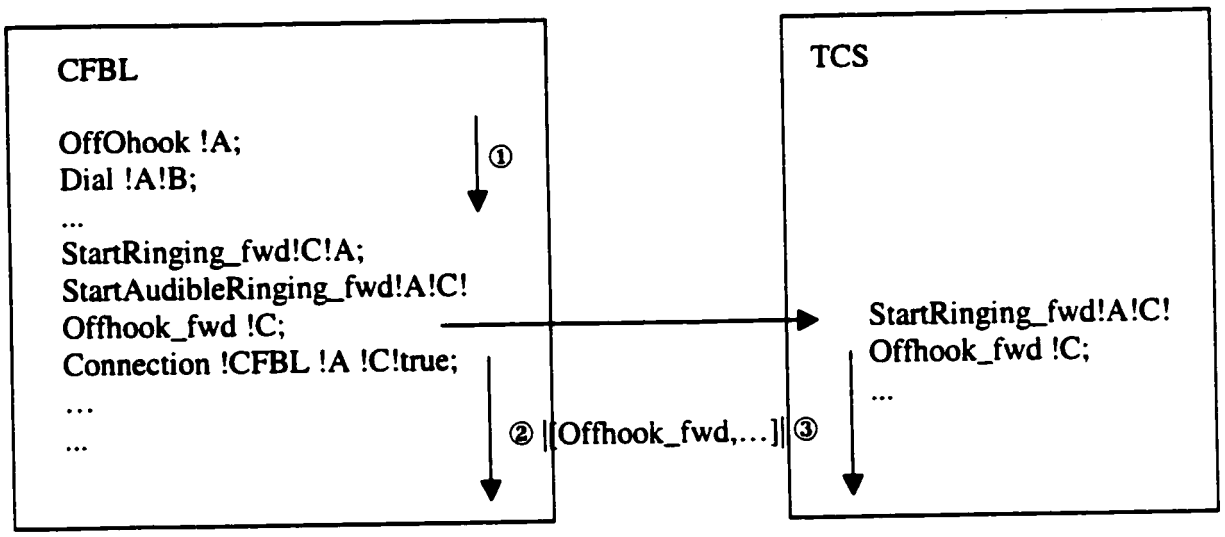


Figure 45: Example of a scenario involving TCS and CFBL

6.4 FI Scenario generation: Goal Oriented Execution (GOE)

In this step, the goal-oriented tool is applied to generate the traces leading to interactions. We call trace a sequence of observable actions that a LOTOS process can offer to the environment.

The method is based on the goal oriented execution tool developed within the LOTOS group of the University of Ottawa [46], [47]. Goal-oriented execution allows one to look for execution traces according to several properties. In this type of execution, the user specifies an action to be reached, usually an action that is not immediately derivable. The system then proceeds in a sort of selective eager execution, being able to select traces likely to reach the action. These traces can be found with the help of a static analysis of the behaviour expression. For example, if the behaviour expression is:

(a ; b ; stop ||| b ; c ; stop) [] c ; d ; f ; stop

and the user wants to be given an (or all) execution trace(s) reaching f, then the goal oriented execution algorithm is able to see that the left-hand side of the behaviour expression does not need to be expanded at all, because it does not contain action f. A considerable saving in computing time and space is obvious from the example.

Goal oriented execution can be used to find sequences of actions corresponding to certain criteria. The system proceeds to select traces that contain this sequence starting by the first action in the sequence. For example, if the behaviour expression is:

(a ; c ; stop [] a ; e ; c ; stop [] a ; stop)
and if the sequence of actions to be reached is: [a, c], then the possible traces that can be selected by the system are a ; c and a ; e ; c.

Events can be associated with actions in the sequence defining the goal to be reached. For example, if the sequence contains an action with gate name "a" and an offer value (!) x1, the selected trace must contain the action with gate a and with the offer of value x1. If the event associated with the action is an accept of a parameter (?), the system will instantiate all the possible values of that parameter.
An example of a goal to be reached is the following:
[a !x1 ?x2 , b~, c] \\ [e, f].

This goal is satisfied by all traces including a sequence of actions starting by an action with gate name a, with an offer of the value x1 and an acceptance of the value x2, leading to the action represented by gate c (without any event), and having as intermediate action an action with gate name b with an arbitrary event (~). Traces must not include actions with gates "e" and "f".

In addition, the tool has many characteristics that can speed up the search. In fact, the search can be guided by the user by setting limits for the number of instantiations of processes and by avoiding some branches of the corresponding LTS.

If some processes are instantiated recursively, leading to infinite LTS (Labelled Transition System: a notation where behaviours are represented by edges), GOE cannot guarantee the absence of a trace corresponding to the specified goal, because the tool limits the number of instantiations of processes.

Note that, after finding a trace, the tool will ask the user whether another one is desired. To accelerate the search we can always guide it by adding some intermediate actions that we know must exist in a trace satisfying the specified goal. For example, if we are looking for a trace leading to an action specifying a connection establishment between two network users, we can add in the goal an action where a user dials a number, since it is evident that before an establishment of a connection, a user must dial a number. If we want to exclude the search from

some branches of the behaviour tree, we can add some specific gates and exclude them from the search. Then, the search process will not go in those branches where these specific gates are inserted.

In our work, we apply Goal Oriented Execution in order to obtain traces containing the gate "VR". Those traces lead to a feature interaction.

These traces (excluding observation gates) are used as test suites to test pairwise feature interaction within the implementation.

6.5 Application

In the following we study the interactions between OCS, CFBL and INFB. The LOTOS specifications are described in the Appendix.

6.5.1 Interactions between OCS and CFBL

This is an example of scenario leading to interactions between OCS and CFBL:
User A subscribes to OCS with C within his screening list. User B is a CFBL subscriber.
B is busy, A calls B so the call is forwarded to C.
The trace obtained with the goal-oriented tool is:
Data is received on the observation gate "connection" indicating that feature OCS doesn't allow connections between A and C.

```
Connection  ! OCS ! A !C ! false;
Offhook  ! A;  (* synchronization between CFBL and OCS *)
Dialtone  ! A;  (* synchronization between CFBL and OCS *)
Dial  ! A ! B;  (* synchronization between CFBL and OCS *)
LineBusyTone ! A; (* From OCS *)
```

A LineBusyTone signal is received on the observation gate "signal" indicating that this signal is received by A and is caused by OCS.

```
Signal ! OCS ! LineBusyTone ! A ! none;
detect_forward  ! C; (*from CFBL*)
```

Data is received on the observation point "connection" indicating that feature CFBL allows the connection between A and C.

```
Connection !CFBL! A !C !true;
```

The new connection information causes a conflict with the existing connection information. The conflict is reported to the environment.

VR ! violationofconnections; (* from FIDetector *)

A StartRinging signal is received on the observation gate "signal" indicating that this signal is received by C where A is involved and is caused by CFBL.

StartRinging_fwd ! C !A; (from CFBL *)*
Signal ! CFBL !ringing !C !A ;

A StartAudibleRinging signal is received on the observation gate "signal" indicating that this signal is received by A where C is involved and is caused by CFBL.

StartAudibleRinging_fwd !A ! C; (from CFBL *)*
Signal ! CFBL !audibleringing !A !C ;

The new signal information causes a conflict with the existing signal information. A is receiving two different signals: LineBusyTone from OCA and StartAudibleRinging from CFBL.

VR ! violationofsignals; (* from FIDetector *)
...etc.

Two interactions are notified:

1. Violation of connection: Connection (CFBL, A, C, true) and Connection (OCS, A, C, false)
2. Inconsistency of signals: Signal (CFBL, audibleringing, A, C) and Signal (OCS, LineBusyTone, A, none)

6.5.2 Interactions between INFB and CFBL

This is an example of scenario leading to interactions between INFB and CFBL

User C subscribes to INFB and user B is a CFBL subscriber. A calls B, B is busy so the call is forwarded to C.

The trace obtained with the goal-oriented tool is:

Offhook ! A; (From CFBL *)*
Dialtone ! A; (From CFBL *)*
Dial ! A ! B; (From CFBL *)*
detect_forward ! C; (from CFBL*)*

Data is received on the observation point "connection" indicating that feature CFBL allows the connection between A and C.

Connection !CFBL! A !C !true;

A StartRinging signal is received on the observation gate "signal" indicating that this signal is received by C where A is involved and is caused by CFBL.

StartRinging_fwd ! C !A;
Signal ! CFBL !ringing !C !A ;

A StartRinging signal is received on the observation gate "signal" indicating that this signal is received by C where A is involved and is caused by INFB.

StartRinging_fwd ! C !A;
Signal ! INFB !ringing !C !A ;
...

A StartAudibleRinging signal is received on the observation gate "signal" indicating that this signal is received by A where C is involved and is caused by CFBL.

StartAudibleRinging_fwd !A ! C;
Signal ! CFBL !audibleringing !A !C ;

A StartAudibleRinging signal is received on the observation gate "signal" indicating that this signal is received by A where C is involved and is caused by INFB.

StartAudibleRinging_fwd !A ! C;
Signal ! INFB !audibleringing !A !C ;

Data is received on the observation point "connection" indicating that feature CFBL allows the connection between A and C.

Connection !CFBL !A !C !true;
...

Data is received on the observation point "connection" indicating that feature INFB allows the connection between A and C.

Connection !INFB !A !C !true;

Offhook_fwd !C;
StopRinging_fwd !C !A;
StopAudibleRinging_fwd !A !C;

Data is received on the observation point "billing" indicating that for feature CFBL A should pay for AB leg.

LogBegin_fwd !A!B!A;
Billing ! CFBL! A!B!A;

Data is received on the observation point "billing" indicating that for feature CFBL B should pay for BC leg.

LogBegin_fwd !B !C !B;
Billing ! CFBL! B ! C !B;

Data is received on the observation point "billing" indicating that for feature INFB B should pay for BC leg.

LogBegin !B!C!C;
Billing !INFB ! B !C! C;

VR ! violationofBilling; (* from FIDetector *)
...etc.

One interaction is notified: Violation of Billing: Billing (CFBL, B, C, B) and Billing (INFB, B, C, C) are conflicting. B and C should not pay for the same leg BC.

6.5.3 Interactions between INFB and OCS

This is an example of scenario leading interactions between INFB and OCS.

User A subscribes to OCS with B within his screening list. User B is an INFB subscriber.

The trace obtained with the goal-oriented tool is:

Connection ! OCS ! A !B ! false;
Offhook ! A; (synchronization between INFB and OCS *)*
Dialtone ! A; (synchronization between INFB and OCS *)*
Dial ! A ! B; (synchronization between INFB and OCS *)*

PlayAnnoucement !A !ScreenedMessageOCS ;
Signal ! OCS ! PlayAnnoucement ! A !none;

StartRinging! B !A; (from INFB *)*
Signal ! INFB !ringing !B !A ;

StartAudibleRinging !A ! B; (from INFB *)*
(Observation point: Signal *)*
Signal ! INFB !audibleringing !A !B ;

VR ! violationofsignals; (* from FIDetector *)
...etc.

Inconsistencies of signals: Signal (INFB, audibleringing, A, C) and Signal (OCS, LineBusyTone, A, none)

6.6 Results

We have investigated the interactions that could occur between the following eight features: Originating Call Screening (OCS), Terminating Call Screening (TCS), IN Free Routing (INFR), Call Forwarding Busy Line (CFBL), Call Number Delivery (CND), IN Freephone Billing (INFB), IN Call Forwarding (INCF) and IN Charge Call (INCC).

Table 5 shows the FI detection results.

	OCS	TCS	INFB	INCC	CND	CFBL	INCF	INFR
OCS	-	-	1	1	-	1, 2	1, 2	1, 2
TCS		-	1	1	4	1, 2	1, 2	1, 2
INFB			-	3	-	1, 3	1, 3	1,3
INCC				-	4	1	1	1
CND					-	4	4	4
CFBL						-	1, 2	1, 2
INCF							1, 2	1, 2
INFR								1, 2

(1) : Signal conflict

(2) : Connection conflict

(3) : Billing conflict

(4) : Display problem

Table 5: Feature Interaction Detection Results

Comparing our results with the benchmark Feature Interaction

The number and type of Feature Interactions detected are two basic factors when evaluating a Feature Interaction detection method. For this reason, the organizing committee of the Feature Interaction Contest [5] published a benchmark document [65], listing the FIs that they believed to exist among the feature to be studied in the contest. In this section, we evaluate our method by comparing the set of interactions detected by FIDS with the one provided in the benchmark.

Before presenting a detailed comparison, we should note that the contest specifications [5] were not specific concerning the composition of the features.

As mentioned in Section 6.2.5, we decided to compose features in parallel and synchronize them only on their common gates, except those representing signals. Thus, they can synchronize on signals in any order and the call process will not be affected if conflicting signals occur.

As mentioned in Section 6.1, we consider only four types of feature interaction: Connection violation, Inconsistency of signals given to users, Incorrectness of Billing, Inconsistency in the display function. The benchmark instead tries a more general classification: Feature Interactions are categorized into corresponding conflict/failure types such as Billing conflict, Call termination conflict, Forwarding conflict, Disconnect conflict, Number delivery failure (Number not displayed), PIN conflicts (over-ride PIN), Flash conflict.

Table 6 lists the mapping relationship from the benchmark FI types to our FI types.

Benchmark FI type	**Our FI type**
Billing conflict	Incorrectness of Billing
Call Termination conflict	Inconsistency of signals given to users
Flash conflict	Not addressed
Disconnect conflict	Inconsistency of signals given to users
Forwarding conflict	Inconsistency of signals given to users Connection violation
PIN conflicts (over-ride PIN)	Not addressed
Number delivery failure	Inconsistency in the display function

Table 6: The Mapping Table of FI Types

From the above mapping, we find that Flash conflict is not addressed in our method since features such as TWC (Three Way Calling) and CW (Call Waiting) are out of the scope of our method. However, most of benchmark FI types can be mapped to a corresponding FI type.

Interactions found under "forwarding conflict" type from benchmark FI types could be either an inconsistency of signals given to users or connection violation.

We failed to detect interactions of type "PIN conflicts" (found between INTL and INCC) because over-riding variables is not considered in our detection method.

The benchmark listed 38 interactions while we succeeded to detect only 36 interactions. Not detected interactions are between:

1.CND-INCF: Scenario causing the interaction: User B subscribed to INCF(C), C subscribed to CND. A calls B and the call is forwarded to C.

Interaction: No number display at C.

2.CND-INFR: Scenario causing the interaction: User B subscribed to INFR(C), C subscribed to CND. A calls B and the call is forwarded to C.

Interaction: No number display at C.

We failed to detect these two interactions related to the display function because in our method we describe only the termination part of the call at user C and we assume that a forwarded call is similar to a normal call.

Table 7 gives a comparison in terms of number of interactions detected for each pair of features. For each feature pair we associate a pair (x, y) where: "x" is the number of feature interactions found with our method and "y" is the number of feature interactions found in the benchmark.

	TCS	INFB	INCC	CND	CFBL	INCF	INFR
TCS	-	(1,1)	(1,1)	(1,1)	(2,2)	(2,2)	(2,2)
INFB		-	(1,1)	-	(2,2)	(2,2)	(2,2)
INCC			-	(1,1)	(1,1)	(1,1)	(2,2)
CND				-	(1,1)	(1,2)	(1,2)
CFBL					-	(3,3)	(3,3)
INCF						(1,1)	(3,3)
INFR							(2,2)

Table 7: Comparison in terms of number of features interactions

According to FI traces described in the benchmark paper, the call process will be terminated when it encounters the first FI. Thus, only one FI can be detected per scenario. However, since we are using the parallel composition and we are introducing observation points to collect data, our method can tolerate any conflicts and the call process continues until all activated features finish. Thus, there is no wonder that our method can detect more than one feature interaction per scenario.

We detected interactions of type "Inconsistency of signals given to users" in some scenarios declared by the filtering method (in Chapter 4) as interaction free scenarios. This is due to:

- The way features are described using our UCM model presented in Section 4.2.2.3. The precedence assumption between stubs in our UCM model solves such interactions. Our FI detection process doesn't consider such assumptions.
- In our FI detection process features don't synchronize on signals.

6.7 Conclusion

The Feature interaction method presented is considered as a complementary step of the filtering process introduced in Chapter 4. Only scenarios with verdict "FI could occur" and "FI occurs" could be analyzed. The method generates also traces, using the goal oriented execution tool, leading to an interaction if one exists. These traces are used further to test the implementation.

The method we have presented uses extensively Abstract Data Types to detect feature interactions.

Chapter 7

Conclusion & Future Work

This thesis proposes a framework for feature Interaction detection.

7.1 Summary

This thesis consists of seven chapters. The background and motivation for our work is given in Chapter 1. In this chapter a list of contributions was also given.

Chapter 2 presents a survey of related work on the formal techniques that are used to specify telephony systems with their features. We also presented some of the Feature Interaction Detection methodologies using FDTs proposed in the literature.

Chapter 3 gives an overview of the existing requirement description techniques relevant to this thesis. We focus on the Use Case Maps notation.

Chapter 4 introduces a Use Case Maps model for describing telephony features at the requirement stage. This model, called also root map, allows us to integrate features, both switch based and IN ones, into the basic call model. Features like CFBL, INTL, INFB, OCS,

TCS are used as examples to illustrate the feature integration mechanism. Based on this model, we propose a method to filter feature interactions at the requirement stage. This method allows the designer to localize where interactions could occur during a call and facilitates their further detection by taking out the interaction-free scenarios. Finally results of the filtering on examples of switch and IN based features were presented. The experimental evaluation shows that more than half of the feature combinations can be filtered (see Section 4.6.1).

As mentioned in Section 4.8, the method is limited to features that can be expressed with a certain structure, however it could be generalized.

Chapter 5 gives an overview of the LOTOS language and of its main operators. It also shows the use of LOTOS as a formal description Technique (FDT) in specifying the telephone system model and features. Our main objective in specifying the system model and features in LOTOS is to provide a specification that can be used for validating and detecting feature interactions.

In Chapter 6, a formal definition of Feature Interaction is provided and a FI Detection system (FIDS) is developed based upon the definition. Based on our experience with feature interaction, we have investigated four kinds of interactions:

- Connection violations
- Inconsistency of signals given to users
- Incorrectness of Billing
- Problems in the display function

Our technique is based on Abstract Data Types (ADT) to detect these violations.

The Feature interaction method presented is considered as a complementary step of the filtering process introduced in Chapter 4.

The methodology presented in this thesis does not give a general solution to the feature interaction problem but a partial solution limited to the detection of the four types of interactions

mentioned above. It allows also the automatic generation of functional test cases that can be used to see whether an interaction exists in the implementation.

We have shown that telecommunication system designers can give precise description and can validate their design with respect to potential feature interaction problems before the implementation stage.

7.2 Future Work

The results of this thesis provide a basis for several future research directions.

As mentioned in Section 4.8, features such as Call Waiting and Three Way Calling are outside of the scope of the filtering method we propose. It would be useful to extend our UCM model to be able to describe these features and to investigate them since they can be in conflict with other features. The filtering quality of our method could also be enhanced using more information (e.g. connectivity of scenario paths and preconditions).

We should also acknowledge a conceptual problem that we could not address in this thesis. The rules we introduced in Chapter 4 to determine that a system is interaction free are not shown to be consistent with the formal definition of feature interactions we used in Chapter 6. In other words, it is possible that a combination of features declared to be interaction free according to the rules of Chapter 4 is in fact not interaction free according to the definition of Chapter 6. However, in all the examples we have analysed we have not found any example showing this possibility. We leave the study of this question to further work.

The current development of telecommunication technology is so intensive and goes in so many directions that probably nobody can really predict what the telecommunication market will look like five years from now. Looking at current developments, one of the things we can be certain is that new types of feature interaction are emerging. Complications arise since functionality tends to be more distributed. Many techniques for resolving interactions in the PSTN are no longer easily applied. Trying to solve these interactions using Use Case Maps and applying formal methods like LOTOS and SDL in this new area is certainly a worthwhile and challenging issue.

Appendix

Specification of the abstract data types

- **Type PIN: Personal Identification Number**
  The type PIN describes the different personal identification numbers

```
type PIN is Boolean
sorts PIN
opns
        PIN_INTL            (*! constructor *),
        Invalid_PIN_INTL    (*! constructor *) : → PIN
        _eq_,
        _ne_                                  : PIN, PIN → Bool

eqns
        forall P1, P2:PIN
        ofsort Bool

                P1 eq P1 = true;
                P2 eq P2 = true;
                P1 eq P2 = false;
                P2 eq P1 = false;
                P1 ne P2 = not (P1 eq P2);

endtype (*Type PIN*)
```

- Type Message

```
type Message is NaturalNumber
sorts Message
opns

        AskForPIN           (*! constructor *),
        ScreenedMessageOCS  (*! constructor *),
        EnterPhoneNumber    (*! constructor *) : → Message
        _eq_,
        _ne_                                  : Message, Message → Bool
        h                                     : Message → Nat

eqns
        forall M1, M2:Message
        ofsort Bool
                M1 eq M2 = h(M1) eq h(M2);
                M1 ne M2 = h(M1) ne h(M2);

        ofsort Nat
                h(AskForPIN) = 0;
                h(ScreenedMessageOCS) = Succ(0);
```

```
        h(EnterPhoneNumber) = Succ (Succ(0));

endtype (*Type Message *)
```

- Type State

```
type State is NaturalNumber
sorts State
opns
        Idle,           (*! constructor *)
        Busy,           (*! constructor *)
        Ringing,        (*! constructor *)
        Audibleringing, (*! constructor *),
        Linebusy,       (*! constructor *),
        Annoucement,    (*! constructor *),
        Talking         (*! constructor *)      : → State

        map                                     : State → Nat
        _ eq _,
        _ ne _                                  : State, State → Bool

eqns
        forall s1, s2: State
        ofsort Nat

                map (Idle)              = 0;
                map (Busy)              = Succ(0);
                map (ringing)           = Succ(Succ(0));
                map (audibleringing)    = Succ(Succ(Succ(0)));
                map (linebusy)          = Succ(Succ(Succ(Succ(0))));
                map (Annoucement)       = Succ(Succ(Succ(Succ(Succ(0)))));
                map (talking)           = Succ(Succ(Succ(Succ(Succ(Succ(0))))));

        ofsort Bool
                s1 eq s2 = map(s1) eq map (s2) ;
                s1 ne s2 = not (s1 eq s2);

endtype (* State *)
```

- Type SignalRecord

```
type SignalRecord is Feature, Number, State, Boolean
sorts SignalRecord
opns

        st                      : Feature, State, Number, Number → SignalRecord

        extract_feature         : SignalRecord → Feature
        extract_Number1         : SignalRecord → Number
        extract_Number2         : SignalRecord → Number
        extract_State           : SignalRecord → State
         _eq_, _ne_             : SignalRecord, SignalRecord → Bool
```

```
eqns

        forall   N1,N2,N3,N4     : number,
                 F1,F2           : Feature,
                 S1,S2           : State,
                 B1,B2           : Bool

        ofsort Feature
                 extract_feature(st(F1,S1,N1,N2)) = F1;

        ofsort Number
                 extract_Number1(st(F1,S1,N1,N2)) = N1;
                 extract_Number2(st(F1,S1,N1,N2)) = N2;

        ofsort State
                 extract_State(st(F1,S1,N1,N2)) = S1;

        ofsort Bool
                 st(F1,S1,N1,N2) eq st(F2,S2,N3,N4) =
                 (F1 eq F2) and (S1 eq S2) and (N1 eq N3)and(N2 eq N4);

                 st(F1,S1,N1,N2) ne st(F2,S2,N3,N4) =
                 not(st(F1,S1,N1,N2) eq st(F2,S2,N3,N4));

endtype (* SignalRecord *)
```

- Type SignalRecord_set

```
type SignalRecord_set is SignalRecord

sorts SignalRecord_sets
opns

(* An empty set*)
        { }                              : → SignalRecord_sets

(* Insert a new signalRecord in the signalRecord set *)
        insert                           : SignalRecord,SignalRecord_sets → SignalRecord_sets

        head                             : SignalRecord_sets → SignalRecord
        tail                             : SignalRecord_sets → SignalRecord_sets

(* Comparing signal sets *)
        _eq_, _ne_                       : SignalRecord_sets, SignalRecord_sets → Bool

(* Checks whether the signal set is empty *)
        empty                            : SignalRecord_sets → Bool

(* Cheks if a signalRecord is in the signalRecord set *)
        isin                             : SignalRecord, SignalRecord_sets → Bool

(* Test if a signalRecord can cause violation within the signalRecord set *)

        CausesViolation_Signal           : SignalRecord, SignalRecord_sets → Bool
```

```
eqns

forall  t1,t2           : SignalRecord,
        F1,F2           : Feature,
        S1,S2           : State,
        S               : SignalRecord_sets,
        n1,n2,n3,n4     : Number
        b1,b2           :Bool

        ofsort SignalRecord_sets
                tail(insert(t1,s))=s;


        ofsort SignalRecord
                head(insert(t1,s))=t1;

        ofsort Bool
                { } eq { } = true;
                { } eq insert(t1,s) = false;
                insert(t1,s) eq { } = false;
                empty(s) = s eq { };
                isin(t1,{ }) = false;
                t1 eq t2 => isin(t1,insert(t2,s)) = true;
                t1 ne t2 => isin(t1,insert(t2,s)) = isin(t1,s);
                CausesViolation_Signal(st(F1,S1,n1,n2),{ }) = false;

                CausesViolation_Signal (st(F1,S1,n1,n2),insert(st(F2,S2,n3,n4),s)) =
                        ((F1 ne F2) and (S1 eq S2) and (n1 eq n3) and (n2 ne n4))
                                or
                        ((F1 ne F2) and (S1 ne S2) and (n1 eq n3))
                                or
                        CausesViolation_Signal(st(F1,S1,n1,n2),s);
```

endtype (* SignalRecord_set*)

- Type BillingRecord

```
type BillingRecord is Feature, Number
sorts BillingRecord
opns

        bt                      : Feature, Number, Number, Number → BillingRecord
        extract_feature         : BillingRecord → Feature
        adr1                    : BillingRecord → Number
        adr2                    : BillingRecord → Number
        ChargedParty            : BillingRecord → Number
        _eq_,
        _ne_                    : BillingRecord, BillingRecord → Bool

eqns
        forall N1,N2,N3,N4,N5,N6        : Number,
        t1,t2                           : BillingRecord,
        B1,B2                           : Bool,
        F1,F2                           : Feature
```

```
        ofsort Feature
                extract_feature(bt(F1,N1,N2,N3)) = F1;

        ofsort Number
                adr1(bt(F1,N1,N2,N3)) = N1;
                adr2(bt(F1,N1,N2,N3)) = N2;
                ChargedParty(bt(F1,N1,N2,N3)) = N3;

        ofsort Bool
                bt(F1,N1,N2,N3) eq bt(F2,N4,N5,N6) =
                ((F1 eq F2)and(N1 eq N4) and (N2 eq N5) and (N3 eq N6)) or
                ((F1 eq F2)and(N1 eq N5) and (N2 eq N4) and (N3 eq N6)) ;

                bt(F1,N1,N2,N3) ne bt(F2,N4,N5,N6) =
                not (bt(F1,N1,N2,N3) eq bt(F2,N4,N5,N6));
```

endtype (* BillingRecord *)

- Type BillingRecord_set

```
type BillingRecord_set is BillingRecord
sorts BillingRecord_sets
opns
        { }     (*! constructor *): → BillingRecord_sets
        insert (*! constructor *) : BillingRecord,BillingRecord_sets → BillingRecord_sets
        head  : BillingRecord_sets → BillingRecord
        tail  : BillingRecord_sets → BillingRecord_sets
        _eq_,
        _ne_  : BillingRecord_sets, BillingRecord_sets → Bool
        empty : BillingRecord_sets → Bool
        eleof : BillingRecord, BillingRecord_sets → Bool
        IsIn  : BillingRecord, BillingRecord_sets → Bool
        CausesViolation_Billing : BillingRecord, BillingRecord_sets → Bool

eqns

 forall t1,t2 : BillingRecord,
        s: BillingRecord_sets,
        bts1,bts2: BillingRecord_sets,
        n1,n2,n3,n4,n5,n6:Number,
        b1,b2:Bool,
        F1,F2: Feature


        ofsort BillingRecord_sets
                tail(insert(t1,s))=s;

        ofsort BillingRecord
                head(insert(t1,s))=t1;

        ofsort Bool
                { } eq { } = true;
                { } eq insert(t1,s) = false;
                bts1 ne bts2 = not(bts1 eq bts2);
```

```
        insert(t1,s) eq { } = false;
        empty(s) = s eq { };
        eleof(t1,{ }) = false;
        t1 eq t2 => eleof(t1,insert(t2,s)) = true;
        t1 ne t2 => eleof(t1,insert(t2,s)) = eleof(t1,s);
        IsIn(t1,{ }) = false;
        t1 eq t2 => IsIn(t1,insert(t2,s)) = true;
        t1 ne t2 => IsIn(t1,insert(t2,s)) = IsIn(t1,s);
        CausesViolation_Billing(bt(F1,n1,n2,n3),{ }) = false;
        CausesViolation_Billing(bt(F1,n1,n2,n3),insert(bt(F2,n4,n5,n6),s)) =
        ((F1 ne F2) and (n1 eq n4) and (n2 eq n5) and (n3 ne n6)) or
        ((F1 ne F2) and (n1 eq n5) and (n2 eq n4) and (n3 ne n6)) or
        CausesViolation_Billing( bt(F1,n1,n2,n3),s);

endtype
```

- Type DisplayRecord

```
type DisplayRecord is Number,Boolean
sorts DisplayRecord
opns (* specify the format of operations *)

dt: Feature, Number, Number, Bool → DisplayRecord

(* num1 and num2 are 2 operations used to extract the user numbers involved in the connection *)

num1        : DisplayRecord → Number
num2        : DisplayRecord → Number

Bvalue      : DisplayRecord → Bool

(* eq, and ne are used to compare displayRecords *)

_eq_, _ne_  : DisplayRecord, DisplayRecord → Bool

eqns (* List of equations *)

    forall  N1,N2,N3,N4   :number,
            t1,t2         : DisplayRecord,
            B1,B2         : Bool
            F1, F2        : Feature
    ofsort Number
            num1(dt(F,N1,N2,B1)) = N1;
            num2(dt(F,N1,N2,B1)) = N2;

    ofsort Bool
            Bvalue(dt(F,N1,N2,B1)) = B1;
            dt(F,N1,N2,B1) eq dt(F,N3,N4,B2) = ((F1 eq F2) and (N1 eq N3) and (N2 eq N4) and (B1 eq B2))
                                                    or
                                        ((F1 eq F2) and  (N1 eq N4) and (N1 eq N2) and (B1 eq B2));

            dt(N1,N2,B1) ne dt(N3,N4,B2) = not( dt(N1,N2,B1) eq dt(N3,N4,B2));

endtype (* DisplayRecord *)
```

- Type DisplayRecord_set

```
type DisplayRecord_set is DisplayRecord
sorts DisplayRecord_sets
opns

(* An empty set*)
        { }                         : → DisplayRecord_sets

(* Insert a new DisplayRecord in the DisplayRecord_set *)
        insert                      : DisplayRecord, DisplayRecord_sets → DisplayRecord_sets
        tail                        : DisplayRecord_sets -> DisplayRecord_sets

(* Comparing Display sets *)
        _eq_, _ne_                  : DisplayRecord_sets, DisplayRecord_sets -> Bool

        empty                       : DisplayRecord_sets → Bool

(* Search for a connection in the display set *)
        isin                        : DisplayRecord, DisplayRecord_sets → Bool

(* Test if a connection can cause violation within the display set *)
        CausesViolation_Display  : DisplayRecord, DisplayRecord_sets → Bool

Eqns (* List of Equations *)

        forall   t1,t2              : DisplayRecord,
                 s                  : DisplayRecord_sets,
                 n1,n2,n3,n4        : Number,
                 b1,b2              : Bool
                 F1, F2             : Feature

ofsort DisplayRecord_sets
        tail (insert(t1,s)) = s;

ofsort Bool
        { } eq { } = true;
        { } eq insert(t1,s) = false;
        insert(t1,s) eq { } = false;
        empty(s) = s eq { };
        isin(t1,{ }) = false;
        t1 eq t2 => isin(t1,insert(t2,s)) = true;
        t1 ne t2 => isin(t1,insert(t2,s)) = isin(t1,s);
        CausesViolation_Display (dt(F1, n1, n2, b1),{ }) = false;

        CausesViolation_Display (dt(F1, n1,n2,b1),insert(dt(F2, n3,n4,b2),s)) =
        ((F1 ne F2) and (n1 eq n3) and (n2 eq n4) and (b1 ne b2))
                or
        ((F ne F2) and (n1 eq n4) and (n2 eq n3) and (b1 ne b2))
                or
        CausesViolation_Display (dt(F1, n1,n2,b1),s);

endtype
```

Some Feature Specifications

• OCS (Originating Call Screening)

```
process OCS_feature[Offhook, DialTone, Onhook, Dial, StartRinging, StartAudibleRinging, StopRinging,
StopAudibleRinging, LineBusyTone, LogBegin, LogEnd, Disconnect, PlayAnnoucement,
Connection, Billing, Signal, Display] (B_State: State, ocs_list: Number) :noexit:=

(* Observation point: Connection *)
Connection ! OCS !A !ocs_list ! false;

Offhook !A;
DialTone !A;
(
        Onhook !A; stop
        [ ]

        ( Dial !A!B;
                (
                [B eq ocs_list] → PlayAnnoucement !A !ScreenedMessageOCS ;
                                (* Observation point: Signal*)
                                Signal ! OCS ! PlayAnnoucement ! A !none;
                                Onhook !A;
                                stop

                        [ ]

                [B ne ocs_list] →
                        (
                        [B_State eq idle] →
                                        StartRinging !B!A ;
                                        (* Observation point: Signal*)
                                        Signal ! OCS ! ringing ! B !A;

                                        StartAudibleRinging !A !B;
                                        (* Observation point: Signal*)
                                        Signal ! OCS ! audibleringing ! A !B;

                                        ( Onhook !A;
                                        StopRinging !B!A;
                                        StopAudibleRinging !A!B;stop

                                                [ ]

                                        Offhook !B;
                                        (* Observation point: Connection*)
                                        Connection ! OCS !A !B !true ;

                                        StopRinging !B !A;
                                        StopAudibleRinging !A !B;
                                        LogBegin !A!B!A;

                                        (* Observation point : Billing*)
                                        Billing ! OCS !A !B !A ;
```

```
                                        (
                                        Onhook !A;
                                        Disconnect !B!A;
                                        LogEnd !A!B;
                                        Onhook !B;
                                        stop

                                        [ ]

                                        Onhook!B;
                                        Disconnect !A!B;
                                        LogEnd !A!B;
                                        Onhook !A;
                                        Stop
                                        )
                                )
                        [ ]
                [B_State eq busy] →     LineBusyTone!A ;
                                        (* Observation point: Signal*)
                                        Signal ! OCS ! LineBusyTone ! B !A;
                                        Onhook!A ;
                                        stop
        )
      )
 )
)
endproc (* OCS Feature *)
```

• CFBL (Call Forwarding Busy Line)

```
process CFBL_feature [Offhook, DialTone, Onhook, Dial,StartRinging, StartAudibleRinging, StopRinging,
StopAudibleRinging, LineBusyTone, LogBegin, LogEnd, Disconnect, Detect_forward, LineBusyTone_fwd,
Onhook_fwd, Offhook_fwd, StartRinging_fwd, StartAudibleRinging_fwd, StopRinging_fwd,
StopAudibleRinging_fwd, LogBegin_fwd, LogEnd_fwd, Disconnect_fwd, Connection, Billing, Signal, Display]
(B_State:InitialState,fwd_number_State:InitialState, fwd_number:Number)
 :noexit:=

Offhook !A;
DialTone !A;
(
 Onhook !A; stop
   []
 (
  Dial !A!B;
   (
     [B_State eq idle] ->   StartRinging !B!A ;
                            (* Observation point: Signal *)
                            Signal ! CFBL !ringing !B !A ;

                            StartAudibleRinging !A !B;
                            (* Observation point: Signal *)
                            Signal ! CFBL !audibleringing !A !B ;
```

```
                    ( Onhook !A;
                    StopRinging !B!A;
                    StopAudibleRinging !A!B;stop

                    [ ]

                    Offhook !B;
                    (*Observation point*)
                    Connection !CFBL !A !B !true ;

                    StopRinging !B !A;
                    StopAudibleRinging !A !B;
                    LogBegin !A!B!A;

                    (* Observation point: Billing *)
                    Billing ! CFBL! A !B!A;
                    (
                    Onhook !A;
                    Disconnect !B!A;
                    LogEnd !A!B;
                    Onhook !B;
                    stop
                      []
                    Onhook!B;
                    Disconnect !A!B;
                    LogEnd !A!B;
                    Onhook !A;
                    stop
            )
           )

      [ ]

[B_State eq busy]-> Detect_forward ! fwd_number;
       (
      [fwd_number_State eq Busy] -> LineBusyTone_fwd !A;

                                (* Observation point: Signal *)
                                Signal ! CFBL !linebusy !A !none;
                                onhook_fwd !A;
                                stop
          []
      [fwd_number_State eq Idle] ->

            (* Observation point: Connection*)
            Connection !CFBL! A !fwd_number !true;

            StartRinging_fwd ! fwd_number !A;
            (* Observation point *)
            Signal ! CFBL !ringing !fwd_number !A ;

            StartAudibleRinging_fwd !A !fwd_number;
            (* Observation point *)
            Signal ! CFBL !audibleringing !A !fwd_number ;
```

```
            ( onhook_fwd !A;
              StopRinging_fwd ! fwd_number !A;
              StopAudibleRinging_fwd !A !fwd_number;
              stop
                [ ]
            Offhook_fwd ! fwd_number;
            StopRinging_fwd ! fwd_number !A;
            StopAudibleRinging_fwd !A !fwd_number;

            LogBegin_fwd !A!B!A;
            (* Observation point *)
            Billing ! CFBL! A!B!A;

            LogBegin_fwd !B!fwd_number!B;
             (* Observation point*)
             Billing ! CFBL! B ! fwd_number !B;

              (
                          onhook_fwd !A;
                          Disconnect_fwd !fwd_number!A;
                          LogEnd_fwd !A!B;
                          LogEnd_fwd !B!fwd_number;
                          onhook_fwd !fwd_number;stop
                        [ ]
                          onhook_fwd !fwd_number;
                          Disconnect_fwd !A !fwd_number;
                          LogEnd_fwd !A!B;
                          LogEnd_fwd !B!fwd_number;
                          onhook_fwd !A;stop
                        )
              )
          )

  )
)
)
```

endproc

• IN Free Billing (INFB)

process INFB_feature[Offhook, DialTone, Onhook, Dial,StartRinging, StartAudibleRinging, StopRinging, StopAudibleRinging, LineBusyTone, LogBegin, LogEnd, Disconnect, Connection, Billing, Signal, Display] (B_State:InitialState) :**noexit**:=

```
Offhook ! A ;
Dialtone ! A ;
( onhook !A; stop
[ ]
Dial !A!B;
  (
    [B_State eq busy]->    LineBusyTone!A ;
                           (* Observation point: Signal *)
                           Signal ! INFB !linebusy !A !none;
                           Onhook!A ;
```

```
                        stop
                          [ ]
[B_State eq idle] ->    StartRinging !B!A ;
                        (* Observation point: Signal *)
                        Signal ! INFB !ringing !B !A ;

                        StartAudibleRinging !A !B;
                        (* Observation point: Signal *)
                        Signal ! INFB !audibleringing !A !B ;

                          ( Onhook !A;
                        StopRinging !B!A;
                        StopAudibleRinging !A!B;
                        stop

                         [ ]

                        Offhook !B;
                        (*Observation point: Connection*)
                        Connection !INFB !A !B !true;

                        StopRinging !B !A;
                        StopAudibleRinging !A !B;

                        LogBegin !A!B!B;
                        (* Observation point*)
                        Billing !INFB ! A !B! B;
                    (
                     Onhook !A;
                     Disconnect !B!A;
                     LogEnd !A!B;
                     Onhook !B;
                      stop
                        [ ]
                     Onhook!B;
                     Disconnect !A!B;
                     LogEnd !A!B;  (*Add The Time*)
                     Onhook !A;
                     stop

                    )
                )
  )
)
endproc
```

Bibliography

[1] 1st International Workshop on Feature Interactions in Telecommunication Software Systems, Florida, December, 1992.

[2] 2nd International Workshop on Feature Interactions in Telecommunications Software Systems, Netherlands, May, 1995.

[3] 3rd International Workshop on Feature Interactions in Telecom. Software Systems, Japan, 1995.

[4] 4th International Workshop on Feature Interactions in Telecom. Software Systems, Montreal, June, 1997.

[5] 5th International Workshop on Feature Interactions in Telecom. Software Systems, Sweden, September, 1998.

[6] 6th International Workshop on Feature Interactions in Telecom. Software Systems, Scotland, May, 2000.

[7] A. Aho, S. Gallagher, N. Griffeth, C. Scheel and D. Swayne. Sculptor with Chisel: Requirements Engineering for Communications services. In: K. Kimbler and W. Bouma (eds.), Fifth international workshop on Feature Interactions in Telecommunications Software Systems, IOS Press 1998, 45-63

[8] A. Dardenne, A. V. Lamsweerde, and S. Fickas. Goal-directed requirements acquisition. Science of Computer Programming XX:3-50, 1993.

[9] A. Grinberg. Seamless Networks: Interoperating Wireless and Wireline Networks. Addison-Wesley, 1996

[10] A. Gammelgaard and J. E. Kristensen. Interaction Detection, a logical approach. In: L.G. Bouma and H. Velthuijsen (eds.) Feature Interactions in Telecommunications Systems. IOS Press, 1994 (Proc. of the 2nd International Workshop on Feature Interactions in Telecommunications Systems, Amsterdam) 178-196.

[11] A. J. Tocher. LOTOS and the Formal Specification of Communication Standards: An example. Formal Methods: Theory and Practice, Chapter 2, edited by P.N. Scarbach. BP Research, 1989.

[12] A. Khoumsi. Detection and Resolution of Interactions between Services of telephone Networks. In: Feature Interactions in Telecommunications and Distributed systems IV, P. Dini et al. (Eds.) IOS Press 1997, 78-92

[13] A. Miga. Application of Use Case Maps to System Design with Tool Support. M.Eng. Thesis, Dept. of Systems and Computer Engineering, Carleton University, Ottawa, Canada, 1998. http://www.UseCaseMaps.org/UseCaseMaps/ucmnav

[14] A. P. Felty and K. S. Namjoshi. Feature Specification and Automatic Conflict Detection. In: M. Calder and E. Magill, Sixth International Workshop on Feature Interactions in Telecommunications and Software Systems, IOS Press, Glasgow 2000, 274-289

[15] B. Ehrig, B. Mahr. Fundamentals of Algebraic Specifications. Springer-Verlag, 1985

[16] B. Ghribi. A Model Checker for LOTOS. Master Thesis, Department of Computer Science, University of Ottawa, 1993.

[17] B. Jonsson, T. Margaria, G. Naeser, J. Nystrom and B. Steffen. Incremental Requirement Specification for Evolving Systems. In: M. Calder and E. Magill, Sixth International Workshop on Feature Interactions in Telecommunications and Software Systems, IOS Press, Glasgow 2000, 145-162

[18] B. Stepien, L. Logrippo. Status-Oriented Telephone Service Specification. IN: T.Rus and C.Rattray (eds.) Theories and Experiences for Real-Time System Developpement. AMAST Series in computing, Vol2, world Scientific, 1994, pages 265-286

[19] B. Stepien and L. Logrippo. Feature Interaction Detection using backward Reasoning with LOTOS. In: S. Vuong (ed.) Protocol Specification, Testing and Verification, XIV Proc. Of the 14th International Symposium on Protocol Specification, Testing and Verification, organised by IFIP WG 6.1, Vancouver, 1995, pages 71-86

[20] B. Stepien and L. Logrippo. Representing and Verifying intentions in Telephony Features using Abstract Data Types. Third International Workshop on Feature Interactions in Telecommunications Software Systems, eds. K. E. Cheng and T. Ohta, IOS Press 1995, pages 141-155

[21] B. Stepien and L. Logrippo. Status-Oriented Telephone Service Specification. Theories and Experiences for Real-Time System Development. AMAST Series in Computing, Vol.2 World Scientific, 1994

[22] D. Amyot. Formalization of Timethreads Using LOTOS. MSc Thesis, 1994, University of Ottawa, Ottawa, Canada.

[23] D. Amyot, L. Charfi, N. Gorse, T. Gray, L. Logrippo, J. Sincennes, B. Stepien T. Ware. Feature Description and Feature Interaction Analysis with Use Case Maps and LOTOS. In: M. Calder and E. Magill, Sixth International Workshop on Feature Interactions in Telecommunications and Software Systems, IOS Press, Glasgow 2000, 274-289

[24] D. Amyot, R. Andrade. Description of Wireless Intelligent Network Services with Use Case Maps. 17o. Brazilian Symposium on Computer Networks proceedings. May 1999.

[25] D. Amyot, R. Andrade, L. Logrippo, J. Sincennes, and Z. Yi. Formal Methods for Mobility Standards. In the IEEE 1999 Emerging Technologies Symposium: Wireless Communications

and Systems proceedings, Richardson (TX) April 1999. Editor: Traci King, Samsung Telecommunications America. Publisher: Steve Bootman, Hitachi Telecom

[26] F. Bordeleau, A Systematic and Traceable Progression from Scenario Models to Communicating Hierarchical State Machines. PhD Thesis, 1999. Carleton University, Ottawa Canada.

[27] G. Booch, Object-Oriented Design with Applications. The Benjamin Cummings Publ. Co., 1991

[28] G. Booch, I. Jacobson, J. Rumbaugh. Unified Modeling Language for Real-time systems design. Documentation set version 2.0, 1997.

[29] G. Bruns, P. Mataga, I. Sutherland. Features as Service transformers. In: K. Kimbler and W. Bouma (eds.), Fifth international workshop on Feature Interactions in Telecommunications Software Systems, IOS Press, 1998, 85-97

[30] G. V. Bochmann. Finite State Description of Communication Protocols. In: Computer Networks, Vol. 2 (1978), 361-372.

[31] I. Jacobson. Object-Oriented Software Engineering. Addison-Wesley, 1992

[32] I. Jacobson, G.Booch, J.Rumbaugh. The Unified software Development Process. Addison-Wesley, 1999.

[33] ITU, Recommendation Z. 120: Message Sequence Chart (MSC). Geneva, 1996

[34] J. Bergstra, W. Bouma. Models for Feature Descriptions and Interactions. In: Feature Interactions in Telecommunications and Distributed systems IV, P. Dini et al. (Eds.) IOS Press 1997

[35] J. Kamoun, L. Logrippo. Goal-Oriented Feature Interaction Detection in the Intelligent Network Model. Feature Interactions in Telecommunications and Software Systems V, eds. K. Kimbler and L. G. Bouma, IOS Press 1998.

[36] J. Lennox, H. Schulzrinne. Feature Interaction in Internet Telephony. In: M. Calder and E. Magill, Sixth International Workshop on Feature Interactions in Telecommunications and Software Systems, IOS Press, Glasgow 2000, 38-50

[37] J. P. Gibson. Feature Requirements Models: Understanding Interactions. In: Feature Interactions in Telecommunications and Distributed systems IV, P. Dini et al. (Eds.) IOS Press 1997, 46-60

[38] J. Q. Fu. Feature Interaction Detection in a Telephony Network Integrated with Switch based Features and IN Features. Master's Thesis, University of Ottawa, 2000.

[39] J. Rumbaugh, M. Blaha, W. Premerlani, F. Eddy, W. Lorensen. Object Oriented Modeling and Design. Prentice-Hall, 1991.

[40] K. E. Cheng. Towards a Formal Model for Incremental Service Specification and Interaction Management Support. In: Feature Interactions in Telecommunications Systems. eds. L.G. Bouma and H. Velthuijsen, IOS Press 1994.

[41] L. Logrippo, M. Faci, M. Haj-Hussein. An Introduction to LOTOS: Learning by Examples. Computer Networks and ISDN Systems, Vol. 23, No 5, 1992 325-342.

[42] M. Faci and L. Logrippo. Specifying Hardware in LOTOS. Proceedings of the IFIP WG10.2. 11th international Conference on computer Hardware Description Languages and their Applications, Ottawa, Canada, April 1993, 305-312

[43] M. Faci, L. Logrippo and B. Stepien. Formal Specifications of Telephone Systems in LOTOS: The Constraints-Oriented Style Approach. Computer Networks and ISDN Systems, 21, North Holland, 1991, 52-67

[44] M. Faci, L. Logrippo and B. Stepien. Formal Specifications of telephone Systems in LOTOS. Protocol Specification, Testing, and Validation IX, eds. E. Brinksma, G. Scolo, and C. Vissers, 1990

[45] M. Faci, L. Logrippo and B. Stepien. Structural Models for specifying Telephone Systems. Computer Networks and ISDN Systems, 1997

[46] M. Haj-Hussein. Goal Oriented Execution for LOTOS. Phd Thesis, Department of Computer Science, University of Ottawa, 1995.

[47] M. Haj-Hussein, L. Logrippo and J. Sincennes. Goal Oriented Execution of LOTOS Specifications. In: M.Diaz and R.Groz (Eds) Formal Description Techniques, V. North Holland, 1993, 311-327

[48] M. Heisel and J. Souquieres. A Heuristic Approach to detect Feature Interactions in Requirements. In: K. Kimbler and W. Bouma (eds.), Fifth international workshop on Feature Interactions in Telecommunications Software Systems, IOS Press, 1998, 165-171

[49] M. Kolberg and E. H. Magill. Service and Feature Interactions in TINA. In: K. Kimbler and W. Bouma (eds.), Fifth international workshop on Feature Interactions in Telecommunications Software Systems, IOS Press 1998, 78-84

[50] M. Nakamura, J. Hassine, L. Logrippo. Feature Interaction Filtering with Use Case Maps at Requirements Stage. In: M. Calder and E. Magill, Sixth International Workshop on Feature Interactions in Telecommunications and Software Systems, IOS Press, Glasgow 2000, 163-178

[51] M. Yoeli and Z. Barzilai. Behavioural Descriptions of Communication Switching Systems Using Extended Petri Nets. In: Digital Processes, Vol. 3, 4, 307-320. 1977.

[52] N. D. Griffeth and H. Velthuijsen, The negotiating agents approach to run-time feature interaction resolution, In: Feature Interactions in Telecommunications Systems, L.G. Bouma and H. Velthuijsen (eds.), IOS Press, Amsterdam, 1994.

[53] P. Combes and S. Pickin. Formalization of a User View of Network and Services for Feature Interaction Detection. Feature Interactions in Telecommunications Systems. Eds. L. G. Bouma and H. Velthuijsen, IOS Press 1994

[54] P. Zave. Classification of research efforts in requirements engineering. ACM Computing Surveys XXIX (4): 315-321, December 1997

[55] P. Zave and M. Jackson. *Four dark corners of requirements engineering*. ACM Trans. Software Eng. and Methodology, 6(1):1-30, Jan. 1997. 85

[56] R. Boumezbeur and L. Logrippo. Specifying Telephone Systems in LOTOS. IEEE Communications Magazine, August, 1993

[57] Recommendation Z100. Specification and Description Language SDL. CCITT SGX, Contribution com X-R 15-E, 1987

[58] R. J. A. Buhr, R. S. Casselman. Use Case Maps for Object-Oriented Systems. Prentice-Hall 1995, USA.

[59] R. J. A. Buhr. Use Case Maps as Architectural Entities for complex systems. In: Transactions on Software Engineering, IEEE, December 1998, pp. 1131-1155

[60] R. J. A. Buhr, D. Amyot, M. Elammari, D. Quesnel, T. Gray, and S. Mankovski, Feature-Interaction Visualization and Resolution in an Agent Environment. In: K. Kimbler and

W. Bouma (eds.), Fifth international workshop on Feature Interactions in Telecommunications Software Systems, IOS Press, 1998, 135-149

[61] R. J. Hall. Feature Combination and Interaction Detection via Foreground/Background Models. In: K. Kimbler and W. Bouma (eds.), Fifth international workshop on Feature Interactions in Telecommunications Software Systems, IOS Press, 1998, 232-246

[62] R. Milner. A Calculus of Communicating Systems. Lecture Notes in Computer Science, (Springer-Verlag, 1980) No 92

[63] R. Zygan-Maus. Feature Interaction Management for Public Switched Networks. In: K. Kimbler and W. Bouma (eds.), Fifth international workshop on Feature Interactions in Telecommunications Software Systems, IOS Press 1998, 32-44

[64] R. Lai and A. Jirachiefpattana. Communication Protocol Specification And Verification. Kluwer Academic Publishers, 1998

[65] R. Blumenthal, N. Griffeth, J. C. Gregoire and T. Ohta, Feature Interaction Contest Interaction Descriptions, 1999

[66] T. Bolognesi and Brinksma. Introduction To The ISO Specification Language LOTOS. In: The Formal Description Technique LOTOS, P.H.J van Eijk, C.A. Vissers and M. Diaz (Eds.). Elsevier Science Publishers B.V, North-Holland, 1989

[67] T. F. LaPorta, D. Lee, Y.-J. Lin, and M. Yannakakis. Protocol Feature Interactions. In S. Budkowski, A. Cavalli, and E. Najm, editors, Proc. Formal Description Techniques And Protocol Specification, Testing and Verification, FORTE XI/PSTV XVIII'98, 3-6 November, Paris, France, pages 59--74, 1998.

[68] Y. Iraqi and M. Erradi. An Experiment for the Processing of Feature Interactions within an Object-Oriented Environment. In: Feature Interactions in Telecommunication Networks IV, eds P. Dini, R. Boutaba and L. Logrippo, IOS Press 1997.

[69] Use Case Maps Web Page and UCM User Group, 1999. http://www.UseCaseMaps.org

[70] V. Whitis and W. Chiang. A State Machine Development Method for Call Processing Software. In Proceedings of the IEEE Electro 81, IEEE Press, Washington D.C., April 1981

[71] W. Bouma and H. Zuidweg. Formal Analysis of Feature Interactions by Model Checking. PTT research, the netherlands, December, 1992

MQ

66050

UMI

MICROFILMED 2002

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

Université d'Ottawa • University of Ottawa

PERMISSION DE REPRODUIRE ET DE DISTRIBUER LA THÈSE

PERMISSION TO REPRODUCE AND DISTRIBUTE THE THESIS

NOM DE L'AUTEUR / NAME OF AUTHOR:	SHANNON HAYWARD-McCLELLAND
ADRESSE POSTALE / MAILING ADDRESS:	314 CON. 8 CARLETON PLACE ON. K7C 3P1
GRADE / DEGREE:	**ANNÉE D'OBTENTION / YEAR GRANTED**
MSc. (MICROBIOLOGY AND IMMUNOLOGY)	2001
TITRE DE LA THÈSE / TITLE OF THESIS:	INVESTIGATION INTO THE POTENTIAL INVOLVEMENT OF APOPTOSIS IN THE PATHOGENIC MECHANISMS OF THE PROTOZOAN PARASITE TRICHOMONAS VAGINALIS

L'auteur permet, par la présente, la consultation et le prêt de cette thèse en conformité avec les règlements établis par le bibliothécaire en chef de l'Université d'Ottawa. L'auteur autorise aussi l'Université d'Ottawa, ses successeurs et cessionnaires, à reproduire cet exemplaire par photograhie ou photocopie pour fins de prêt ou de vente au prix coûtant aux bibliothèques ou aux chercheurs qui en feront la demande.

Les droits de publication par tout autre moyen et pour vente au public demeureront la propriété de l'auteur de la thèse sous réserve des règlements de l'Université d'Ottawa en matière de publication de thèses.

The author hereby permits the consultation and the lending of this thesis pursuant to the regulations established by the Chief Librarian of the University of Ottawa. The author also authorizes the University of Ottawa, its successors and assignees, to make reproductions of this copy by photographic means or by photocopying and to lend or sell such reproductions at cost to libraries and to scholars requesting them.

The right to publish the thesis by other means and to sell it to the public is reserved to the author, subject to the regulations of the University of Ottawa governing the publication of theses.

N.B. LE MASCULIN COMPREND EGALEMENT LE FEMININ

Nov. 9/01
DATE

Shannon Hayward-McClelland
(AUTEUR) SIGNATURE (AUTHOR)



Université d'Ottawa · University of Ottawa

Université d'Ottawa • University of Ottawa

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

HAYWARD-MCCLELLAND, Shannon

AUTEUR DE LA THÈSE - AUTHOR OF THESIS

M.Sc. (Microbiology and Immunology)

GRADE - DEGREE

Department of Biochemistry, Microbiology and Immunology

FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

Investigation Into the Potential Involvement of Apoptosis in the Pathogenic Mechanisms of the Protozoan Parasite *Trichomonas Vaginalis*

Gary Garber and Lionel Filion

DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

K. Copeland

J. Webb

J.-M. De Koninck, Ph.D.

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

SIGNATURE

DEAN OF THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

INVESTIGATION INTO THE POTENTIAL INVOLVEMENT OF APOPTOSIS IN THE PATHOGENIC MECHANISMS OF THE PROTOZOAN PARASITE *TRICHOMONAS VAGINALIS*

A Thesis Submitted to the

School of Graduate Studies

University of Ottawa

In Partial Fulfillment of the Requirements for the

Degree of Master of Science

Department of Biochemistry, Microbiology, & Immunology

Submitted by

Shannon Hayward-McClelland

©Shannon Hayward-McClelland, Ottawa, Canada, 2001

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-66050-8

Canada

ABSTRACT

The protozoan parasite *Trichomonas vaginalis* is the causative agent of the common sexually transmitted disease trichomoniasis. *T. vaginalis* infection varies greatly with respect to clinical severity and, while not life-threatening, can lead to a number of serious urogenital complications. Despite a great deal of research, the exact pathogenic mechanisms employed by the parasite in its attack on the human urogenital tract remain largely undefined. It is likely, given the wide variations in symptomology seen in those infected with *T. vaginalis*, that the parasite employs a number of different pathogenic mechanisms. These mechanisms include the ability to evade host immune responses, cellular adhesion, haemolysis, the excretion of soluble factors, and interaction with the resident urogenital flora. *In vitro* studies have shown that *T. vaginalis* causes a number of pathological changes in host cells and effectively destroys a variety of host cell monolayers. It has also been documented that many microorganisms, including a number of protozoan pathogens, are capable of inducing apoptosis in host cells. Our hypothesis tied together these two areas and proposed that *T. vaginalis* would induce apoptosis in host cells as part of its pathogenic assault. This hypothesis was tested in an established co-culture system; previous research demonstrated that the parasite destroyed McCoy cell monolayers. The TUNEL assay was used to identify cells undergoing apoptotic DNA fragmentation as a result of co-incubation with *T. vaginalis*. Early experiments suggested that apoptosis may have been occurring in the McCoy monolayers, but this was later disproved with the help of a sandwich ELISA to detect DNA/histone fragments that characteristically result from apoptosis. Instead, the cells that had originally been identified as apoptotic were found to be adherent trichomonads. While McCoy cell apoptosis in response to co-culture with *T. vaginalis* was not observed, *T. vaginalis* did, over time, cause detachment of almost all monolayer cells, and also caused irreparable damage in a portion of these cells. Induction of McCoy monolayer detachment was determined to be contact-independent, given that trichomonad culture supernatants had a similar effect.

ACKNOWLEDGEMENTS

I would like to take this opportunity to say thank you to many people who have, in one way or another, made the completion of my Master of Science thesis possible. Firstly, I want to thank my supervisor Dr. Gary Garber for his continual support throughout the time I've spent in his lab. Along with being a wonderful physician, he took the time to create a great working environment in his research lab, and I am in his debt for the opportunities he has provided. Thanks also to Cathy Moreau for the visits and for making life so much easier on a day to day basis. I also want to thank Dr. Andrew Badley and Dr. Lionel Filion who acted as my co-supervisors and who provided both technical and moral support throughout my research. My thesis advisory committee members, Dr. Kathryn Wright and Dr. Jonathan Angel, have also given me a great deal of their time and have helped to provide me with an excellent departmental support network.

I have had the good fortune to be able to work with great people in other departments as well, and I would like to extend to them my gratitude. Dr. Douglas Franks and Louise Pelletier, members of the Pathology Department, have been instrumental in allowing me to effectively process and capture images of the many tissue and cell samples involved in my research. Dr. Mary Senterman kindly took the time to teach me how to interpret the mouse vaginal smears that were a large part of the mouse model study. The staff of the Animal Care Department, particularly Dr. Marilyn Keaney and Dr. Susan Kilborn, taught me a great deal about small animal handling and the ethical considerations of such work.

Thanks to the friends I have made in the Garber/Cameron lab. Isabelle, Wally, Deb, Kiera and Dino have all been wonderful sources of support, both technical and personal, and have just made life a lot more fun. Finally, thank you to my family, my husband Jamie and especially to my daughter Peyton, who have made it all worth doing.

TABLE OF CONTENTS

ABSTRACT II

ACKNOWLEDGEMENTS III

TABLE OF CONTENTS IV

LIST OF FIGURES VII

LIST OF TABLES VII

LIST OF ABBREVIATIONS VIII

INTRODUCTION 1

TRICHOMONIASIS: SIGNIFICANCE AND EPIDEMIOLOGY 1
CLINICAL PICTURE OF UROGENITAL TRICHOMONIASIS IN WOMEN 2
CLINICAL PICTURE OF UROGENITAL TRICHOMONIASIS IN MEN 3
DETECTION AND DIAGNOSIS OF *T. VAGINALIS* INFECTION 4
TREATMENT OF *T. VAGINALIS* INFECTION 6
HOST DEFENSE MECHANISMS 8
TRICHOMONAS VAGINALIS- THE ORGANISM 9
Taxonomy 9
Morphology 9
Reproduction and Life Cycle 14
Nutritional Requirements and Cultivation 15
Strain Heterogeneity 16
T. VAGINALIS PATHOGENESIS 17
Immune Evasion 17
Surface Adhesins and Adherence to Host Cells 19
Haemolysis 24
Proteinases 26
Cell Detaching Factor (CDF) 27
Interaction with Vaginal Flora 30
APOPTOSIS-A POTENTIAL PATHOGENIC MECHANISM EMPLOYED BY *T. VAGINALIS*? 31
Defining Apoptosis 33
The Caspase Family of Cysteine Proteinases 36
HYPOTHESIS AND OBJECTIVES 39
Hypothesis 39
Objectives 39

MATERIALS AND METHODS 41

CELL CULTURE 41
Growth and Maintenance of McCoy Cells 41
Growth and Maintenance of Jurkat Cells 42
CULTURE OF MICROORGANISMS 42
Trichomonas vaginalis 42
IN VITRO CO-INCUBATION EXPERIMENTS 43
T. vaginalis and McCoy Cell Monolayers 43
T. vaginalis Growth Media (Cell Free Filtrate) and McCoy Cell Monolayers 44
CYTOSPINS TO RECOVER DETACHED CELLS 44
INDUCTION OF APOPTOSIS IN MCCOY CELLS USING CAMPTOTHECIN (POSITIVE CONTROL) 44
CASPASE INHIBITION IN MCCOY CELLS 45
JURKAT APOPTOSIS CONTROL USING ANTI-CD95 ANTIBODY 45

Detection of Apoptosis Using the TUNEL Assay 46
Detection of Apoptosis Using the Sandwich ELISA 46
Cell Enumeration and Image Capture 47

RESULTS 48

Measurement of DNA Strand Breaks in McCoy Cells Using the TUNEL Assay 48
Role of Contact Between *T. vaginalis* and McCoy Cells in Monolayer Destruction and Induction of McCoy Cell Apoptosis 53
Role of Caspases in *T. vaginalis*-Induced McCoy Cell Apoptosis 56
Adherence of *T. vaginalis* to Monolayer Following Detachment of McCoy Cells 63
Induction of Apoptosis in McCoy Cells Using Camptothecin (Positive Apoptosis Control) .. 66
Measurement of DNA/Histone Fragments Resulting from Apoptosis by ELISA Assay (An Alternate Means of Assessing Apoptosis) 69
Assessment of Adherent *T. vaginalis* via the TUNEL Assay 71

DISCUSSION 73

Detection of Apoptosis 73
McCoy Monolayer/*T. vaginalis* Co-Cultures: TUNEL and *T. vaginalis* Staining 75
Investigating Involvement of Caspases in the McCoy Cell Death Process 77
A Positive Apoptosis Control in McCoy Cells 78
ELISA Experiments Support a Lack of Apoptosis in McCoy Cells 79
McCoy Monolayer/*T. vaginalis* Co-Culture: Pathology 79
Stages of Monolayer Destruction Following Exposure to *T. vaginalis* 80
Appearance of *T. vaginalis* On Glass Substrate 82
Timing of Monolayer Destruction and Parasite/Cell Ratios 82
Ultimate Fate of Detached McCoy Cells Following Co-Culture With *T. vaginalis* 85
Cell Free Filtrate Experiments-Cellular Detachment Without Trichomonad Contact 88
Conclusions 92

REFERENCE LIST 94

APPENDIX I: 114

THE INTRAVAGINAL MOUSE MODEL OF *TRICHOMONAS VAGINALIS* INFECTION REVISITED 114

INTRODUCTION 114

T. vaginalis Vaccine Potential 114
In vivo Models of *T. vaginalis* Infection 115
The Intravaginal Mouse Model of Infection 117
The Intravaginal Mouse Model of Infection and the Anti-*T. vaginalis* Vaccine 120
Hypothesis and Objectives 122
Hypothesis 122
Objectives 122

MATERIALS AND METHODS 123

Microorganisms 123
T. vaginalis 123
Lactobacillus acidophilus 123
Animal Care 125
Pretreatment of Mice with Estrogen 125
Determination of Mouse Estrous Stage 126
Vaginal Inoculation of Mice 126
Vaginal Washes in Mice 127

RESULTS 128

THE ROLE OF THE *T. VAGINALIS* ISOLATE IN THE MOUSE INFECTION MODEL 128
THE IMPORTANCE OF ESTROUS STAGE IN ESTABLISHMENT OF THE INTRAVAGINAL INFECTION 129
L. ACIDOPHILUS AND ITS ROLE IN THE MAINTENANCE OF A LONG-TERM INFECTION IN THE MICE 137
INFLUENCE OF TIMING OF VARIOUS ELEMENTS OF THE INFECTION PROTOCOL 139

DISCUSSION 143

DEVELOPMENT OF A MOUSE MODEL FOR VACCINE STUDIES IN OUR LAB 143
PROBLEMS ENCOUNTERED IN REESTABLISHMENT OF THE INFECTION MODEL 145
ASSESSMENT OF THE IMPORTANCE OF THE *T. VAGINALIS* ISOLATE IN ESTABLISHMENT OF AN INFECTION 145
ESTROGENIZATION AND ESTROUS STAGE AND THEIR PART IN THE INFECTION PROTOCOL 146
IMPORTANCE OF *L. ACIDOPHILUS* IN ESTABLISHMENT OF INFECTION 149
OVERALL TIMING OF VARIOUS ELEMENTS IN PROTOCOL 150
CONCLUSIONS 150

APPENDIX II 152

RECIPE FOR DIAMOND'S TYI-S 33 MEDIUM 152
RECIPE FOR PHOSPHATE BUFFERED SALINE (PBS) 10X STOCK SOLUTION 153

LIST OF FIGURES

Figure 1: Scanning Electron Micrographs Illustrating Surface Morphology of *T. vaginalis* 10
Figure 2: Diagrams of *T. vaginalis* Illustrate Important Internal and External Features 11
Figure 3: Light Microscopy and Transmission Electron Microscopy Illuminate Key Internal Structures in the *T. vaginalis* Cell 13
Figure 4: Co-Culture Illustrating Three Cellular Morphologies 49
Figure 5: Rat Mammary Gland Stained with TUNEL Assay Reveals Apoptotic Cells 51
Figure 6: Addressing the Need For Contact in *T. vaginalis* Pathology 57
Figure 7: Effect of Caspase Inhibitors on McCoy Cells 58
Figure 8: Control Experiment to Verify Efficacy of Caspase Inhibitors 61
Figure 9: Lack of Apoptosis in Detached McCoy Cells 65
Figure 10: Apoptosis of McCoy Cells Induced by CAM 68
Figure 11: Cell Death Detection ELISA 70
Figure 12: Morphology of *T. vaginalis* Adherent to Glass Substrate (Alone and in Co-Culture with McCoy Cells) 72
Figure 13: Mouse Vaginal Smears to Determine Stage of Estrous 131
Figure 14: Effect of Estrogen Delivery Method on Estrous Stage 133
Figure 15: Effect of Extra Estrogen Injections on Estrous Stage 136

LIST OF TABLES

Table 1: Initial TUNEL Experiments 52
Table 2: Timecourse Experiment 54
Table 3: Examining the Role of Contact in *T. vaginalis* Pathogenesis 55
Table 4: Investigating the Involvement of Caspases 59
Table 5: Caspase Inhibitor Control 62
Table 6: Assessing Detached Monolayer Cells 64
Table 7: Apoptosis Control in McCoy Cells 67
Table 8: Intravaginal Mouse Infection Protocol 124
Table 9: Importance of the *T. vaginalis* Isolate in the Mouse Model 130
Table 10: Estrogen Delivery Method and Establishment of Intravaginal Infection 135
Table 11: Estrogen Dosage and Establishment of Intravaginal Infection 138
Table 12: *L. acidophilus* Concentration and Establishment of Intravaginal Infection 140
Table 13: Timing of *T. vaginalis* Inoculation and Establishment of Intravaginal Infection 142

LIST OF ABBREVIATIONS

ABTS	2,2'-azino-di- [3-ethylbenzthiazoline sulfonate]
AIDS	acquired immunodeficiency syndrome
Apaf-1	apoptotic protease-activating factor
ATCC	American Type Culture Collection
Ca^{2+}	calcium ion
CAM	camptothecin
caspase	cysteinyl aspartate-specific proteinase
CDC	Centers for Disease Control and Prevention
CDF	cell detaching factor
cfu	colony forming units
CHO	Chinese hamster ovary
CO_2	carbon dioxide
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
EIA	enzyme-linked immunoassay
ELISA	enzyme-linked immunosorbent assay
FBS	fetal bovine serum
FITC	fluorescein isothiocyanate
g	gram
GST	glutathione transferase
^3H	tritium
Hep-2	human epithelial
HEPES	N-[2-hydroxyethyl] piperazine-N'-[2-ethanesulfonic acid]
HIV	human immunodeficiency virus

HSV	herpes simplex virus
IAP	inhibitor of apoptosis protein
IgA	immunoglobulin A
IgG	immunoglobulin G
IgM	immunoglobulin M
ISNT	*in situ* nick translation
kb	kilobase
KCl	potassium chloride
kDa	kilodalton
KH_2PO_4	potassium dihydrogen phosphate
L	liter
mg	milligram
mL	milliliter
mm	millimeter
mM	millimolar
MRS	deMan, Rogosa, and Sharpe (lactobacilli media)
N	normal
NaCl	sodium chloride
Na_2HPO_4	sodium monohydrogen phosphate
NaOH	sodium hydroxide
NH_2	amino group
nm	nanometer
O_2	oxygen
OH	hydroxyl group
PBS	phosphate buffered saline
P:C	parasite to target cell ratio

PFA	paraformaldehyde
PI	propidium iodide
POD	peroxidase
PVC	polyvinyl chloride
rpm	revolutions per minute
RPMI	Roswell Park Memorial Institute
STD	sexually transmitted disease
TNFR	tumor necrosis factor receptor
TUNEL	terminal deoxynucleotidyl transferase (TdT)-mediated 2'-deoxyuridine-5'-triphosphate (dUTP) nick end labeling
Tv	*Trichomonas vaginalis*
TYI	Diamond's TYI-S-33 medium (Trypticase, yeast extract, iron-serum)
U	units
µg	microgram
µL	microliter
µm	micrometer
µM	micromolar
WHO	World Health Organization
z-IETD-fmk	z-Ile-Glu(OMe)-Thr-Asp(OMe)-fluoromethylketone
z-VAD-fmk	z-Val-Ala-Asp(OMe)-fluoromethylketone

INTRODUCTION

Trichomoniasis: Significance and Epidemiology

Trichomoniasis is a sexually transmitted disease (STD) of global importance (World Health Organization, 1995) and it is caused by the protozoan parasite *Trichomonas vaginalis*. While once considered to be a disease of little importance, trichomoniasis has now been accepted to have economic, cultural and medical repercussions (Hook, 1999). With an annual incidence of more than one hundred and seventy million cases, it has become the most prevalent non-viral STD (World Health Organization, 1995). The same WHO document shows that eight million new cases of *T. vaginalis* infection are reported each year in North America alone. Given that up to fifty percent of infections may exist without any symptoms (Fouts and Kraus, 1980), these infection rates are likely underestimated. Asymptomatic infections are especially problematic because carriers may transmit the parasite unknowingly to others (Nicoletti, 1961). *T. vaginalis* is an obligate parasite that is almost always transmitted via sexual contact (Heine and McGregor, 1993). Although the organism lacks a cystic stage, it has been shown to survive outside its host for a number of hours (Honigberg, 1978; Lossick, 1990a). While it is theoretically possible for urogenital trichomonads to be transmitted non-sexually, there has been no documentation of this happening (Heine and McGregor, 1993). Trichomoniasis has a cosmopolitan distribution and cases have been documented throughout all continents, races, and socio-economic strata. It has been suggested, however, that certain populations are more affected by the disease than others. Risk factors for the acquisition of trichomoniasis appear to include multiple sexual partners, low education and income, a previous history of STDs, and the non-use of contraceptives (Lossick, 1990a; Heine and McGregor, 1993). The true prevalence of trichomoniasis is difficult to determine since sampling is not random and is biased by the fact that most surveys involve those people visiting STD clinics with symptoms of *T. vaginalis* infection or other STDs (Lossick,

1990a). Unfortunately, *T. vaginalis* is not a reportable disease in most countries, so accurate data with respect to disease morbidity and infection trends are difficult to collect (Lossick, 1990a).

Clinical Picture of Urogenital Trichomoniasis in Women

T. vaginalis primarily infects the squamous epithelium of the female urogenital tract. The infection is multi-focal and parasites can be detected in the vagina, cervix, urethra, bladder, and in the Skene's, Bartholin's, and periurethral glands (Honigberg, 1978; Wolner-Hanssen *et al.*, 1989; Krieger, 1995). Trichomoniasis is generally a disease of the reproductive years, and is rarely reported before puberty or after menopause (Honigberg, 1978). The incubation period is generally in the range of a few days to a month; infection may persist for a protracted period of time. Clinical presentation in women can range from asymptomatic to flagrant vaginitis to a chronic carrier state (Wolner-Hanssen, 1989; Rein, 1990; Heine and McGregor, 1993). In addition, symptoms can vary over time in a given patient. Rein (1990) has suggested that over thirty percent of asymptomatically infected women become symptomatic within a period of six months. While the phenomenon of disparity among patient clinical profiles is not completely understood, the explanation likely involves differences in the inherent virulence of individual strains and also variation in host susceptibility (Rein, 1990).

The prevalence of specific symptoms in *T. vaginalis*-infected women differs greatly from study to study. Acute trichomoniasis has been reported to present with any or all of the following symptoms: vaginal discharge, vaginitis, vulvitis, vulvovaginal irritation or pruritis, dyspareunia, and dysuria (Rein, 1990). Vaginal, and less frequently, vulvular erythema has also been found in women infected with *T. vaginalis*. Colpitis macularis or "strawberry cervix", which is created by capillary dilatation and punctate hemorrhages on the vaginal walls and the cervix, can be detected by the naked eye in only two percent of cases, but can be found by colposcopy in as many as

ninety percent of *T. vaginalis*-infected women (Fouts and Kraus, 1980; Wolner-Hanssen *et al*, 1989; Krieger *et al.*, 1990; Heine and McGregor, 1993). In general, the pH of the vagina also changes through the course of infection, with a rise from the normal 4.5 to greater than 5.0 (Rein and Chapel, 1975; Rein, 1990).

Trichomoniasis symptoms and parasite load have been shown to be cyclical and tend to worsen during menstruation (Rein and Chapel, 1975; Spiegel, 1990; Heine and McGregor, 1993). Menstrual blood may serve as a stimulant or nutrient for the parasites. The elevation in vaginal pH, with a coincident decrease in oxidation-reduction potential seen during menses, may also enhance growth of anaerobic microorganisms like *T. vaginalis* (Spiegel, 1990).

T. vaginalis infection can lead to a number of serious complications in women. These include premature rupture of the placental membranes, premature labour, low birth weight infants, adnexitis, pyosalpinx, endometritis, atypical pelvic inflammatory disease (Heine and McGregor, 1993), cervical cancer (Wolner-Hanssen *et al.*, 1989; Viikki *et al.*, 2000), cervical erosion (Rein and Chapel, 1975) and infertility (Rein, 1990). It has also been suggested that infection with urogenital trichomonads can increase the risk of transmission of other STDs such as herpes simplex virus (HSV) (Pindak *et al.*, 1989) and human immunodeficiency virus (HIV) (Cameron and Padian, 1990; Laga *et al.*, 1993, 1994).

Clinical Picture of Urogenital Trichomoniasis in Men

Although male trichomoniasis has generally been considered asymptomatic, it has been concluded that thirty to fifty percent of all men with trichomonal infections show signs of mild symptomatic disease. These symptoms include urethral discharge, urethritis, and prostatitis (Krieger, 1995). The infection has been linked to infertility and a variety of other complications

and inflammatory diseases of the male genitourinary tract (Krieger, 1995).

It has not been determined with any certainty why the clinical picture of infection is so different in males and females. Zinc, which has been shown to be present in the prostatic fluid, may be cytotoxic to *T. vaginalis* and hence limit infection in men (Krieger and Rein, 1982). The reducing environment of the vagina may also contribute to the symptoms seen in women, via the activation of certain pathogenic mechanisms. The oxidative environment of the male genital tract may, on the other hand, inhibit such mechanisms (Alderete and Provenzano, 1997).

Detection and Diagnosis of *T. vaginalis* Infection

Diagnosis of trichomoniasis can be difficult due to the fact that many of the symptoms mimic those of candidiasis and bacterial vaginosis (Rein and Holmes, 1983) or of other STDs (Jirovec and Petru, 1968). Furthermore, some of the "hallmark" signs of *T. vaginalis* infection, like the "strawberry cervix", are actually observed in only a minority of patients (Fouts and Kraus, 1980). Diagnosis should therefore be made based on clinical presentation and the positive identification of trichomonads in the urogenital tract via a laboratory investigation.

The traditional method of *T. vaginalis* detection has been the direct microscopic examination of motile trichomonads in vaginal secretions. This technique is simple and quick, but has been shown to have widely varying specificity, depending on the collection of the specimen, the amount collected, and the number of motile parasites present; characteristic *T. vaginalis* motility can be lost rapidly. The sensitivity of the wet mount preparation ranges from thirty-eight to eighty-two percent (McMillan, 1990) and the technique requires the presence of at least 10^4 organisms per mL to give positive results.

T. vaginalis can also be detected by microscopic observation of fixed specimens. A variety of staining techniques has been reviewed in the literature (see McMillan, 1990 for details and references). These techniques include Gram-, Papanicolaou-, Romanowsky-, Acridine orange-, silver, Diff-Quick, and Periodic acid-Schiff staining. While these procedures are quick and permanent, their diagnostic specificity and sensitivity is questionable, even when the slides are reviewed by a skilled technician (Lossick and Kent, 1991). One limiting factor to consider is the fact that the typical morphological characteristics of the trichomonads may be lost during staining and fixation (Bhatt *et al.*, 1996).

Having the ability to amplify the number of trichomonads from an original sample, the broth culture technique (with a high degree of sensitivity (ninety-two to ninety-five percent)) has been dubbed the "gold standard" for diagnosis of *T. vaginalis* infection (Bhatt *et al.*, 1996). Interpretation is simple and only a few hundred trichomonads/mL are required to initiate growth (Garber *et al.*, 1987). The drawbacks to this technique include: elevated cost associated with the procedure, lack of media on hand in many labs, and a delay in diagnosis to allow for growth of the specimen (Bhatt *et al.*, 1996). Other researchers have used cell-culture technique to recover trichomonads from clinical samples. Garber *et al.* (1987) found that the cultivation of *T. vaginalis* in McCoy cell monolayers was extremely sensitive (detecting as few as three organisms per mL), but this method is not routinely used because of the associated inconvenience and higher cost.

More sophisticated methods for the diagnosis of *T. vaginalis* infection have been introduced, but few are in routine laboratory usage due to the fact that they are not always cost-effective. These *T. vaginalis*-detecting techniques include latex agglutination (Carney *et al.*, 1988), specific dot-enzyme immunoassays (Gombosova and Valent, 1990), direct immunofluorescence assays (Krieger *et al.*, 1988; McMillan, 1990), and monoclonal-based enzyme-linked immunoassay

(EIA) (Lisi *et al.*, 1988). Other diagnostic procedures rely on detection of anti-trichomonal antibodies (Street *et al.*, 1982; Sibau *et al.*, 1987) or trichomonal DNA (Rubino *et al.*, 1991; van Der Schee *et al.*, 1999) in a given sample.

Treatment of *T. vaginalis* Infection

Prior to 1959, people infected with *T. vaginalis* relied on local topical preparations to relieve some of the symptoms of trichomoniasis. Protracted infection and re-infection was commonplace, however, because topical treatments did not penetrate to all sites harbouring trichomonads, and were unsuitable for treatment of infected men (Lossick and Kent, 1991).

In the early 1960s, the treatment of trichomoniasis was radically altered with the introduction of metronidazole (α-hydroxyethyl-2-nitroimidazole), a 5-nitroimidazole derivative of a Streptomyces spp. antibiotic, azomycin. Metronidazole is the only nitroimidazole currently available in North America, but other structurally related nitroimidazoles with varying pharmacokinetic and therapeutic activities, including tinidazole, secnidazole, ornidazole and mebendazole, are available elsewhere (Rossignol *et al.*, 1984; Lau *et al.*, 1992; Mathisen and Finegold, 1992; Gillis and Wiseman, 1996).

A number of oral metronidazole treatment regimens for *T. vaginalis* infection have been described, and with proper compliance, most are equally effective in both males and females (Lossick, 1982; Lossick, 1990b). The Center for Disease Control (CDC) guidelines currently suggest that metronidazole be given orally in a single 2 g dose or 250 mg twice a day for seven days (CDC, 1998). Successful treatment of infection is observed in ninety to ninety-five percent of cases; simultaneous treatment of sexual partners increases the success rate (CDC, 1998).

Metronidazole enters the *T. vaginalis* cell by a process of passive diffusion. The drug, itself inactive, is then reduced anaerobically in the hydrogenosome by pyruvate-ferredoxin oxidoreductase to yield short-lived nitro radical intermediates. A concentration gradient is formed, allowing more metronidazole to enter the cell. The highly reactive intermediates damage the parasite's DNA (Ings *et al.*, 1974; Muller, 1981; Moreno *et al.*, 1984; Edwards, 1993). The cell responds rapidly with the cessation of cell division and motility; within eight hours, cell death occurs (Nielson, 1976).

Metronidazole treatment is curative in a large majority of cases, but occasionally, treatment failure can occur. A variety of explanations have been put forth to try to explain refractory cases of trichomoniasis. Once non-compliance and re-infection were eliminated, researchers proposed ineffective delivery of the drug to the urogenital area, inactivation of the drug by vaginal bacteria (McFadzean *et al.*, 1969; Ingham *et al.*, 1979; Edwards *et al.*, 1979), and incomplete drug absorption (Kane *et al.*, 1961) as possible explanations for treatment failure.

Finally, it was proposed that while these explanations were possible, the likely reason for treatment failure was that *T. vaginalis* isolates could possess true metronidazole resistance (Lumsden *et al.*, 1988). In 1989, the CDC estimated that five percent of all *T. vaginalis* isolates possessed metronidazole resistance of varying degrees (Narcisi and Secor, 1996). Resistance can be caused by a number of different mutational changes in the *T. vaginalis* genome and can affect both aerobic and anaerobic mechanisms of metabolism (Upcroft and Upcroft, 2001). In aerobic metronidazole resistance, the transcription of the ferredoxin gene is diminished, thereby hampering the cell's ability to activate the drug (Quon *et al.*, 1992). In examples of anaerobic resistance, the activities of pyruvate-ferredoxin oxidoreductase and hydrogenase are either impaired or absent (Kulda *et al.*, 1993).

Host Defense Mechanisms

There exists a complicated host defense system, which plays an important role in restricting a *T. vaginalis* infection to the genitourinary compartment. Prior exposure induces no real host protection against re-infection, despite the presence of an initial measurable host immune response (humoral and cellular) (Honigberg, 1986; Ackers, 1990). Why protection does not occur is largely unknown, but may involve induction of an inappropriate or low-grade response or may be related to immune evasion by the parasite itself.

The host defense network is comprised of three components: nonimmunologic factors (zinc concentration and iron availability), innate immune responses (complement, neutrophils, natural antibodies, and macrophages), and acquired immune responses (antibody production and cell-mediated immunity) (Landolfo *et al.*, 1980; Demes *et al.*, 1987; Ackers, 1990; Shaio and Lin, 1995).

While the specific host immune response remains the least well-defined aspect of *T. vaginalis* immunology, many researchers have documented the presence of circulating serum and local cervicovaginal antibodies (IgG, IgA, and IgM) to *T. vaginalis* (Mason, 1979; Su, 1982; Street *et al.*, 1982; Alderete, 1984; Cogne *et al.*, 1985; Sibau *et al.*, 1987 and Sharma *et al.*, 1991). There is limited evidence of what effect these antibodies may have on the parasite. According to Honigberg (1986), antibody titers in *T. vaginalis* infection progressively decline after eradication of the parasite by treatment. Within six to twelve months, all antibodies disappear from circulation, leaving the host with no specific immune defense. The existence of cell mediated immunity in response to *T. vaginalis* infection has also been documented, but its role remains largely undefined (Yano *et al.*, 1983; Mason and Patterson, 1985).

Trichomonas vaginalis- The Organism

Taxonomy

The eukaryotic, aerotolerant anaerobic (Paget and Lloyd, 1990) protozoan parasite *T. vaginalis* was first identified by Donné (Donné, 1836). The taxonomic status of *T. vaginalis* can be described as: Subkingdom Protozoa, Phylum Sarcomastigophora, Subphylum Mastigophora, Class Zoomastigophorea, Superorder Parabasalidea, Order Trichomonadida, Family Trichomonadidae, and Subfamily Trichomonadinae (Schmidt and Roberts, 1981; Marquardt and Demaree, 1985; Honigberg, 1990).

Morphology

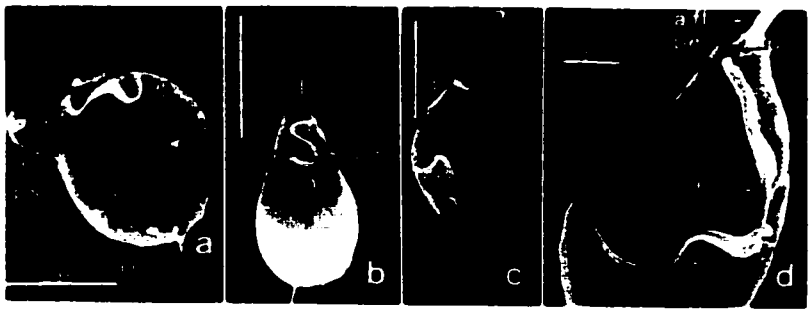
There have been a number of studies into *Trichomonas* structure and morphology. Both light and electron microscopic examinations of the organism have contributed to the current body of information. The shape of *Trichomonas vaginalis* tends to be variable depending upon whether the specimen is fixed or living. The method of fixation itself can influence the shape and size of the trichomonad (Honigberg and King, 1964). On average, *T. vaginalis* has been shown to be approximately 10 μm in length and 7 μm in width (Honigberg and King, 1964) (Figure 1a-d). Shape tends to be more uniform among trichomonads grown in non-living axenic culture than in those found in vaginal secretions. In axenic culture, they tend to be ovoid, ellipsoidal, or spheroidal (See Figure 1e). *In vivo*, however, *T. vaginalis* tends to be more ameboid with distinct pseudopodial extensions (Heath, 1981; Arroyo *et al.*, 1993) (Figure 1f).

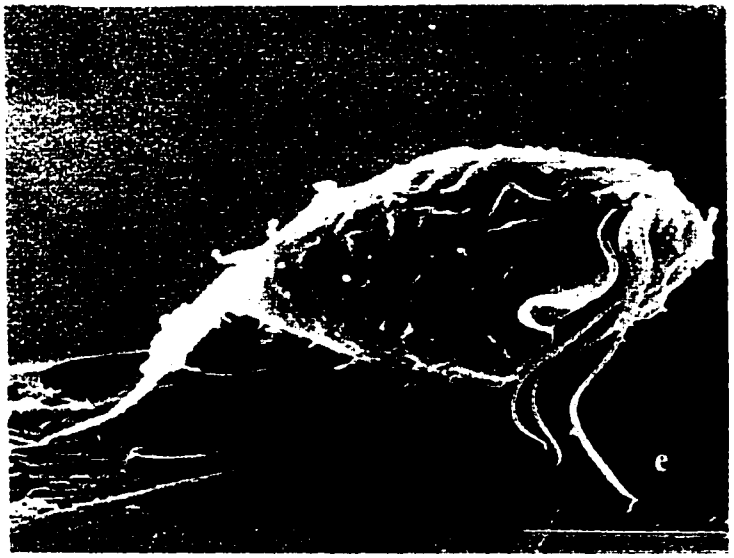
Regardless of its shape, the *T. vaginalis* cell bears many interesting surface and internal features. One of the most striking features of the trichomonad is its collection of flagella (Figures 1a-f, 2a and 2b). Emerging from the anterior pole of the cell are the four anterior flagella which originate

Figure 1: Scanning Electron Micrographs Illustrating Surface Morphology of *T. vaginalis*

The anterior flagella (a.fl.) can be seen emerging from the periflagellar canal (p.c.) in both the dorsolateral **(a)** and dorsal **(b-d)** views of *T. vaginalis*. The recurrent flagellum (r.fl. in **d**) emerges from the periflagellar canal posterior and dorsal to the anterior ones. The wall of the canal is reinforced by the pelta (pe. in **d**). The margin of the undulating membrane (u.m. in **b**) is made up of the attached portion of the recurrent flagellum (inner component) and the outer accessory filament (ac.f.)(outer component) **(d)**. (Reproduced, with permission, from Warton and Honigberg, 1979).

T. vaginalis grown to late logarithmic phase in suspension media possesses an elipsoidal, ovoid shape **(e)** while a transformation to an ameboid, "fried-egg" appearance can be observed in the same isolate after only five minutes of contact with vaginal epithelial cells **(f)**. (Reproduced from Arroyo *et al.*, 1993 with permission of publisher).


a
b
c
d

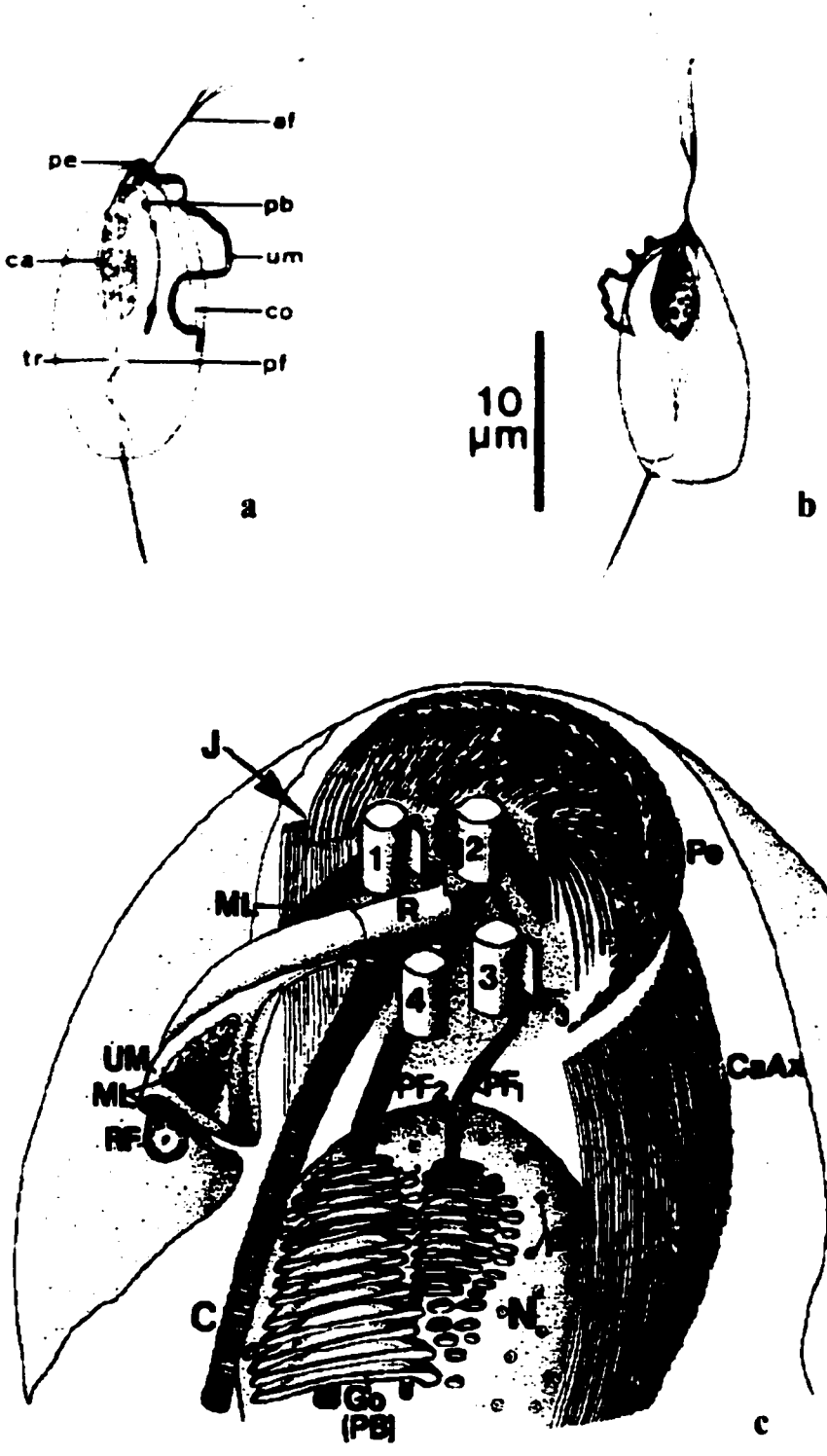

e


f

Figure 2: Diagrams of *T. vaginalis* Illustrate Important Internal and External Features

Left **(a)** and right **(b)** views of *T. vaginalis* highlight all the mastigont organelles: these include the anterior flagella (af), pelta (pe), V-shaped parabasal body (pb) with unequal arms, and parabasal filaments (pf), undulating membrane (um), along with the capitulum (ca), and trunk (tr) of the axostyle. (Reproduced, with permission, from Honigberg and Brugerolle, 1990).

(c) Schematic diagram of *T. vaginalis* karyomastigont system (including the nucleus). The kinetosomal complex, which includes the parallel kinetosomes (#1 to #4) of the anterior flagella and the kinetosome R of the recurrent flagellum (RF) can be seen clearly. The filaments or lamellae (F_1, F_2, F_3, and X) connect the kinetosomes with one another or with other organelles, and anchor the complex to various areas of the cytoplasm. The costa (C) supports the undulating membrane (UM) along the dorsal surface of the cell. The undulating membrane (UM) is composed of a dorsal fold of the cytoplasmic membrane (which encloses the marginal lamella (ML)) and of the attached recurrent flagellum (RF). The parabasal apparatus is made up of parabasal filaments (PF_1 and PF_2) and a parabasal body (PB) (two Golgi complexes (Go)). The peltar-axostylar consists of the pelta (Pe), which supports the wall of the periflagellar canal from which emerge the flagella, and of the axostyle. The pelta and axostyle are connected along the peltar-axostylar junction (J). The anterior portion of the axostyle, the capitulum (CaAx), can be seen here. Finally, the nucleus (N), with an envelope containing many pores (P), is apposed to the dorsal concave surface of the capitulum. (Reproduced, with permission, from Honigberg and Brugerolle, 1990).


ef
pe
pb
ca
um
co
tr
pf
a
10
µm
b
J
1
2
Pe
ML
R
4
3
UM
ML
RF
PF2
PF1
CaAx
C
N
Gb
(PB)
c

in the kinetosomal complex (Figure 2c). Slightly posterior and dorsal to where the anterior flagella originate, the kinetosomal complex gives rise to the undulating membrane and the supporting costa (Figure 2c). The undulating membrane, is made up of the outer accessory filament (an extension of the outer membrane and the cytoplasm of the organism) and a fifth flagellum (termed the recurrent flagellum). Together, they extend a portion of the way along the body of the trichomonad. The recurrent flagellum ends with the undulating membrane and does not continue beyond as a free posterior flagellum. *T. vaginalis* derives its characteristic corkscrew motility from the action of the flagella and the undulating membrane (Honigberg, 1978; Honigberg and Brugerolle, 1990).

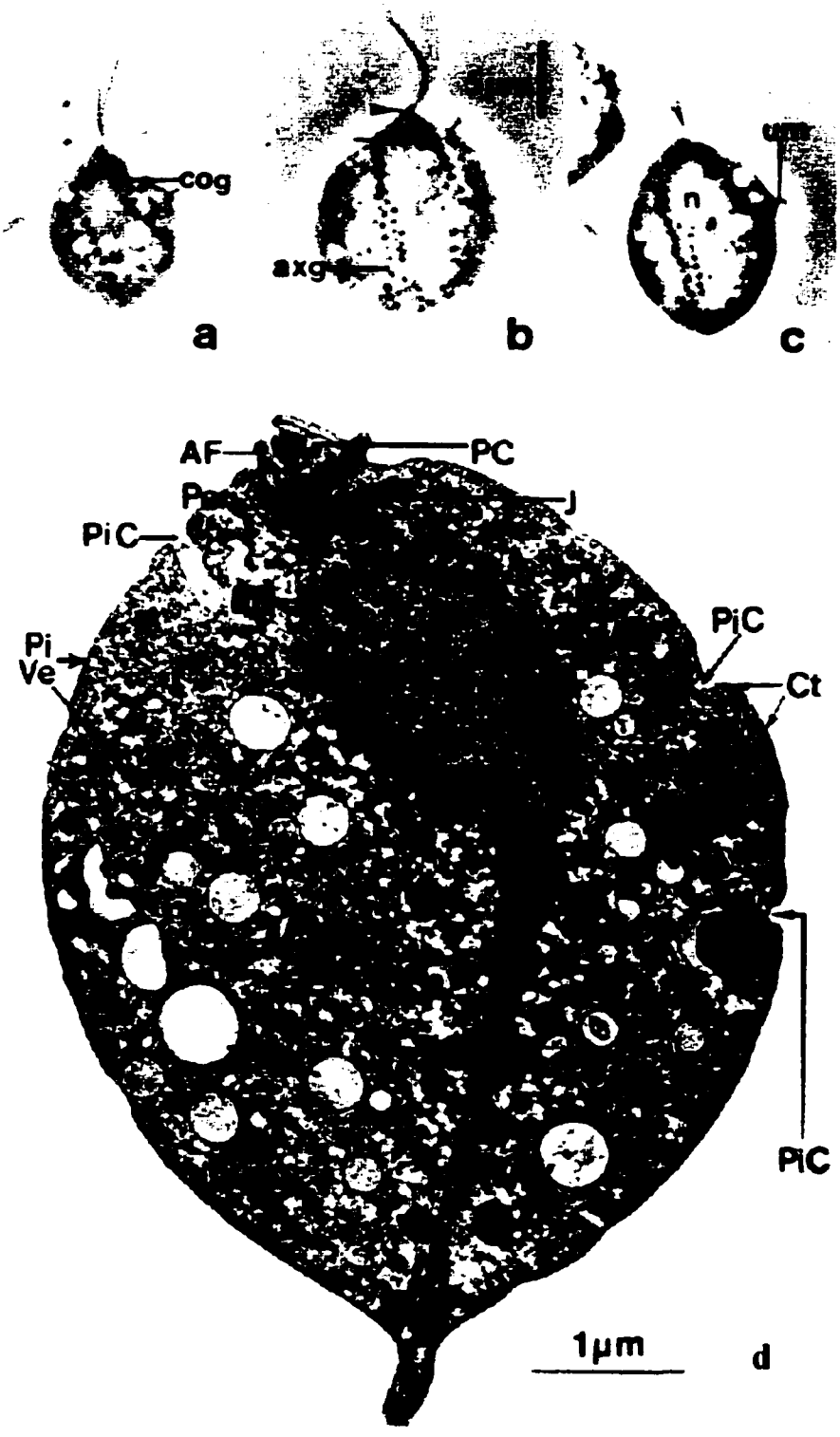
Located anteriorly, an elongated nucleus with dimensions of approximately 4 μm by 2 μm, can be found with the left ventral surface next to the spatulate axostylar capitulum. At its anterior end, the capitulum is connected to the pelta, a crescent-shaped structure that engulfs the area from which the anterior flagella protrude (Figure 2c). The capitulum continues, posterior to the nucleus, and narrows to give rise to a slender rod-like axostylar trunk down the centre of the cell. The axostyle protrudes through the posterior end of the cell body and eventually tapers to a point that is believed to play a role in anchoring the trichomonad to host tissues (Figure 3d). The nucleus is also surrounded by endoplasmic reticulum (Figure 3d). Situated dorsally and to the right of the nucleus, there can be found a parabasal apparatus. Typically V-shaped, it consists of a well-developed parabasal body (two Golgi complexes) and two parabasal filaments linking the apparatus to the kinetosomal complex. Each Golgi complex consists of stacks of cisternae surrounded by vesicles (Honigberg and Brugerolle, 1990) (Figure 2c).

One of the most consistent features of the *T. vaginalis* cell is the collection of granules termed the paracostal and paraxostylar granules (Figure 3a-d). The paracostal granules are found on either side of the filamentous costa, while the paraxostylar granules are located in three rows that

Figure 3: Light Microscopy and Transmission Electron Microscopy Illuminate Key Internal Structures in the *T. vaginalis* Cell

(a-c) Motile organisms under the light microscope show nuclei (n), anterior flagella, undulating membrane (um), as well as paraxostylar (axg) and paracostal (cog) granules or hydrogenosomes. Note the parallel rows of the paraxostylar granules, which constitute a diagnostic characteristic of *T. vaginalis*. The flagella emerge from the periflagellar canal (arrowhead in **b**). The region of the parabasal apparatus is also visible (small arrow in **b**). (Reproduced, with permission, from Honigberg and Brugerolle, 1990).

(d) A transmission electron micrograph of the organism. The anterior flagella (AF) originate from the periflagellar canal (PC), which is supported by the pelta (Pe). The nucleus (N) is lodged in the depression of the axostylar capitulum (CaAx). Endoplasmic reticulum (ER) can be found in the nuclear region. The axostylar capitulum narrows into the axostylar trunk (TrAx), which projects beyond the posterior surface of the cell. Hydrogenosomes (H) are lined up along the axostylar trunk, while vesicles (V) can be seen throughout the cytoplasm. The cell surface is marked with depressions leading to pinocytotic canals (PiC). Note the cell coat (Ct) visible along the right side of the cell. Pinocytotic vesicles (PiVe) can be found in the cytoplasm. (Reproduced, with permission, from Honigberg and Brugerolle, 1990).


cog
axg
n
a
b
c
AF
PC
J
PiC
Pi
Ve
PiC
Ct
PiC
1µm
d

parallel the axostyle, from its anterior originating point to the posterior area where it emerges from the cell body. These granules can be seen in living trichomonads and in many stained preparations. Most of the paraxostylar and paracostal granules are in fact electron-dense hydrogenosomes, which have been shown to play a vital role in cell metabolism (Honigberg and King, 1964; Honigberg, 1978; Honigberg and Brugerolle, 1990). The hydrogenosome is a double membrane-bound organelle; unlike mitochondria, the inner membrane of the hydrogenosome does not form cristae (Muller, 1988). Being a primitive eukaryotic organism, *T. vaginalis* lacks mitochondria and it is the function of the hydrogenosomes to carry out analogous metabolic functions (Muller, 1993). Studies have shown that the major role of these organelles is the anaerobic conversion of pyruvate to acetate, malate, CO_2, and H_2. This process is accompanied by the substrate-level phosphorylation of ADP to ATP (Muller, 1988).

The trichomonad cell membrane is covered with a thin cell coat. Structural and metabolic studies have indicated that endocytosis encompasses both phagocytosis and pinocytosis. Throughout the cytoplasm there can be found food vacuoles of varying sizes. Pinocytic canals and pinocytic vesicles can be seen at or near the trichomonad cell surface (Honigberg and Brugerolle, 1990) (Figure 3d).

Reproduction and Life Cycle

Like many other protozoan parasites, *T. vaginalis* exists only in the trophozoite stage and lacks a true cystic stage (Honigberg and Brugerolle, 1990). When faced with unfavourable environmental conditions, trichomonads may round up and internalize their locomotor organelles, forming what have been described by some as "pseudocysts". Since these forms have not been found to give rise to normal motile trophozoites, however, they may represent a degenerate form of the organism (Honigberg and Brugerolle, 1990).

As in all trichomonads, reproduction in *T. vaginalis* occurs via a process termed cryptopleuromitotic division. This process involves longitudinal binary fission without the concomitant disappearance of the nuclear membrane (Brugerolle, 1975). The process begins with the appearance of a duplicate kinetosome and flagellum as well as the development of two attractophores (described as "bell-clapper shaped" organelles) which give rise to the poles of division. Originating from these attractophores, microtubules join to form one or more bundles. The extranuclear spindle created by these bundles is termed the paradesmose. Chromosomal microtubules, also originating from the attractophores, attach to the nuclear chromosomal kinetochores or centromeres. Elongation of the extranuclear spindle leads to the separation and migration of the daughter mastigont systems. During telophase, two daughter nuclei are formed. Once the daughter cells have separated, each produces the necessary organelles to replace those lost in the division process (Brugerolle, 1975; Honigberg and Brugerolle, 1990).

Nutritional Requirements and Cultivation

T. vaginalis is considered to be an aerotolerant anaerobe (Paget and Lloyd, 1990) that has become optimally adapted to survive in the vaginal milieu where high levels of CO_2 and traces of O_2 are the norm. *T. vaginalis* is also an obligate parasite that must glean many necessary nutrients from its surroundings; it lacks the ability to synthesize many of these macromolecules, acquiring them instead from vaginal secretions (Huggins and Preti, 1981) or via the ingestion of bacterial and host cells (Francioli *et al.*, 1983; Street *et al.*, 1984). Given the paucity of *de novo* synthesis of such essential nutrients as carbohydrates (ter Kuile, 1994; ter Kuile and Muller, 1995), amino acids (Rowe and Lowe, 1986; Zuo *et al.*, 1995), fatty acids (Roitman *et al.*, 1978), vitamins (Hollander and Leggett, 1985), iron (Gorrell, 1985), pyrimidines and purines (Heyworth *et al.*, 1982; Wang and Cheng, 1984; Harris *et al.*, 1988; Wang, 1990), appropriate *in vivo* culture media must therefore incorporate all these factors.

In keeping with this, Diamond (1986) outlined the essential components of complex liquid media for the axenic cultivation of trichomonads: (1) peptones (combinations of peptides and amino acids), (2) liver or yeast extracts which provide a source of B vitamins, purines and pyrimidines, (3) maltose or glucose which act as the main source of fermentable carbohydrates, (4) buffer, (5) reducing compounds, (6) agar to aid in the maintenance of low redox potential, (7) serum providing essential lipids and fatty acids, and (8) trace metals. Addition of an iron supplement, in particular, favours trichomonad growth. In axenic culture, *T. vaginalis* will grow in a relatively wide range of pH, but tends to flourish at a pH of 6.0 to 6.3 (Diamond, 1986).

Strain Heterogeneity

Much research has been conducted in the area of antigenic differences among isolates of *T. vaginalis*. Given that many of these studies focused on the development of serologic tests for trichomoniasis, antigenically defined populations of *T. vaginalis* have come to be referred to as serotypes (Ackers, 1990). It has been estimated that there exist between two and eight different serotypes (Honigberg, 1978; Garber *et al.*, 1986). These distinct "antigenic types" were demonstrated by precipitation, agglutination, complement fixation, and hemagglutination tests. Honigberg (1978) proposed that, for the most part, the number of antigenic types discovered tends to correlate with the number of strains examined.

A number of studies (Alderete *et al.*, 1985, 1986a, 1986b; Garber *et al.*, 1986) have indicated that the antigenic heterogeneity observed in *T. vaginalis* is a product of the surface location of a group of high molecular weight proteins, a number of which have been identified and analyzed. Interestingly, they also found that all isolates possess the capability to synthesize the same complement of highly immunogenic proteins; it is their deposition onto the cell surface that separates the strains.

T. vaginalis Pathogenesis

Even though *T. vaginalis* has become one of the most intensely-studied members of the trichomonad family and gives rise to a common STD, the pathogenic mechanisms whereby the protozoan invades the human urogenital tract, establishes residency there and induces pathologic effects are largely undefined. Given the wide gamut of clinical symptoms seen in trichomoniasis sufferers, it is unlikely that any single mechanism could be responsible. More likely, there are many mechanisms in the pathogenic arsenal of the parasite. Research to date suggests that these mechanisms include (1) the ability to evade host immune responses, (2) cellular adhesion, (3) haemolysis, (4) the excretion of soluble factors like extracellular proteinases and cell detaching factor (CDF), and (5) interaction with the normal urogenital flora.

Immune Evasion

T. vaginalis has developed specialized mechanisms that allow it to survive and flourish despite the best efforts of the host immune system. The ability of *T. vaginalis* to evade destruction by the host immune system and to survive in such a hostile and variable environment is an important element of its pathogenesis.

One way in which *T. vaginalis* manages to evade the host immune response is via the degradation of host antibodies. It has been documented that all trichomonal isolates synthesize numerous cysteine proteinases of varying molecular weights (Coombs and North, 1983; Lockwood *et al.*, 1987; Neal and Alderete, 1990; Alderete *et al.*, 1991). Provenzano and Alderete (1995) demonstrated that cysteine proteinases produced by *T. vaginalis* possess the ability to degrade human IgG, IgM, and IgA. They were able to detect immunoglobulin-degrading proteinases in the vaginal washes of patients with trichomoniasis, and went on to confirm proteolytic activity against immunoglobulin in the vaginal washes, suggesting *in vivo* importance. The production of

antibody-degrading proteinases by *T. vaginalis* appears to be constitutive and is independent of vaginal iron status. This may be particularly advantageous since it allows for parasite resistance to antibody regardless of the host immune status at the time of infection (Provenzano and Alderete, 1995).

Another important means of immune evasion is in the ability of *T. vaginalis* to coat itself with host plasma proteins, rendering the protozoan virtually invisible to the host immune response (Peterson and Alderete, 1982). Various assays and experiments have indicated that these proteins are bound, some loosely, others avidly, to the trichomonad surface via specific membrane sites (Peterson and Alderete, 1982). Specific and non-specific immune responses could be effectively blocked if the parasite is camouflaged in host plasma proteins.

Alderete and Garza (1984) have suggested that *T. vaginalis* may also elude the host immune response through the secretion of highly immunogenic antigens. They reasoned that the continuous release of these membrane protein antigens might neutralize the host's anti-*T. vaginalis* immune response by blocking antibody or cytotoxic lymphocyte receptors and hence interfering with specific antibody or cell-mediated killing.

Rendon-Maldonado *et al.* (1998) used transmission and scanning electron microscopy to show that *T. vaginalis* has the ability to internalize and degrade host leukocytes. Researchers found that both strains studied (one highly virulent, the other moderately virulent) were capable of internalizing the host cells, with the process occurring more rapidly in the more virulent strain. This may provide *T. vaginalis* with both an efficient means of nutrient acquisition as well as a way of counter-attacking the host immune response.

Another method of trichomonal immune evasion lies in the organism's capacity for complement avoidance. It has been well documented that *T. vaginalis* activates the alternative complement pathway (Gillin and Sher, 1981). Research has shown, however, that fresh isolates differ with respect to their sensitivity to complement-mediated killing (Demes *et al.*, 1988). Furthermore, it was demonstrated that during most of the menstrual cycle, *T. vaginalis* encounters very little complement activity. Demes *et al.* (1988) found an almost total lack of complement activity in cervical mucus samples studied. They also noted that while complement activity does increase with the onset of menses, menstrual blood possesses considerably less complement activity than does serum. Therefore, it appears that *T. vaginalis* colonizes a ecological niche where complement is not a real factor until the onset of menses, at which point, the trichomonad implements an iron-mediated system which protects it from complement-mediated killing. With increasing levels of iron in the menstrual blood, *T. vaginalis* upregulates synthesis of proteinases which are capable of degrading the C3 portion of host complement. This helps to prevent complement-mediated killing of the parasite and thus allows persistence of infection (Alderete *et al.*, 1995c).

These tactics further demonstrate how *T. vaginalis* is able to elegantly respond to changing environmental factors in its bid to survive in the vaginal environment. They also reinforce the genuine complexity of the parasite-host relationship.

Surface Adhesins and Adherence to Host Cells

The surface of a given *T. vaginalis* cell plays a vital role in its interaction with the surrounding environment. On the surface of the cell can be found a myriad of proteins and other molecules, some of which have been shown to play a part in the parasite's adherence to epithelial cells in the human vagina. It has been reported by a number of researchers (Heath, 1981; Alderete and

Pearlman, 1984; Krieger *et al.*, 1985; Graves and Gardner, 1993; Alderete *et al.*, 1995a) that cytoadherence of the trichomonad to host epithelial cells is essential for the effective colonization and persistence of the pathogen. Alderete and Garza (1985) described the parasite-epithelial cell association as time, temperature and pH-dependent in nature. They also found that *T. vaginalis* adhered with more avidity and possessed a higher degree of cytotoxicity to epithelial cell lines than they did to fibroblast cell lines. This finding is noteworthy because it parallels the situation seen *in vivo* where virulent trichomonads predominantly parasitize the vaginal epithelium.

According to Alderete *et al.* (1995a), cytoadherence, in its simplest form, involves the interaction of molecules on the surface of the microbe (adhesins) with specific molecules on the host cell surface (receptors). Trichomonal adherence is ligand-receptor in nature (Alderete and Garza, 1988) and depends upon four distinct surface proteins. These proteins have been designated with names reflecting their respective molecular masses: AP65, AP51, AP33, and AP23 (Arroyo *et al.*, 1992; Alderete *et al.*, 1995b; Engbring *et al.*, 1996). Anti-adhesin antibodies have been shown to specifically recognize and bind to their respective protein. Further, these antibodies inhibited the binding of live trichomonads to epithelial cells and protected the host cells from contact-dependent cytotoxicity (Arroyo *et al.*, 1992). The same study showed that pretreatment of epithelial cells with purified adhesins had a similar inhibitory effect on adherence.

Gene expression of the four adhesins has been shown to be regulated in a coordinated fashion at the transcriptional level. Lehker *et al.* (1991) and Ryu *et al.* (2001) have suggested that this regulation is, at least in part, under the control of environmental iron levels. Alderete *et al.* (1995b) estimate that such regulation may be a mechanism by which *T. vaginalis* adapts to the constantly changing environmental conditions found in the vaginal compartment. Downregulation of adhesins in response to a limited exogenous supply of iron may allow trichomonads to migrate more freely toward sites that are richer in iron. This would allow the

parasite to persist despite the flushing action of the mucosal secretions, desquamation of the mucosal epithelium, and a general shortage of nutrients (Alderete *et al.*, 1995a).

Surface adhesins are very sensitive to proteinases (Alderete and Garza, 1988; Arroyo *et al.*, 1992), yet strangely enough, cysteine-proteinase activity is a prerequisite for effective cytoadherence (Arroyo and Alderete, 1989; Arroyo and Alderete, 1995; Alvarez-Sanchez, 2000; Mendoza-Lopez *et al.*, 2000). The exact nature and functions of the surface proteinases are unknown, but it has been proposed that they may carry out some unmasking function. Alderete *et al.* (1995a) suggest that they contribute to adherence by degrading the proteins which protect the adhesins on the parasite surface.

While the exact details have not been completely elucidated, it appears that there exists a sophisticated signal transduction system in the *T. vaginalis* cell. The system initiates a number of changes in the parasite as it adheres to a host cell, the first being a transformation in the shape of the parasite. Arroyo *et al.* (1993) found that the typical ellipsoidal "pear-shape" (with a rough surface, four anterior flagella, a lateral undulating membrane and posterior axostyle) of the *in vitro*-grown *T. vaginalis* (Figure 1e) gave way rapidly, after contact with a vaginal epithelial cell, to a thin, flat ameboid shape (with a smooth surface) (Figure 1f). The ameboid shape seemed to allow the parasite to maximize the area of adhesion to the surface of the target cell. These "amoebic" parasites formed pseudopodia and filopodia that were able to interdigitate at specific sites on the plasma membrane of the vaginal epithelial cell. The trichomonads attached to the target cell in such an orientation that the side opposite the undulating membrane always faced the host cell, with the flagella and undulating membrane remaining free (Alderete and Garza, 1985). Once the morphological transformation had taken place, however, the axostyle was no longer apparent (Arroyo *et al.*, 1993). Rasmussen *et al.* (1986) demonstrated that once the morphological transformation occurred, the ameboid *T. vaginalis* were found to contain a dense

network of microfilaments in the area of the parasite in contact with the target epithelial cell. Krieger *et al.* (1985) supported the notion that the microfilaments concentrated at the site of contact played a role in the adherence to and parasitism of the target cell, in that inhibitors of microfilament and microtubule assembly dramatically reduced the cytopathic effects of trichomonads on Chinese Hamster Ovary (CHO) cell monolayers.

The second contact-dependent alteration was increased synthesis of the four surface adhesin proteins (AP65, AP51, AP33, and AP23), possibly strengthening attachment to the target cell, and thus allowing for more efficient parasitism of the target cell. Furthermore, the contact-initiated enhancement of surface adhesin synthesis may allow for the translocation of adhesins to the expanding filopodia and pseudopodia as they attach to newly-recognized contact sites on the target cell (Arroyo *et al.*, 1993). It was noted that the morphological transformation to an ameboid form occurred in response to contact with vaginal epithelial cells (the *in vivo* target cell), but not in response to contact with HeLa epithelial cells. The increase in adhesin synthesis, thought to result from a signal separate from the signal to change shape (Arroyo *et al.*, 1992), however, occurs in response to contact with either vaginal or HeLa epithelial cells (Arroyo *et al.*, 1993). The fact that increased adhesin synthesis occurs as a result of contact with HeLa epithelial cells may suggest that it is the interaction itself, between the adhesins and their receptors, which provides the signal for synthesis upregulation (Arroyo *et al.*, 1992, 1993).

Arroyo *et al.* (1993) also showed that while a given vaginal epithelial cell was originally parasitized by just one parasite, it was quickly covered by numerous trichomonads with multiple membrane interdigitations adjacent to one another. Alderete *et al.* (1995a), in reference to this study, have suggested that perhaps the original adherent trichomonad recruits other parasites to the target cell via the generation of a chemoattractant type of signal.

Other surface molecules have been implicated in the adherence mechanisms of *T. vaginalis*. Silva Filho *et al.* (1988) report that *T. vaginalis* displays, on its surface, laminin-binding proteins which may aid in attachment to target cells. These receptors are thought to recognize the glycoprotein laminin, which is located in the basement membranes of the urogenital tract and is believed to promote cellular adhesion, differentiation, shape, and motility (Silva Filho *et al.*, 1988). Crouch and Alderete (1999) have suggested that *T. vaginalis* may persist at sites in the vaginal compartment beneath the epithelial surface via interaction with laminin and fibronectin. They hypothesize that this may explain the non-self-limiting nature of trichomoniasis in the face of exfoliating vaginal epithelial cells from the vaginal epithelium. They have noted that receptors on the trichomonad surface interact with both laminin and fibronectin, and propose that these are independent of the four adhesin proteins.

In vitro pathological studies have shown that *T. vaginalis* possesses the ability to destroy monolayers and does so via an ordered process of events (Hogue, 1943; Christian *et al.*, 1963; Farris and Honigberg, 1970; Heath, 1981; Alderete and Pearlman, 1984; Alderete and Garza, 1985; Silva Filho and de Souza, 1988). Essentially, when first introduced into a co-culture with monolayer cells, trichomonads swim freely in the growth media, and eventually settle onto the target cells. Once *T. vaginalis* establishes direct contact with the target/host cells, the pathological effects can be observed. Small holes or "plaques" in the monolayer can be seen as the monolayer cells detach from areas where the trichomonads have adhered (the timing of this process varies depending on the target cell type, trichomonad virulence and concentration, and other culture conditions). As the parasites infiltrate more of the monolayer, these plaques grow in size until eventually the entire monolayer detaches. These plaques may have the same etiology as the hemorrhagic lesions seen in the cervix and vaginal mucosa in severe trichomoniasis (Christian *et al.*, 1963).

A number of researchers have assessed the viability of detached monolayer cells following contact with *T. vaginalis*, and have shown that irreversible damage and death of the target cell is the norm (Hogue, 1943; Heath, 1981; Alderete and Pearlman, 1984; Krieger *et al.*, 1985; Rasmussen *et al.*, 1986). For example, Alderete and Pearlman (1984) demonstrated that exposure of monolayer cultures of human urogenital and vaginal, human epithelial, normal baboon testicular, and monkey kidney cells to *T. vaginalis* resulted in disruption of monolayers. Via Trypan blue exclusion and ^{3}H-thymidine release they also showed that irreversible damage to the cells comprising these monolayers took place. Mirhaghani and Warton (1996) showed, with electron microscopic analysis, that direct contact between trichomonads and amnion membrane epithelial cells resulted in the destruction of the target cells, and eventually led to the desquamation of the epithelial layer from the basement membrane.

Haemolysis

It has been well documented that *T. vaginalis* possesses the ability to bind and lyse human erythrocytes (Krieger *et al.*, 1983; Dailey *et al.*, 1990; Lehker *et al.*, 1991; Fiori *et al.*, 1993). Furthermore, it has been shown that this β-haemolytic activity is correlated with virulence in patients, animal models, and tissue culture (Krieger *et al.*, 1983). The clinical observation that trichomoniasis is frequently exacerbated during menses may be due to increased substrate and therefore enhanced levels of haemolytic activity (Krieger *et al.*, 1983).

Given the fact that *T. vaginalis* does not possess the ability to synthesize, *de novo*, a number of molecules necessary for survival, the parasite must obtain these nutrients from exogenous sources. Erythrocytes and other host cells may therefore become an attractive source of nutrients for the trichomonad. *T. vaginalis* lacks the necessary enzymatic machinery for the biosynthesis of lipids (Peterson and Alderete, 1984; Beach *et al.*, 1990, 1991), so the β-haemolytic activity of

trichomonads may be a very important means of nutrient acquisition. Haemolysis of erythrocytes also provides *T. vaginalis* with a supply of iron, which is integral to its growth and multiplication (Lehker *et al.*, 1990; Lehker and Alderete, 1992; Fiori *et al.*, 1993) and is a limiting factor in the vaginal environment.

Haemolysis is a contact-dependent process. Fiori *et al.* (1993) demonstrated that haemolysis occurs when trichomonads are co-cultured with target cells, but not when they are separated by a semi-permeable membrane. The process is also sensitive to temperature and pH. Dailey *et al.* (1990) showed that haemolysis is optimal at 37°C and at a level of acidity reflecting the pH range (5.0-6.0) found in the vagina during trichomoniasis.

Erythrocyte lysis is mediated by protein receptors found on the surface of both the red blood cells and the trichomonads. Fiori *et al.* (1993) identified five trichomonad surface adhesins that facilitate the binding of the *T. vaginalis* cell to the host erythrocyte. Molecular mass determination has shown that three of the five polypeptides are identical to the adhesins found by Arroyo *et al.* (1992) (AP65, AP51, and AP33). The two other surface proteins defined by Fiori *et al.* (1993) possessed molecular masses of 140 and 42 kDa. On the basis of this work, Fiori *et al.* (1993) have divided the haemolytic process into at least three distinct steps. Firstly, the trichomonad recognizes and adheres to the target erythrocyte via the adhesins described previously. Secondly, the parasite releases toxic molecules that are believed to form functional pores in the target membrane. It has been suggested that these toxic molecules are soluble perforin-like proteins (Fiori *et al.*, 1993). While in contact with the target erythrocyte, *T. vaginalis* has also been shown to cause disruption of the target cell cytoskeleton via the degradation of the molecule spectrin. It is believed that when the parasite is in close intimate contact with the erythrocyte, it transfers a 30 kDa proteinase effector into the target cell, which then suffers cytoskeletal disruption (Fiori *et al.*, 1997). (It has been suggested that the same type

of proteinase could be used by *T. vaginalis* to induce cytoskeletal disruption in nucleated target cells as well. This would leave the host cell membrane more vulnerable to lysis by perforin proteins released by the trichomonad (Fiori *et al.*, 1999)). Finally, the *T. vaginalis* cell detaches from the erythrocyte prior to target cell lysis (Fiori *et al.*, 1993). Apolipoprotein CIII-specific receptors on the trichomonad surface enable it to bind and incorporate lipids from the lysed erythrocyte (Peterson and Alderete, 1984). As well, *T. vaginalis* has specific surface receptors that allow it to bind and internalize iron-containing molecules like haemoglobin (Lehker *et al.*, 1990).

Proteinases

T. vaginalis has the ability to produce a variety of different proteinases (Coombs and North, 1983; Lockwood *et al.*, 1987). It has been shown that *T. vaginalis* lysates contain proteinases and that the parasite secretes, (both *in vitro* and *in vivo*), a variety of proteinases (Alderete *et al.*, 1991). Neale and Alderete (1990) identified twenty-three distinct proteinase activities in their evaluation of fresh and cultured *T. vaginalis* isolates. They demonstrated that some proteinases were differentially expressed and underwent phase variation, suggesting that they were regulated by environmental factors.

It seems reasonable that some of the proteinases produced by *T. vaginalis* play a role in the pathogenesis of the organism. In order for cysteine proteinases to be functional, they must first be activated by reducing agents. Alderete and Provenzano (1997) have shown that the human vagina possesses a reducing environment sufficient for the activation of *T. vaginalis* proteinases, supporting the hypothesis that they play a role *in vivo*.

It has been established that while the pH of vaginal secretions in healthy individuals is less than or equal to 4.5 (Paavonen, 1983; Larsen, 1993), a rise in pH to greater than 5.0 correlates with *T. vaginalis* infection (Hanna *et al.*, 1985; Rein, 1990). The proteinases characterized by Garber and Lemchuk-Favel (1989) were found to be active over a wide pH range (4.0-8.0), suggesting a potential role for proteinases in the establishment of infection. As has been discussed in the preceding sections, cysteine proteinases appear to be involved in different aspects of *T. vaginalis* pathogenesis, including adherence to host cells (Arroyo and Alderete, 1989; Arroyo and Alderete, 1995; Alvarez-Sanchez, 2000; Mendoza-Lopez *et al.*, 2000), cytoskeletal disruption in erythrocytes (Fiori *et al.*, 1997) and in nucleated target cells (Fiori *et al.*, 1999), and immune evasion via immunoglobulin degradation (Provenzano and Alderete, 1995). Lehker and Sweeney (1999) have demonstrated another interesting pathogenic role played by *T. vaginalis* cysteine proteinases. They showed that secreted trichomonad proteinases were able to degrade mucin, the framework molecule of mucus, which helps to form a physical barrier to microbial colonization in the human urogenital tract.

Cell Detaching Factor (CDF)

Hogue (1943) first proposed that *T. vaginalis* may produce soluble cytotoxic mediators *in vitro* via a contact-independent mechanism when she reported that cell-free filtrates caused severe pathology in tissue culture. The issue has been debated ever since, but evidence is accumulating in support of this hypothesis (Farris and Honigberg, 1970; Pindak *et al.*, 1986; Silva Filho and de Souza, 1988; Garber *et al.*, 1989; Garber and Lemchuk-Favel, 1990; Pindak *et al.*, 1993). Heath (1981) suggested that *T. vaginalis* likely employs both contact-mediated and contact-independent pathogenic mechanisms. In the *in vivo* setting, the importance of contact-mediated pathogenesis is readily apparent. Adhesion of trichomonads to the vaginal epithelium is important in the disruption of the superficial layers of the squamous epithelium. Heath (1981) observed, however,

that many infected women display a notable increase in the sub-epithelial vascularity of the cervicovaginal walls which is frequently out of proportion to the numbers and location of trichomonads adhering to the epithelium. This idea is supported by the findings of Nielson and Nielson (1975) who employed electron microscopy to look at biopsied vaginal epithelial cells from patients with trichomoniasis. These findings may suggest that *T. vaginalis* is able to exert pathogenic effects on host cells without actually coming into contact with them.

A number of researchers have examined contact-independent *T. vaginalis* cytotoxicity and have verified the original findings. Many of these studies give direct evidence showing cytopathic effects in culture cells in response to incubation with cell-free filtrates of *T. vaginalis* (Farris and Honigberg, 1970; Pindak *et al.*, 1986; Silva Filho and de Souza, 1988; Garber *et al.*, 1989; Garber and Lemchuk-Favel, 1990; Pindak *et al.*, 1993). Others have used a semi-permeable filter system which prevents contact between the trichomonads and the culture cells, but permits passage of media. Garber and Bowie (1990) used this system and concluded that trichomonad pathology still occurs in the absence of contact.

Pindak *et al.* (1986) first designated the trypsin-like component of cell free filtrate, which caused rounding and detachment of a number of different types of monolayer cells, as cell detaching factor (CDF). CDF was also shown to prevent attachment of fresh culture cells to the culture plate. It has been demonstrated that although the monolayer cells become detached, they remain viable for an extended period of time, and are even capable of re-forming a monolayer if washed and transferred to fresh media (Pindak *et al.*, 1986; Garber *et al.*, 1989).

The role of CDF in pathogenesis is unknown, but the *in vitro* detaching of monolayer cells may mirror the *in vivo* observation that patients with trichomoniasis often display sloughing of the vaginal epithelial cells (Garber *et al.*, 1989). Garber and Lemchuk-Favel (1990) have concluded

that CDF activity correlates strongly with clinical disease and thus may be an important marker of virulence in *T. vaginalis*. Cell-free filtrates of *Pentatrichomonas hominis*, a non-pathogenic gastrointestinal trichomonad, do not display any CDF activity, also suggesting that this component is relevant to an isolate's virulence mechanism (Garber *et al.*, 1989).

CDF has been identified as a metabolic product of *T. vaginalis*; lysate of the parasite is inactive, suggesting that the factor is not an internal component (Pindak *et al.*, 1986). It is a heat and acid labile 200 kDa glycoprotein (Pindak *et al.*, 1986; Garber *et al.*, 1989). Garber *et al.* (1989) found that the amount of CDF in growth filtrate varied with duration of trichomonad growth, the initial inoculum of *T. vaginalis*, and the pH of the filtrate at the time of harvesting. Research has indicated that, due to its acid lability, CDF activity is influenced by the pH of the surrounding media (Pindak *et al.*, 1986; Garber *et al.*, 1989). Garber *et al.* (1989) have found that CDF is active from pH 5.0 to pH 8.5, with peak activity at 6.5. Despite the fact that low pH (<5.0) causes the inactivation of CDF, the compound may still play a role in the *in vivo* setting. While the normal pH of the healthy vaginal compartment stays at or below 4.5, the pH rises to greater than 5.0 with *T. vaginalis* infection. Thus, once an infection has been established, the altered vaginal pH becomes permissive for the functioning of CDF (Garber *et al.*, 1989). CDF activity may also be affected by vaginal estrogen concentration. Garber *et al.* (1991) found that while estrogen has no real effect on *T. vaginalis* growth or on the activity of CDF, it does inhibit the production of CDF by the parasite. The diminished levels of estrogen found in the vaginal environment around the time of menses may allow for augmented production of CDF, leading to the exacerbation of the disease (Garber *et al.*, 1989).

These reports suggest that contact-independent factors likely play a part in the *T. vaginalis* pathogenic repertoire. A number of researchers question the importance of contact-independent mechanisms, however, and have reported a lack of such pathology in their experiments (Kotcher

and Hoogasian, 1957; Christian *et al.*, 1963; Kulda, 1967; Alderete and Pearlman, 1984; Krieger *et al.*, 1985; Rasmussen *et al.*, 1986). Given that CDF displays sensitivity to acid concentration, some (Pindak *et al.*, 1986; Garber *et al.*, 1989; Garber and Bowie, 1990; Pindak *et al.*, 1993) have suggested that perhaps *T. vaginalis* metabolic end-products interfered with these experiments by lowering the pH to levels (<5.0) that rendered CDF inactive. Garber *et al.* (1989) showed that monolayer cells die rapidly when exposed to pH levels below 5.0 which is in contrast to the effects induced by CDF (i.e. rounding and detachment, but not cell death). Garber and Bowie (1990) acknowledged that contact-dependent mechanisms play an important role in trichomonad pathogenesis, but suggested that some of the contact-dependent cytotoxicity seen in previous research could be attributed to low pH levels. These same low pH levels would prevent contact-independent cytotoxicity by inactivating CDF. Support for this idea came when, using a filter system to keep the cells separate, Garber and Bowie (1990) rigidly controlled the pH in a co-culture of *T. vaginalis* and McCoy cells. They found that McCoy cells detached but remained viable. When the cells were cultured in contact with one another, monolayer cells did die eventually, but pH adjustment led to a decreased rate of target cell death (Garber and Bowie, 1990).

Interaction with Vaginal Flora

The vaginal flora of a healthy woman consists of a wide variety of anaerobic and aerobic bacterial genera and species, but the most dominant is the facultative, microaerophilic, anaerobic genus Lactobacillus (Redondo-Lopez *et al.*, 1990). It has been shown that lactobacilli possess many properties and produce a number of metabolites that are important in the maintenance of the normal vaginal environment (Redondo-Lopez *et al.*, 1990). Another hallmark of a healthy human vaginal environment is an acidic pH (4.5 or less). However, once the cervicovaginal region is colonized with *T. vaginalis*, there is almost always a rise in the pH to levels where the parasite

can thrive and multiply more easily (pH of 5.0 or greater) (Hanna *et al.*, 1985; Rein, 1990). Another pathological change associated with *T. vaginalis* infection is the reduction in numbers or even total loss of resident lactobacilli (Jirovec and Petru, 1968). How *T. vaginalis* eliminates the resident lactobacilli is not well understood, but *in vitro* studies have demonstrated that *T. vaginalis* does have a deleterious effect on *L. acidophilus* in co-cultures (McGrory *et al.*, 1994). While the exact mechanisms have not been elucidated, *T. vaginalis* has been shown to phagocytize lactobacilli *in vitro* (Juliano *et al.*, 1991; Rendon-Maldonado *et al.*, 1998). Similarly, trichomonad-secreted metabolites (such as CDF or proteinases) have been shown to inhibit lactobacilli growth *in vitro* (McGrory and Garber, 1992; McGrory *et al.*, 1994). Either or both of these mechanisms may occur *in vivo*. They both lead to the removal of a factor that helps to maintain a pH that is unhealthy for the colonizing trichomonads, and phagocytosis may provide a source of nutrition for the parasite.

Apoptosis-A Potential Pathogenic Mechanism Employed by *T. vaginalis*?

The apoptotic process is known to serve many vital functions, such as cellular deletion during embryonic development, balancing mitosis in continuously regenerating tissues, hormone-dependent involution in the adult, immune system maturation, selective deletion of inappropriate immune cells, and a number of other physiological processes (Allen *et al.*, 1997). Dysregulation of apoptosis can be dangerous, however, and has been linked to a wide variety of diseases and conditions. A lack of apoptosis may constitute one possible route to carcinogenesis or may be involved in the development of some autoimmune disorders. Conversely, elevated levels of apoptosis have been linked to neurodegenerative conditions like Alzheimer's and to the T cell destruction characteristic of AIDS (Wyllie, 1997).

A number of microorganisms can induce apoptosis dysregulation in host cells. In fact, it has been documented to be an important aspect of pathogenesis for many viruses, bacteria, and protozoa. Some viruses have been shown to induce apoptosis in host cells. Examples include coxsackievirus B3, an enterovirus in the family *Picornaviridae* (Carthy *et al.*, 1998), and HIV (Badley *et al.*, 1998). Many other viruses express anti-apoptotic proteins that effectively slow or block the apoptotic cascade (Young *et al.*, 1997) possibly leading to enhanced viral survival. Such viral interference with the apoptotic process may promote cancer (Young *et al.*, 1997; Tschopp *et al.*, 1998). Tschopp *et al.* (1998) have outlined a variety of viruses, including herpesviruses, poxviruses, papovaviruses, adenoviruses, and baculoviruses, that encode a number of immunomodulatory molecules that are capable of interfering with the host's apoptotic signaling pathways.

Researchers have described a number of bacteria that survive, in part, via the ability to induce apoptosis in host immune cells. Some do so by producing pore-forming proteins that induce subtle biochemical and physiological changes in the cell, which lead to apoptosis. Bacterial pathogens that produce such toxins include *Actinobacillus actinomycetemcomitans*, *Staphylococcus aureus*, and *Escherichia coli* (Chen and Zychlinsky, 1994). Other bacteria, such as *Corynebacterium diphtheriae*, *Pseudomonas aeruginosa*, and *Shigella dysenteriae*, induce apoptosis in host immune cells by producing toxins that enter into the host cells and inhibit protein synthesis (Chen and Zychlinsky, 1994). Monack *et al.* (1996) have shown that *Salmonella typhiumurium* induces apoptosis in host macrophages by causing "membrane ruffling" upon entry into the cell. Hersh *et al.* (1999) have shown that *Salmonella* spp. induce apoptosis in host macrophages by binding directly to caspase-1. Group A *Streptococcus pyogenes* has been found to induce apoptosis in host epithelial cells by initiating mitochondrial dysfunction (Nakagawa *et al.*, 2001).

Apoptotic dysregulation has also been documented in cells exposed to various protozoan pathogens. Toure-Balde *et al.* (1995, 1996, 2000) have shown that *Plasmodium falciparum*, a causative agent of malaria, induces apoptosis in human mononuclear cells. Researchers have found that *Trypanosoma cruzi*, (which causes Chagas' disease) evades the host immune response by inducing apoptosis in T cells (DosReis *et al.*, 1995; Lopes *et al.*, 1995) and in B cells (Zuniga *et al.*, 2000). Das *et al.* (1999) found that experimental infection with the intracellular protozoan parasite *Leishmania donovani* leads to selective deletion of host CD4+ T cells, which results in a decrease in the secretion of interleukin-2 and interferon-gamma. Conversely, Moore and Matlashewski (1994) found that *Leishmania donovani* inhibits macrophage apoptosis, thus facilitating the spread of infection by boosting the number of host cells available for parasitization, and by augmenting the number of infected macrophages available for uptake by the sandfly vector. They believe that the parasites do this by inducing host macrophages to secrete a soluble factor that may act in an autocrine manner to prevent apoptotic cell death (Moore and Matlashewski, 1994).

These examples help to illustrate the importance of apoptosis in the pathogenic complement of numerous different microorganisms, including a number of protozoan pathogens. It was proposed, therefore, that *T. vaginalis* might possess the ability to induce apoptosis in its pathogenic attack on monolayer cells in an *in vitro* co-culture system. This relationship has not been investigated before, but seems plausible given the wide scope of the elements incorporated in the pathogenic arsenal of *T. vaginalis*.

Defining Apoptosis

Apoptosis generally takes place in scattered single cells and is considered to be a form of self-directed cellular suicide. It is characterized by a specific set of biochemical and morphological

changes in the cell. The term was first used by Kerr *et al.* in 1972 when they noted that morphologically similar cell deaths could be detected in many pathological conditions as well as in normal tissue.

Morphologically, the apoptotic cell shrinks and becomes denser. The chromatin becomes pyknotic and is packed into smooth masses against the nuclear membrane. This phenomenon leads to margination of the nuclear chromatin such that it takes on a distinctive half-moon, or sickle-shape. During the condensation phase, there is a breakdown of cell-cell interactions as the cell undergoing apoptosis is isolated from its neighbours (Bursch *et al.*, 1990). There is little or no swelling of the cellular organelles. The nucleus may break into small pieces (karyorhexis) and the cell often emits extensions which tend to break off and become apoptotic bodies. These apoptotic bodies, formed via the "budding phenomenon", often contain pyknotic nuclear fragments (Majno and Joris, 1995). Interestingly, there appears to be no leakage of cellular contents during the budding process (Bursch *et al.*, 1990). Once they have separated from the main apoptotic body, the smaller buds are free to be phagocytized by macrophages or by neighbouring cells. If budding does not occur, the cell may shrink into a single compact, spherical mass (Majno and Joris, 1995).

Biochemically, DNA is fragmented by endogenous endonucleases via a two-stage process. Initially, the DNA is chopped up into large fragments (50-300 kb in size) and then ultimately a portion of the DNA is further fragmented into oligo- and mononucleosomal-size fragments (Arends *et al.*, 1990; Kokileva, 1994; Walker *et al.*, 1994; Zhivotovsky *et al.*, 1994). Walker and Sikorska (1997) have suggested that depending on the cell type, DNA fragmentation is likely controlled by either Ca^{2-} ions or a decrease in pH. They also stated that while DNA degradation into oligonucleosomes is not essential for apoptosis, all cells must undergo degradation of DNA into large fragments. Another biochemical hallmark of apoptosis is the implementation of

changes in the plasma membrane. One such change, namely the translocation of phosphatidyl serine from the inner to the outer leaflet of the plasma membrane, has been well documented (Allen *et al.*, 1997). Researchers have suggested that this, as well as a number of other surface markers, likely play an important role in the recognition of the apoptotic cell by professional phagocytes.

The process of apoptosis is under genetic control and can be instigated by an internal genetic clock (as in programmed cell death), or conversely, can be induced by extracellular agents such as hormones, cytokines, killer cells, microorganisms, or a variety of different chemical or physical agents (Majno and Joris, 1995). Apoptosis can occur very rapidly, and for this reason tends to be relatively unobtrusive in tissue sections. It is generally suggested that apoptosis does not induce an inflammatory response, but this may not be completely accurate. It has been well documented that the apoptotic bodies resulting from cellular suicide are often recognized and phagocytized by macrophages (Majno and Joris, 1995). This would indicate that the dying cells must release some form of chemoattractant, allowing professional phagocytes to find their target. It could be that since apoptotic cells often die singly, and not *en masse*, the level of chemoattractant emitted attracts only a few nearby macrophages and does not induce a large-scale inflammatory response (Majno and Joris, 1995).

Often, the term programmed cell death is used as a synonym for apoptosis, but this is misleading. There are many situations where a programmed cell death occurs by a mechanism other than apoptosis. Programmed cell death merely indicates that a genetic clock selects a given time for the death of certain cells, while the genetic program for apoptosis specifies the means by which the suicide will occur (Wick, 1994; Hockenberry, 1995; Majno and Joris, 1995).

The Caspase Family of Cysteine Proteinases

The central effectors of the apoptotic process are the cysteinyl aspartate-specific proteinases (caspases). The caspase gene family has thus far been shown to include at least fourteen mammalian members, of which eleven human enzymes have been identified (Nicholson, 1999). Caspases generally function in an ordered enzymatic cascade. In the typical caspase cascade, the upstream caspases (initiator caspases including caspases-6, -8, -9, and -10) are activated by association with one or more caspase adapters. This initial activation step has an important regulatory role and is controlled by a complex network of pro- and antiapoptotic proteins. Once the initiator caspase is transformed to its active state, it processes and activates one or more downstream caspases (executioner or effector caspases-usually caspases-2, -3, and -7). The active effector caspases then cleave various cellular substrates, which leads to the ultimate apoptotic demise of the cell (Chang and Wang, 2000).

Caspases are first synthesized as inactive precursors, or procaspases. All procaspases contain a highly conserved protease domain. The protease domain is comprised of a large subunit (17-21 kDa) and a small enzyme subunit (10-14 kDa). The procaspases also contain a prodomain or NH_2-terminal peptide of varying lengths; the prodomain can be longer than one hundred amino acids in the initiator caspases or may contain fewer than thirty amino acids in the effector caspases. Some procaspases contain a brief linker between the small and large subunits. A two-step proteolytic processing mechanism gives rise to a mature caspase tetramer (a homodimer of the large and small heterodimers arranged in twofold rotational symmetry) (Rathmell and Thompson, 1999; Chang and Yang, 2000).

One of the most remarkable enzymatic properties of the caspases is their specificity with respect to substrate binding. Caspases recognize a short tetrapeptide sequence within targeted

polypeptides and have demonstrated an absolute requirement for an aspartic acid residue in the P_1 position (Thornberry, 1997; Nicholson, 1999). The variability in residue preference for the P_2-P_4 (P_4 in particular) sites has helped researchers to group the caspases in the following manner: the group I caspases (including caspases-1, -4 and -5) prefer substrates with bulky hydrophobic residues at P_4; the group II enzymes (including caspases -2, -3, and -7) prefer aspartic acid at the P_4 site; group III caspases (including caspases-6, -8, and -9) bind preferentially to branched-chain aliphatic amino acids at the P_4 site. While their substrate preferences have not been as well defined, phylogenetic evidence suggests that caspases-11, -12, and -13 resemble the group I caspases, and caspase -10 is similar to the group III caspases (Rathmell and Thompson, 1999).

There are at least three different mechanisms for caspase activation in the mammalian cell. The first, called recruitment activation, involves the interaction of a procaspase in an oligomeric activating complex via its prodomain. An example of this kind of activation is the recruitment of procaspase-8 to an oligomeric activation complex following ligation of the surface CD95 (Fas) receptor. The second mechanism, transactivation, occurs when one caspase is activated by another. An example would be the cleavage and activation of a downstream effector caspase by an initiator caspase that has recently undergone activation by recruitment. The third mechanism, where a caspase initiates its own activating cleavage, is termed autoactivation (Nicholson, 1999; Wolf and Green, 1999; Kruidering and Evan, 2000).

In mammalian cells, there are two fundamental ways by which a caspase cascade is known to be initiated. The first is called the intrinsic pathway and is generally thought to occur in cells that have been damaged or neglected. Essentially, cytochrome *c*, under the control of the Bcl-2 family proteins, is released from the mitochondria; cytochrome *c* then binds and activates apoptotic protease-activating factor (Apaf-1). Apaf-1 binds and activates caspase-9, which, in turn, activates caspase-3, initiating the downstream caspase cascade (Porter and Janicke, 1999;

Rathmell and Thompson, 1999; Chang and Yang, 2000). The second mechanism is the extrinsic pathway, which can occur following the binding of surface death receptors. The most thoroughly described surface receptors belong to the tumor necrosis factor receptor (TNFR) family. The TNFR family includes apoptosis-inhibiting and apoptosis-promoting receptors. A well-characterized apoptosis-promoting receptor, CD95 (Fas), can directly recruit and activate caspases-8 and -10 via its intracellular death domain. Once initiated, the upstream caspases can then proceed to bind and activate the downstream effector caspases (Rathmell and Thompson, 1999; Chang and Yang, 2000).

The substrates associated with the downstream effector caspases are varied and numerous; researchers have identified more than sixty mammalian caspase substrates. The ultimate consequences of substrate cleavage include the impairment of homeostatic and repair processes, the inactivation of apoptosis inhibitors, the disassembly of structural integrity, induction of morphological alterations, and the labeling of the dying cell for phagocytosis. To do this, the caspases may inactivate the normal function of a substrate, activate the substrate via influence on regulatory domains, alter the function of the target protein, or become directly involved in the proteolysis of the structural components of the cell (Nicholson, 1999).

Regulation of the complex apoptosis cascade involves both apoptosis-promoting and apoptosis-inhibiting proteins. Apoptosis inhibitors may be produced endogenously by the cell. The inhibitor of apoptosis (IAP) molecules, which carry out their function either by binding directly to the caspases or by influencing signal transduction pathways, are important examples of endogenous caspase inhibitors. Viruses have also been shown to produce caspase inhibitors that prevent host cell death, thus allowing for viral survival and propagation (Chang and Yang, 2000). In the laboratory, several classes of reversible and irreversible peptide-based caspase inhibitors have been designed for use in various apoptosis models. Reversible inhibitors include aldehydes,

nitriles, and ketones. Irreversible inhibitors include chloromethylketones, fluoromethylketones, and diazomethylketones. The peptide portions of these compounds determine their caspase-binding preference (Thornberry, 1997).

Hypothesis and Objectives

Hypothesis

T. vaginalis causes a number of pathological changes in host cells and effectively destroys a variety of host cell monolayers *in vitro*. The pathology induced by *T. vaginalis in vitro* likely mirrors the *in vivo* situation, where host tissues display trichomonad-induced lesions and cell sloughing. It has been well established that *T. vaginalis* uses haemolysis of erythrocytes and phagocytosis of urogenital flora as an effective nutrient acquisition system. Many microorganisms, including a number of protozoan pathogens, possess the ability, as part of their pathogenic arsenal, to induce apoptosis in host cells. Our hypothesis is that *T. vaginalis* induces apoptosis in host cells during the course of infection. We speculate that it may be to the parasite's advantage to induce apoptosis in host tissues or target monolayer cells. The break down into apoptotic bodies of a target cell might make it a more suitable candidate for phagocytosis, thus facilitating the salvage of a variety of nutrients (carbohydrates, amino acids, fatty acids, vitamins, and iron) and nucleic acids which *T. vaginalis* cannot synthesize.

Objectives

Using an *in vitro* system (McCoy cell monolayer), the objective was to look at the possible induction of apoptosis by *T. vaginalis*. The fragmented DNA characteristic of apoptosis would be detected with the use of TUNEL and ELISA assays. If apoptosis was confirmed, the mechanism of induction would be investigated by examining the involvement of caspases, and by

determining if the process was contact-dependent or independent. The co-culture nature of the experiments would also allow us to examine the contact-dependent and independent interaction of *T. vaginalis* and McCoy cell monolayers with respect to pathology other than the induction of apoptosis.

MATERIALS AND METHODS

Cell Culture

Growth and Maintenance of McCoy Cells

The monolayer cells used in the co-culture experiments were mice fibroblast (McCoy) cells that had been used extensively in our lab over the past decade. McCoy cells were grown at 37°C / 5% CO_2 in Minimum Essential Medium (Gibco BRL-Life Technologies Inc. Grand Island, NY) adjusted to a pH of 7.5 and supplemented to provide the following concentrations: 2 mM L-glutamine (Gibco-BRL), 7 mM sodium bicarbonate (Fisher Scientific, Whitby, ON), 24 mM glucose (Fisher Scientific), 8 μg gentamicin (Gibco-BRL) per mL, 25 U nystatin per mL. Heat inactivated fetal bovine serum (FBS) (Gibco-BRL) was added at the end to provide 10% of the total volume of the media. This media will be referred to as CMGA (Garber *et al.*, 1987)

McCoy monolayers were regularly grown in 75 cm^3 and 25 cm^3 vent cap plastic tissue culture flasks with 0.2 μm membrane cap (Corning, Fisher Scientific). Every two to three days, media in the flasks was decanted, cells were trypsinized (0.25% Trypsin (Gibco-BRL)), resuspended in CMGA, spun at 500 x g for five minutes and then were transferred into fresh media.

For co-incubation experiments, McCoy cells were first cultured in flasks, then transferred to four-well Nunc Lab Tek II Chamber Slides (Gibco-BRL). These consisted of four sterile plastic chambers attached to a glass slide. Normally, $1x10^5$ McCoy cells, in 1 mL of CMGA, were placed in each chamber and were allowed to form a monolayer over forty-eight hours. Healthy cells were firmly attached to the glass slide after twenty-four hours, at which point, the media was replenished.

Growth and Maintenance of Jurkat Cells

Jurkat cells (clone E6-1 (T cell leukemia, human) ATCC TIB-152) were cultured in plastic tissue culture flasks in RPMI 1640 medium (Gibco-BRL) supplemented to yield the following concentrations: 2 mM L-glutamine, 1.5 g/L sodium bicarbonate, 4.5 g/L glucose, 10 mM HEPES, and 1.0 mM sodium pyruvate (Gibco-BRL). Heat inactivated FBS was added to make up 10% of the final culture media volume. Cells were enumerated and sub-cultured every two to three days in order to maintain the cell density between 10^5 and 10^6 cells per mL.

Culture of Microorganisms

Trichomonas vaginalis

The *T. vaginalis* isolates used in these experiments were originally harvested from the vaginal secretions of a female patient suffering from vaginitis. An isolate designated "202" was employed for these experiments. In a ranking of clinical isolates, "202" has been shown to cause moderate pathology in patients (Garber and Lemchuk-Favel, 1990). Organisms were grown in glass screw-capped tubes (16 x 125 mm) in 10 mL of Diamond's TYI-S-33 medium (TYI-See Appendix II) (pH 6.2) (Diamond *et al.*, 1978) supplemented with 10% heat inactivated FBS, 100 U/mL penicillin, 100 µg/ml streptomycin (penicillin/streptomycin solution, Gibco-BRL) and 2.5 µg/mL amphotericin B (Fungizone, Gibco BRL). Tubes containing *T. vaginalis* cultures were incubated at 37°C with 5% CO_2, at a 45° angle. Cultures were passaged every two to three days. Viable organisms were enumerated on a Bright Line hemacytometer (VWR Canlab, Mississauga, ON), using trypan blue exclusion (1% Trypan Blue (Sigma, Sat. Louis, MO)) to distinguish between dead and healthy cells (Takahashi *et al.*, 1970).

In vitro Co-Incubation Experiments

T. vaginalis and McCoy Cell Monolayers

As described above, McCoy monolayers were allowed to grow to confluency in sterile chamber slides. Prior to the addition of *T. vaginalis*, monolayers were washed three times with sterile prewarmed PBS. *T. vaginalis* grown to mid-log phase was prepared in the following manner: *T. vaginalis* (isolate 202) growth culture was transferred to a centrifuge tube and the cells were spun at 900 x g for ten minutes. The pellet was washed three times and the final pellet was then suspended in two parts CMGA to one part TYI. The resuspended sample was counted with a hemacytometer and the concentration was adjusted accordingly. Depending on the desired concentration for each chamber (10^4, 10^5, or 10^6 *T. vaginalis* per mL), proportions of the prepared *T. vaginalis* and of the CMGA-TYI media were added to each of the prewashed McCoy monolayers, so that each chamber held 1 mL. One chamber on each slide was designated the control and was incubated with media only. The chamber slides were then placed in a 37°C incubator (with 5% CO_2) and were incubated for six hours. After the six hour incubation period, the pH of the culture supernatant was checked (pHydrion Vivid 6-8 (Fisher Scientific)), and removed from the chambers for transfer to pre-labeled cryovials for storage at -80°C.

Once the media was transferred, the chambers were removed from the glass slides, and were fixed for ten minutes in 1% paraformaldehyde (PFA-BDH supplied by VWR Canlab), in preparation for immunocytochemical staining.

For the time-course experiment, a portion of the slides was removed from the incubator at two hours and four hours. These slides were fixed in PFA and were then transferred to PBS (at 4°C) for temporary storage until all slides could be recovered from the incubator.

T. vaginalis Growth Media (Cell Free Filtrate) and McCoy Cell Monolayers

T. vaginalis cells were passaged into CMGA-TYI (2:1) at a concentration of 10^4 trichomonads per mL and were incubated for forty-eight hours, allowing the culture to reach mid-log phase. The sample was centrifuged at 900 x g for ten minutes. The supernatant was poured through a bottle top filter (0.22 μm pore size (Corning-Fisher Scientific)) to remove any debris or cells still present after centrifugation. pH of the cell free filtrate was checked (pHydrion Vivid 6-8) and adjusted with 1N NaOH to a pH of 6.6 to 6.8.

McCoy cells were grown to confluency in chamber slides, as described above, and immediately before the addition of the cell free filtrate, were washed three times with sterile PBS. Cell free filtrate and fresh media were added in the appropriate proportions to each chamber, with a final volume of 1 mL in each. One chamber was again designated the control and to this chamber CMGA-TYI (2:1) media only was added. Monolayers were incubated at 37°C / 5% CO_2 for six hours.

Cytospins to Recover Detached Cells

Cells that had detached from the chamber slides during co-incubation were pelleted from the co-incubation media by centrifugation at 500 x g for ten minutes. The supernatant was then aspirated and the pellet was resuspended in 500 μL 1% PFA for ten minutes to fix cells. The samples were then spun (Shandon cytospin centrifuge (Fisher Scientific)) onto glass slides at 1200 rpm for ten minutes.

Induction of Apoptosis in McCoy Cells Using Camptothecin (Positive Control)

McCoy cell monolayers were first allowed to grow to confluency as previously described. Monolayers were treated with camptothecin (concentrations of 5 μM and 10 μM) in CMGA

media for four hours; fresh CMGA was added to monolayer cells, which were then incubated for an additional forty-eight hours. After treatment, all chambers were removed and the slides were fixed for ten minutes in 1% PFA, in preparation for immunocytochemical staining.

Caspase Inhibition in McCoy Cells

McCoy monolayers were pre-treated with caspase inhibitors z-Ile-Glu(OMe)-Thr-Asp(OMe)-fluoromethylketone (z-IETD-fmk) (Enzyme Systems Products, Livermore, CA) and z-Val-Ala-Asp(OMe)-fluoromethylketone (z-VAD-fmk) (Enzyme Systems Products) prior to co-incubation with *T. vaginalis*. McCoy monolayers were incubated with the inhibitors for twelve hours, and then with inhibitors and *T. vaginalis* for an additional six-hour period. Ten μM and 100 μM concentrations of z-VAD-fmk and z-IETD-fmk were used to treat different monolayers. Chamber slides were incubated at 37°C in 5% CO_2. Once the incubation time had elapsed, the chambers were removed, and the slides were fixed for ten minutes in 1% PFA. Immunocytochemical staining was then undertaken.

Jurkat Apoptosis Control Using Anti-CD95 Antibody

Healthy Jurkat cells were first transferred to chamber slides (3×10^5 cells per chamber). One-quarter of the chambers were then pre-treated with caspase inhibitor z-VAD-fmk and one-quarter with caspase inhibitor z-IETD-fmk (1 mL of RPMI media supplemented with inhibitor to give a concentration of 100 μM in each chamber) for ten hours.

After the ten hour pre-treatment with caspase inhibitors alone, CH11 (0.33 μg/mL) (anti-CD95 antibody clone IgM (Beckman Coulter-Immunotech, Burlington, ON)) was added to all chambers except one, which was untreated. Following fifteen hours of incubation, the contents of each chamber were transferred to microcentrifuge tubes and were spun at 300 x g for ten minutes. The

cells recovered in each pellet were then fixed in 500 μL of 1% PFA for ten minutes. The fixed cell samples were then cytospun as described above and the resulting slides were stained.

Detection of Apoptosis Using the TUNEL Assay

The ApopTag® Plus Peroxidase *In Situ* Apoptosis Detection Kit (Intergen Company, Purchase, NY) was used in all TUNEL assays as suggested by manufacturer's protocol, except that counterstaining was carried out for twenty minutes with methyl green solution at a pH of 5.0.

The reagents provided in the kit are designed to label the free 3' OH DNA termini *in situ* with chemically labeled and unlabeled nucleotides. These nucleotides are enzymatically added to the DNA ends by terminal deoxynucleotidyl transferase (TdT) via a template-independent process. The incorporated nucleotides form an oligomer of digoxigenin nucleotide and unlabeled nucleotide in a random sequence. DNA fragments that have been labeled with the digoxigenin-nucleotide are then allowed to bind to an anti-digoxigenin antibody conjugated to peroxidase. The bound peroxidase antibody conjugate enzymatically generates a permanent, localized stain from chromogenic substrates, providing sensitive detection in tissues, or in individual cells (The Complete ApopTag® Manual, Intergen Company).

Detection of Apoptosis Using the Sandwich ELISA

As an alternate method of apoptosis detection, the Cell Death Detection ELISA^{Plus} Kit (Roche Molecular Biochemicals, Mannheim, Germany) was used as described by the manufacturer's instructions. This assay is used for the qualitative and quantitative *in vitro* determination of cytoplasmic histone-associated DNA fragments (mono- and oligonucleosomes) formed as a result of induced cell death (From the Cell Death Detection ELISA^{Plus} Kit Instruction Manual, Roche Molecular Biochemicals Website http://biochem.roche.com).

The samples (in this case, co-culture supernatant) were placed in the wells of a streptavidin coated microtiter plate. To these wells, a mixture of anti-histone-biotin and anti-DNA-peroxidase (POD) was added. Once unbound components were washed out of the wells, nucleosomes in the samples were quantified by detecting the peroxidase still bound in the complexes. Peroxidase concentration was determined photometrically, five minutes after the addition of peroxidase substrate 2,2'-Azino-di- [3-ethylbenzthiazoline sulfonate] (ABTS) (from the Cell Death Detection ELISA^{Plus} Kit Instruction Manual, Roche Molecular Biochemicals Website http://biochem.roche.com).

Co-culture growth supernatants that had been frozen at -80°C after earlier TUNEL experiments were used as samples in these experiments. Samples were thawed and spun at 200 x g for ten minutes to remove cells and debris. Supernatant aliquots were then added to the appropriate wells of the microtiter plate and processed as per the manufacturer's instructions. Culture supernatant from camptothecin-treated McCoy cells was used as a positive control specific for these cells (See "Positive Apoptosis Control in McCoy Cells" above).

Cell Enumeration and Image Capture

A set of criteria was established, in order to gain consistency in the enumeration of cells. For each chamber or sample, at least five separate microscope fields were counted, with the minimum acceptable number of enumerated cells per chamber set at three hundred.

Fixed cells were observed under an Olympus BH-2 microscope fitted with a Sony Power HAD 3CCD Colour Video Camera. The video camera allowed for the capture of the desired still images, which were then transferred into Image Pro Plus software for digital preservation. All images were captured at a magnification of 400x.

RESULTS

Measurement of DNA Strand Breaks in McCoy Cells Using the TUNEL Assay

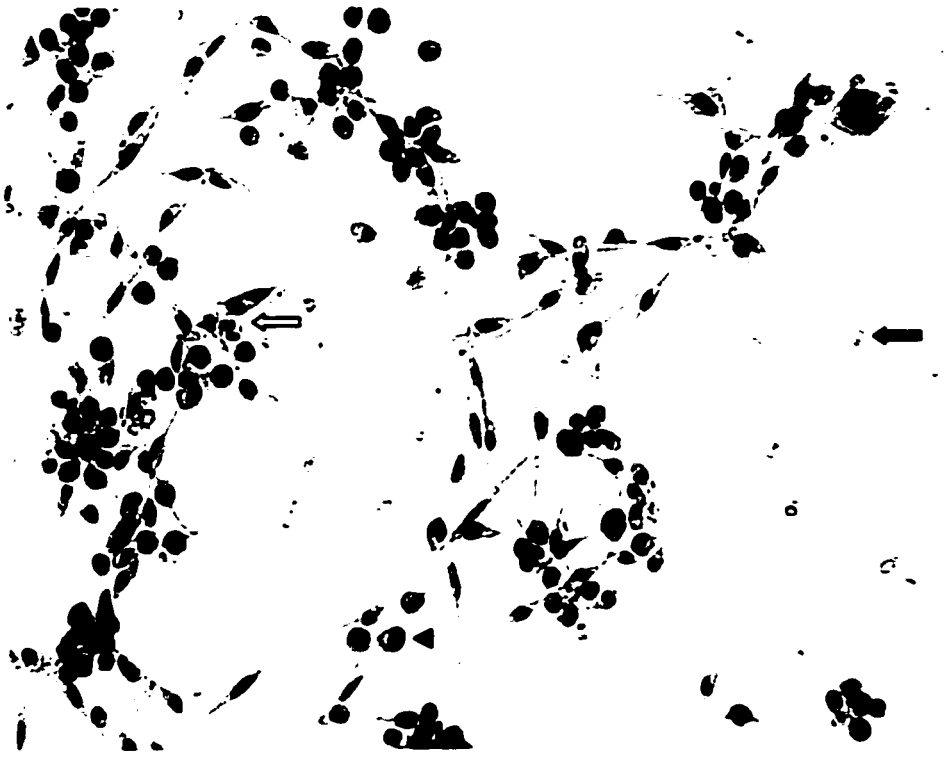
The TUNEL assay, which allows for the differentiation between apoptotic and non-apoptotic cells via a colourimetric reaction, was used to detect *T. vaginalis*-induced apoptosis in McCoy cells.

McCoy cell monolayers were allowed to grow to confluency and were incubated with fresh media (control) or media containing *T. vaginalis* in specific concentrations (ranging from 10^4 to 10^6 cells/mL). Following six hours of co-incubation in divided chamber slides, monolayers were rinsed with PBS and the remaining cells were fixed on the glass. Fixed slides were then treated with a TUNEL assay which employs a colourimetric reaction to distinguish between apoptotic (brown) cells and non-apoptotic (blue counterstained) cells.

Figure 4 shows the TUNEL-stained monolayer following six hours of co-culture *with T. vaginalis*. There were three distinct cell types present. The most prevalent cell type was blue and fibroblast-like in shape with a large nucleus in relation to the size of the whole cell (an example is indicated by the arrowhead in Figure 4). Based on staining observed in untreated control monolayers (not shown), these appeared to be non-apoptotic, adherent McCoy cells. The second cell type, indicated by the open arrow, was dark brown and had a large nucleus. These were relatively rare and appeared to be apoptotic McCoy cells. The third type of cell observed in the TUNEL experiments, indicated by the solid arrow, had not taken up any of the blue counterstain, but were very lightly stained a greyish/beige colour and were much smaller than the other cells observed. In addition, the nuclei seen in these cells were quite small and appeared to have a spindle or pinpoint shape.

Figure 4: Co-Culture Illustrating Three Cellular Morphologies

TUNEL assay on adherent cells remaining after co-culture. The first population consisted of fibroblast-shaped, blue cells with large dark-staining nuclei (arrowhead). The second type of cell, clearly apoptotic, also had a large nucleus, but it was stained dark brown (open arrow). The third cell type was smaller, did not counterstain well, but were light greyish/beige in colour (closed arrow). These cells also contained small pinpoint or spindle-shaped nuclei.



To address the possibility that the beige cells were trichomonads, *T. vaginalis*, trypsinized McCoy cells, and a mixture of both, were cytospun and stained with the TUNEL assay. This allowed for comparison of cellular morphology. The two cell types were found to be in the same approximate size range, although *T. vaginalis* showed more variability. With respect to staining, the McCoy cells counterstained well, while the *T. vaginalis* did not. For this reason, the trichomonads were more difficult to see. Non-specific brown staining of *T. vaginalis* was not observed, however. It was also discovered that the cytospun *T. vaginalis* did not resemble the lightly stained, small, beige cells found attached to the chamber slides following co-culture of *T. vaginalis* and McCoy cells (not shown).

It was speculated that perhaps the poorly-counterstained, beige cells were apoptotic McCoy cells. When these cells were compared with the positive control tissue (fixed rat mammary gland section (Strange *et al.*, 1995)) supplied with the TUNEL kit, however, the faint staining seen in the cells in question (solid arrow in Figure 4) was unlike the darker brown colour of the positive cells in the kit (solid arrows in Figure 5). This led to some doubt as to the accuracy of counting the small beige cells as apoptotic. Therefore, they were counted as separate, "possibly apoptotic", cells in all subsequent experiments. Hence, there are three cell types in all tables: blue non-apoptotic cells, brown apoptotic cells, and beige cells which, at the time, were assessed as possibly apoptotic.

Initial TUNEL experiments showed that when incubated in media alone (i.e. no *T. vaginalis*), the McCoy monolayers remained largely intact and non-apoptotic McCoy cells made up 99.7% of the cells counted (Table 1). With the addition of trichomonads to the co-incubation, there were found to be fewer dark blue apoptosis-negative cells. As can be seen in Table 1, when the McCoy cells were incubated with 10^6 *T. vaginalis*, the majority of the cells counted after the six-hour co-incubation were the lightly positive greyish/beige cells (98.7%). Very few large blue McCoy

Figure 5: Rat Mammary Gland Stained with TUNEL Assay Reveals Apoptotic Cells

Panels **(a)** and **(b)** are two fields of TUNEL-stained sections of rat mammary gland that act as positive controls for the ApopTag® Plus Peroxidase *In situ* Apoptosis Detection Kit (Intergen Company). The dark brown apoptotic cells are easily detected (arrow).


a


b

Table 1: Initial TUNEL Experiments

Early TUNEL experiments showed that when increasing numbers of trichomonads were added to the co-culture, fewer healthy McCoy cells resulted. After co-incubation, the majority of cells remaining were found to be stained a light beige colour, suggesting that apoptosis may have been occurring in McCoy monolayer cells as a result of the six hours of exposure to *T. vaginalis*.

Co-Culture	blue cells negative for apoptosis-% of total cells counted (raw number)	brown cells positive for apoptosis-% of total cells counted (raw number)	slightly beige cells which may be apoptotic-% of total cells counted (raw number)
McCoy cells alone	99.7% (1517)	0.3% (4)	0% (0)
McCoys and 10^4 *T. vaginalis*	97.0% (1189)	0% (0)	3.0% (37)
McCoys and 10^5 *T. vaginalis*	48.0% (547)	0% (0)	52.0% (593)
McCoys and 10^6 *T. vaginalis*	1.3% (27)	0% (0)	98.7% (2017)

cells remained (1.3% of cells counted).

A timecourse experiment (Table 2) showed that after four and six hours of co-incubation, almost all of the dark blue McCoy cells were replaced by the lightly staining beige cells. The faintly stained cells made up 99.7% of the cells counted on slides left to co-incubate for four hours with 10^6 *T. vaginalis* and made up 99.1% of cells counted on slides left to incubate for six hours with 10^6 *T. vaginalis*. A similar trend was seen at two hours, where 70.2% of the cells counted were light beige in colour. Throughout the TUNEL experiments, the percentage of cells that were dark brown and therefore definitely positive for apoptosis, remained uniformly low (generally less than 1%) (Tables 1 and 2).

Role of Contact Between *T. vaginalis* and McCoy Cells in Monolayer Destruction and Induction of McCoy Cell Apoptosis

Based on the hypothesis that apoptosis was taking place in the McCoy cells, the mechanism of its induction was assessed. It has been well established that *T. vaginalis* exerts a negative effect on cell monolayers, but there is some debate as to the need for contact in this process. Thus, McCoy cell monolayers were grown to confluency, but rather than being incubated with whole trichomonads, *T. vaginalis* culture supernatant was added to the monolayers. Culture supernatant was prepared by growing *T. vaginalis* to mid-log phase and removing the cells by centrifugation. The *T. vaginalis* supernatant was filtered and the pH was adjusted prior to co-incubation. Control monolayers incubated with fresh media showed no signs of damage after six hours (100% blue McCoy cells-Table 3). Monolayer cells incubated with the *T. vaginalis* culture supernatant, however, had to be recovered with a cytospin technique because the monolayer almost completely detached from the glass slide. Once these cells were collected, spun onto fresh glass slides, and stained with the TUNEL assay, it was observed that 100% of them were negative for

Table 2: Timecourse Experiment

Co-cultures were conducted as before, but this time cells were left for two, four and six hours. Similar to initial TUNEL experiments, the timecourse showed an inverse relationship between the number of trichomonads and the number of healthy McCoy monolayer cells remaining after four and six hours of co-culture. (The same trend was seen, to a lesser extent, after two hours of co-culture). This experiment also showed that with increasing co-incubation times, there were fewer healthy blue McCoy cells and more lightly stained beige (and possibly apoptotic) cells.

Co-Culture	blue cells negative for apoptosis-% of total cells counted (raw number) after 2, 4, and 6 h			brown cells positive for apoptosis-% of total cells counted (raw number) after 2, 4, and 6 h			slightly beige cells which may be apoptotic-% of total cells counted (raw number) after 2, 4, and 6 h		
	2 h	**4 h**	**6 h**	**2 h**	**4 h**	**6 h**	**2 h**	**4 h**	**6 h**
McCoy cells alone	99.8% (1316)	99.7% (997)	99.0% (900)	0.2% (3)	0.3% (3)	1.0% (9)	0% (0)	0% (0)	0% (0)
McCoys and 10^4 *T. vaginalis*	99.8% (1370)	98.0% (904)	95.6% (986)	0% (0)	0.4% (4)	0.1% (1)	0.2% (3)	1.5% (14)	4.3% (44)
McCoys and 10^5 *T. vaginalis*	87.2% (1072)	80.4% (701)	64.7% (738)	0.2% (3)	0.2% (2)	0.2% (2)	12.6% (155)	19.4% (169)	35.1% (400)
McCoys and 10^6 *T. vaginalis*	29.8% (248)	0.3% (2)	0.9% (9)	0% (0)	0% (0)	0% (0)	70.2% (584)	99.7% (763)	99.1% (996)

Table 3: Examining the Role of Contact in *T. vaginalis* Pathogenesis

McCoy cell monolayers were incubated with *T. vaginalis* supernatant for six hours. Monolayer cells detached from the glass slide after co-culture with the growth media but once recovered by cytospin, were found to be negative for apoptosis.

Co-Culture	blue cells negative for apoptosis-% of total cells counted (raw number)	brown cells positive for apoptosis-% of total cells counted (raw number)	slightly beige cells which may be apoptotic-% of total cells counted (raw number)
McCoys attached to chamber slide following incubation in fresh media only	100% (839)	0% (0)	0% (0)
McCoys recovered via cytospin following incubation with 1:4 supernatant dilution	99.9% (1003)	0.1% (1)	0% (0)
McCoys recovered via cytospin following incubation with 1:2 supernatant dilution	100% (1000)	0% (0)	0% (0)
McCoys recovered via cytospin following incubation with full strength supernatant	100% (1000)	0% (0)	0% (0)

apoptosis (Table 3 and Figure 6). Similarly, 1:4 and 1:2 dilutions of the *T. vaginalis* conditioned media yielded no apoptotic cells after the TUNEL assay.

Role of Caspases in *T. vaginalis*-Induced McCoy Cell Apoptosis

Co-incubation experiments were repeated in the presence of caspase inhibitors in order to explore the mechanism of apoptosis induction in the McCoy cells. Caspases play a central role in the majority of apoptotic cascades, therefore it was postulated that caspase inhibitors would prevent the induction of McCoy cell apoptosis by *T. vaginalis*. This would yield important information about the mechanism of apoptosis occurring in these cells and would also act as an alternate way of confirming that the McCoy cells were indeed dying by apoptosis.

As expected from previous experiments, McCoy cells incubated with media alone were healthy and 99.9% were stained blue after the TUNEL assay (Figure 7a and Table 4). Also following the pattern observed earlier, non-apoptotic blue cells made up only 84.3% and 0% of cells counted, when incubated with 10^5 (Figure 7b) and 10^6 *T. vaginalis*/mL respectively. The lightly staining beige cells (solid arrow) made up 15.7% and 100% of cells counted when the co-incubation included 10^5 (Figure 7b) and 10^6 *T. vaginalis*/mL respectively.

The lightly staining beige cells were also observed in similar numbers in chambers treated with the caspase inhibitors. As illustrated in Table 4, co-incubations of McCoy cells and 10^6 *T. vaginalis* in the presence of 10 μM and 100 μM z-IETD-fmk (a caspase-8 inhibitor) resulted in a uniform population (100% of cells counted) of the lightly staining beige cells. When McCoy cells were grown with 10^5 *T. vaginalis* in the presence of 10 μM and 100 μM z-IETD-fmk, it was observed that 41% and 49.6% of cells, respectively, stained light beige (solid arrows in Figures 7c and 7d) (Table 4). After co-incubation of McCoy cells and 10^6 *T. vaginalis* in the presence of 10

Figure 6: Addressing the Need For Contact in *T. vaginalis* Pathology

Incubation of McCoy cell monolayers with *T. vaginalis* supernatant was found to cause widespread detachment of the McCoy cells from the glass substrate. Once the detached McCoy cells were recovered via a cytospin, and stained with the TUNEL assay, it was discovered that they were rounded but not apoptotic.

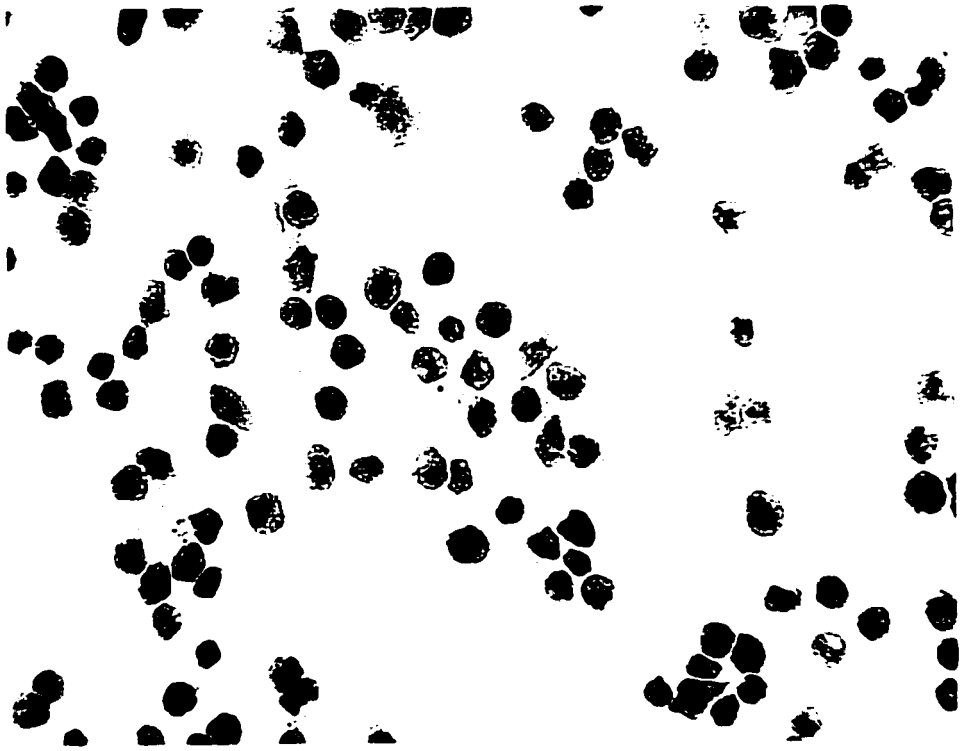


Figure 7: Effect of Caspase Inhibitors on McCoy Cells

The co-culture experiments were repeated in the presence of two different caspase inhibitors - z-IETD-fmk (a caspase-8 inhibitor) and z-VAD-fmk (a general caspase inhibitor). **(a)** McCoy cells incubated in media alone (control) and then stained with TUNEL assay were found to be non-apoptotic. **(b)** McCoy cells co-cultured with 10^5 *T. vaginalis*; some of the cells were stained light beige (arrow). The lightly stained beige cells (arrow) were also observed when McCoy cells were co-cultured with 10^5 *T. vaginalis* in the presence of 10 μM **(c)** or 100 μM **(d)** concentrations of z-IETD-fmk. **(e)** The same phenomenon was observed when the McCoy cells were co-cultured with 10^5 *T. vaginalis* in the presence of 10 μM z-VAD-fmk.


b
c
d
e

Table 4: Investigating the Involvement of Caspases

Co-culture experiments were repeated in the presence of caspase inhibitors (z-IETD-fmk and z-VAD-fmk). Caspase inhibitors did not have any effect on the numbers of slightly beige (and possibly apoptotic) cells present.

Caspase Inhibitor	**Co-Culture**	**blue cells negative for apoptosis-% of total cells counted (raw number**	**brown cells positive for apoptosis-% of total cells counted (raw number)**	**slightly beige cells which may be apoptotic-% of total cells counted (raw number)**
Control-No drug	McCoys in media alone	99.9% (2855)	0.1% (3)	0% (0)
	McCoys and 10^5 *T. vaginalis*	84.3% (307)	0% (0)	15.7% (57)
	McCoys and 10^6 *T. vaginalis*	0% (0)	0% (0)	100% (300)
10 μM z-IETD-fmk	McCoys with drug but no *T. vaginalis*	100% (650)	0% (0)	0% (0)
	McCoys with drug and 10^5 *T. vaginalis*	59.0% (453)	0% (0)	41.0% (315)
	McCoys with drug and 10^6 *T. vaginalis*	0% (0)	0% (0)	100% (314)
100 μM z-IETD-fmk	McCoys with drug but no *T. vaginalis*	100% (557)	0% (0)	0% (0)
	McCoys with drug and 10^5 *T. vaginalis*	50.4% (227)	0% (0)	49.6% (223)
	McCoys with drug and 10^6 *T. vaginalis*	0% (0)	0% (0)	100% (472)
10 μM z-VAD-fmk	McCoys with drug but no *T. vaginalis*	100% (583)	0% (0)	0% (0)
	McCoys with drug and 10^5 *T. vaginalis*	48.5% (236)	0% (0)	51.5% (251)
	McCoys with drug and 10^6 *T. vaginalis*	0.2% (1)	0% (0)	99.8% (568)

μM z-VAD-fmk (a general caspase inhibitor), 99.8% of the cells enumerated were found to be light beige. A co-culture of McCoy cells and 10^5 *T. vaginalis* under the same conditions showed that 51.5% of those cells enumerated were light beige in colour (solid arrow in Figure 7e) (Table 4). It should be noted that in this experiment, higher concentrations of z-VAD-fmk were used in some chambers, but it was found that these concentrations exerted direct negative effects on the McCoy cells, resulting in disruption of the monolayer in control as well as test chambers.

Given that the caspase inhibitors did not block what we had postulated to be apoptosis in the McCoy cells, it was important that the inhibitors be tested in a cell system where apoptosis was confirmed. A Jurkat control, in which apoptosis was induced and inhibited, was established to demonstrate the ability of the caspase inhibitors to block caspase-8 (z-IETD-fmk) and caspases in general (z-VAD-fmk). The Jurkat cells were cultured and were then pretreated with the caspase inhibitors (z-IETD-fmk and z-VAD-fmk) for ten hours. After the pretreatment, an anti-CD95 antibody (CH11) was added to all chambers and the cells were left for a fifteen-hour incubation period. Controls were also included in which Jurkat cells were cultured in media with no CH11 or inhibitor and in media with CH11 only (no inhibitor). When Jurkat cells were incubated in media alone (with no treatment), 100% of the cells were blue and non-apoptotic (Figure 8a), while those cells treated with CH11 yielded two populations following the TUNEL assay. 24.9% of the cells counted were dark brown and definitely apoptotic (closed arrow in Figure 8b), while 75.1% were found to be blue and non-apoptotic (Table 5). The numbers of non-apoptotic cells returned to control levels when the caspase inhibitors were added to the incubation chambers. 99.7% of Jurkat cells treated with CH11 and z-IETD-fmk were non-apoptotic (Figure 8c) while 100% of those treated with CH11 and z-VAD-fmk were found to be non-apoptotic (Figure 8d and Table 5). This suggested that the caspase inhibitors were working properly and that the findings in the McCoy/*T. vaginalis* caspase inhibition experiment could not be explained by inactivity on the part of the inhibitors.

Figure 8: Control Experiment to Verify Efficacy of Caspase Inhibitors

Jurkat cells were incubated with an anti-CD95 antibody (CH11). **(a)** Jurkat cells incubated in media alone, cytospun, and stained with the TUNEL assay. **(b)** Jurkat cells, a number of which appear to be apoptotic (arrow), following incubation with the anti-CD95 antibody. **(c)** and **(d)** Non-apoptotic Jurkat cells were the result when the culture media contained both the anti-CD95 antibody and a caspase inhibitor (z-IETD-fmk **(c)** or z-VAD-fmk **(d)**).


a
b
c
d

Table 5: Caspase Inhibitor Control

Anti-CD95 induced apoptosis was inhibited in Jurkat cells via the use of caspase inhibitors (z-IETD-fmk and z-VAD-fmk) in the culture media.

Co-Culture	blue cells negative for apoptosis-% of total cells counted (raw number	brown cells positive for apoptosis-% of total cells counted (raw number)	slightly beige cells which may be apoptotic-% of total cells counted (raw number)
Jurkats recovered via cytospin following incubation with media only	100% (600)	0% (0)	0% (0)
Jurkats recovered via cytospin following incubation with CH11	75.1% (426)	24.9% (141)	0% (0)
Jurkats recovered via cytospin following incubation with CH11 and z-IETD-fmk (100 μM)	99.7% (685)	0.3% (2)	0% (0)
Jurkats recovered via cytospin following incubation with CH11 and z-VAD-fmk (100 μM)	100% (765)	0% (0)	0% (0)

Adherence of *T. vaginalis* to Monolayer Following Detachment of McCoy Cells

During the course of one of the six-hour *T. vaginalis*/McCoy co-incubation experiments, prior to fixation, the co-culture was examined under the inverted microscope. It was noted that many of the cells stuck to the monolayer possessed the distinct corkscrew movement patterns of trichomonads. It was hypothesized that *T. vaginalis* was inducing apoptosis in the McCoy cells but that these compromised cells were detaching and allowing for the attachment of the *T. vaginalis*. It was estimated that by collecting detached cells after the usual six-hour co-incubation of McCoy cells and *T. vaginalis*, this possibility could be addressed. The detached cells were cytospun onto fresh glass slides and these slides were stained with TUNEL reagents.

When McCoy cells were incubated with fresh media only, 99.9% of the attached cells counted were blue and non-apoptotic (Table 6). The co-incubation media from the control chambers were not cytospun because of a lack of cellular detachment in these chambers. (The monolayers were fully intact). At a *T. vaginalis* concentration of 10^4 trichomonads/mL, the large majority of cells counted were blue, non-apoptotic McCoy cells (98.7% and 99.1% of attached and detached cells counted respectively) (Table 6). When the monolayers were incubated with 10^5 *T. vaginalis*/mL, 70.5% of the attached cells were assessed to be blue and non-apoptotic, while 84.9% of the detached cells were non-apoptotic. None of the detached cells from the 10^4 *T. vaginalis*/mL or 10^5 *T. vaginalis*/mL co-incubations were found to be dark brown (apoptotic). Following the co-incubation of McCoy cells and 10^5 *T. vaginalis*/mL, 29.5% of the attached cells were found to have been lightly stained and beige in colour, while 15.1% of the detached cells were lightly stained. Figure 9 illustrates the cells captured via cytospin following the co-incubation of a McCoy cell monolayer and 10^5 *T. vaginalis*. Because of the massive destruction of the monolayers when McCoy cells were incubated with 10^6 *T. vaginalis*, the cytospin yielded too many cells to count properly. It was noted, however, that none of the cells in these cytospins was

Table 6: Assessing Detached Monolayer Cells

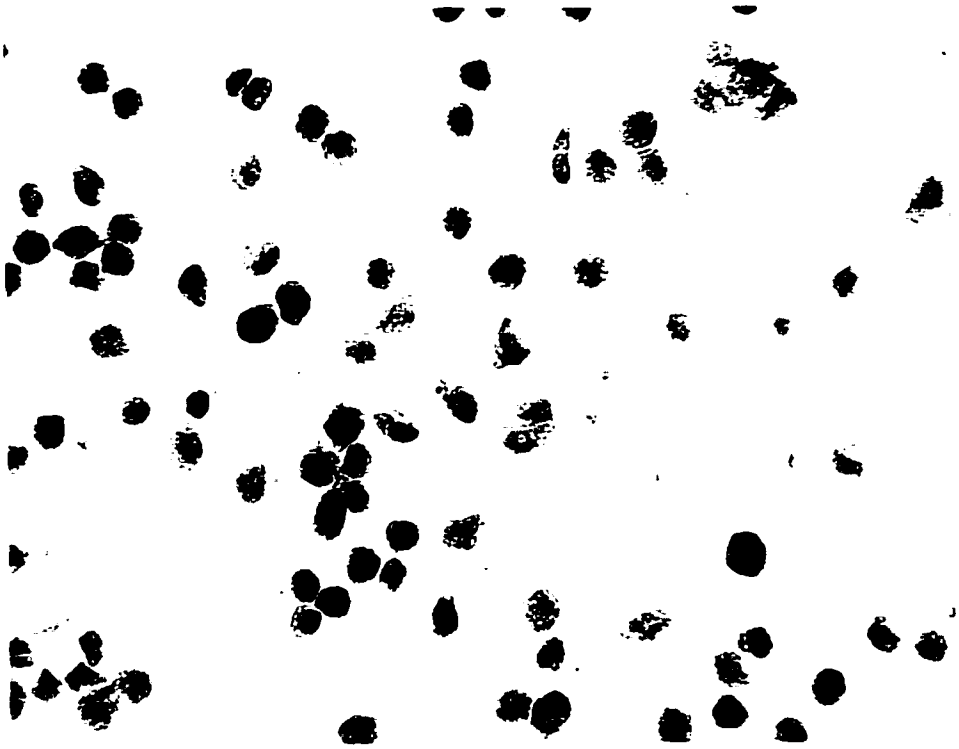
Co-culture experiments were conducted as per the standard protocol. Following the six hour co-incubation, TUNEL staining was done on both attached cells and detached cells recovered by cytospin from the co-culture media. Increasing concentrations of *T. vaginalis* led to augmented detachment of monolayer cells. None of the detached cells were found to stain dark brown.

Co-Culture	blue cells negative for apoptosis-% of total cells counted (raw number)		brown cells positive for apoptosis-% of total cells counted (raw number)		slightly beige cells which may be apoptotic-% of total cells counted (raw number)	
	attached to chamber slide following incubation	recovered via cytospin following incubation	attached to chamber slide following incubation	recovered via cytospin following incubation	attached to chamber slide following incubation	recovered via cytospin following incubation
McCoy cells alone	99.9% (1250)	n/a	0.1% (1)	n/a	0% (0)	n/a
McCoys and 10^4 *T. vaginalis*	98.7% (1072)	99.1% (883)	0.2% (2)	0% (0)	1.1% (12)	0.9% (8)
McCoys and 10^5 *T. vaginalis*	70.5% (659)	84.9% (1095)	0% (0)	0% (0)	29.5% (276)	15.1% (194)
McCoys and 10^6 *T. vaginalis*	0.1% (1)	(>1200)*	0% (0)	(0)	99.9% (949)	(>1200)*

Note: * = too many cells cytospun onto slide to count

Figure 9: Lack of Apoptosis in Detached McCoy Cells

Detached McCoy cells recovered, after co-incubation with 10^5 *T. vaginalis*, by cytospin, and stained with TUNEL assay.



dark brown in colour (Figure 9). When the *T. vaginalis* and McCoy cells contained in the co-incubation media were cytospun, it became quite difficult to count the *T. vaginalis*. The trichomonads did not counterstain well and tended to flatten onto the slide (Note the faint, lightly stained cells in the background of Figure 9).

Induction of Apoptosis in McCoy Cells Using Camptothecin (Positive Apoptosis Control)

In order to determine whether the lightly stained beige cells were apoptotic McCoy cells or adherent *T. vaginalis*, a positive apoptosis control in McCoy cells was required. This was achieved by incubating McCoy cells with camptothecin (CAM), which is used to induce apoptosis in a number of different cell types. McCoy cell monolayers were incubated in media supplemented with CAM (5 μM or 10 μM concentrations) for four hours and were then left for forty-eight hours in regular growth media. At this point, slides were stained with TUNEL and cells were counted.

McCoy cells incubated in media only (no CAM) were largely blue and non-apoptotic (94.3%) (Table 7), with a slightly higher than normal background level of apoptotic cells (5.7%) likely due to natural deterioration of the monolayer after the long post-confluency incubation period. 17.6% of the McCoy cells treated with 5 μM CAM (solid arrow in Figure 10 and Table 7) and 23.3% of the McCoy cells treated with 10 μM CAM were found to be dark brown and apoptosis positive. It should be noted that no lightly staining beige cells with spindle nuclei were seen in the culture. The colour of the positive cells in this experiment was comparable to the brown staining seen in the kit control tissues (solid arrows in Figures 5 and 10) and was easily distinguished from the blue counterstaining seen in the non-apoptotic McCoy cells (open arrow in Figure 10).

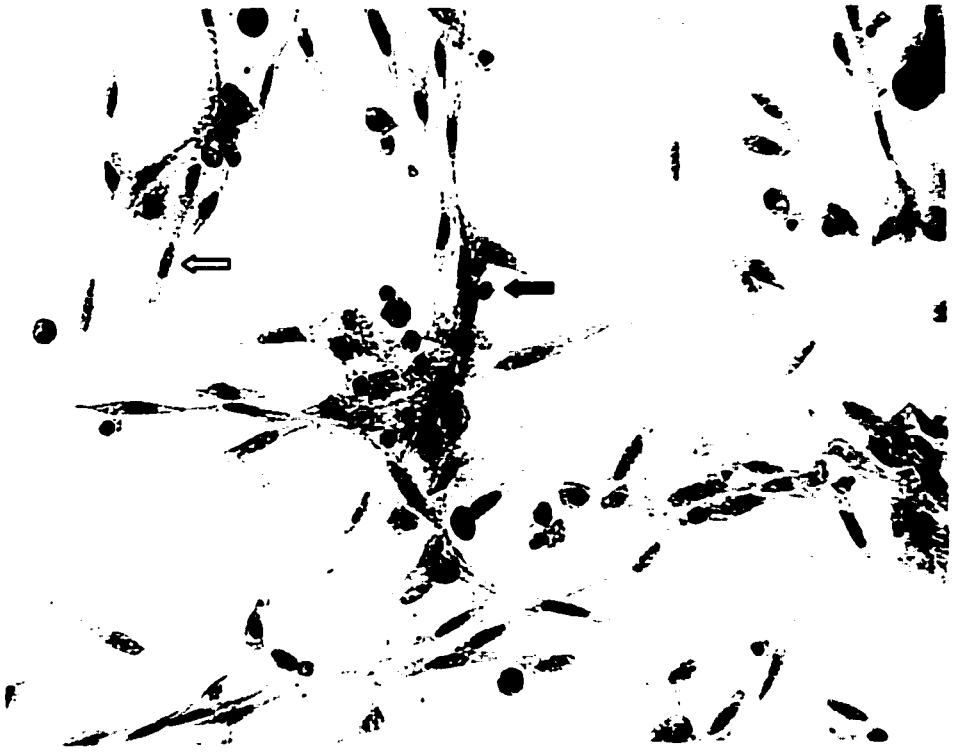
Table 7: Apoptosis Control in McCoy Cells

McCoy cells were incubated with camptothecin (CAM) for four hours before being stained with the TUNEL assay. More apoptotic cells were detected in those monolayers treated with CAM than in those incubated in media alone. Examination of the treated monolayer cells revealed none of the lightly stained beige cells that had been observed in previous experiments.

Co- Culture	blue cells negative for apoptosis-% of total cells counted (raw number)	brown cells positive for apoptosis-% of total cells counted (raw number)	slightly beige cells -% of total cells counted (raw number)
McCoy cells in media only	94.3% (741)	5.7% (45)	0% (0)
McCoys in media with 5 μM CAM	82.4% (575)	17.6% (123)	0% (0)
McCoys in media with 10 μM CAM	76.7% (490)	23.3% (149)	0% (0)

Figure 10: Apoptosis of McCoy Cells Induced by CAM

McCoy cells were incubated with camptothecin (CAM) (5 μM concentration) for four hours. After a further forty-eight hours, cells were fixed and stained with the TUNEL assay. Resulting apoptotic cells (closed arrow) were dark brown and were easily distinguished from the blue, non-apoptotic McCoy cells (open arrow).



Measurement of DNA/Histone Fragments Resulting from Apoptosis by ELISA Assay (An Alternate Means of Assessing Apoptosis)

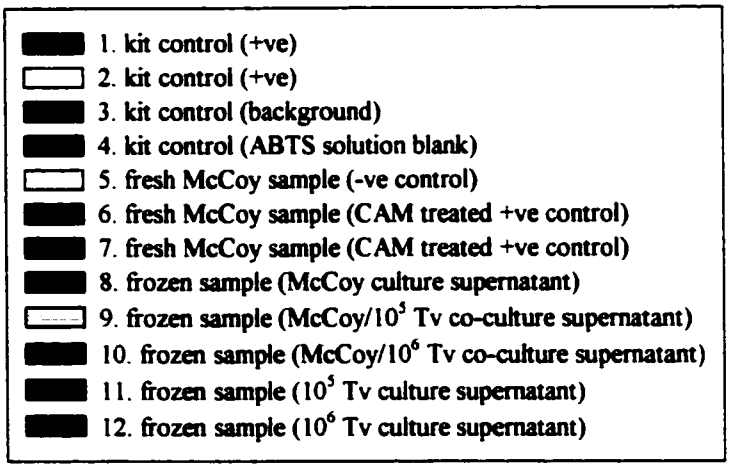
The Cell Death Detection ELISA^{Plus} Kit was employed as an alternative method for identifying apoptosis in the co-cultures of McCoy cells and *T. vaginalis*. If the trichomonads were inducing apoptosis in the McCoy cells during co-culture, the resulting DNA/histone fragments in the culture supernatant would be detected via the sandwich ELISA.

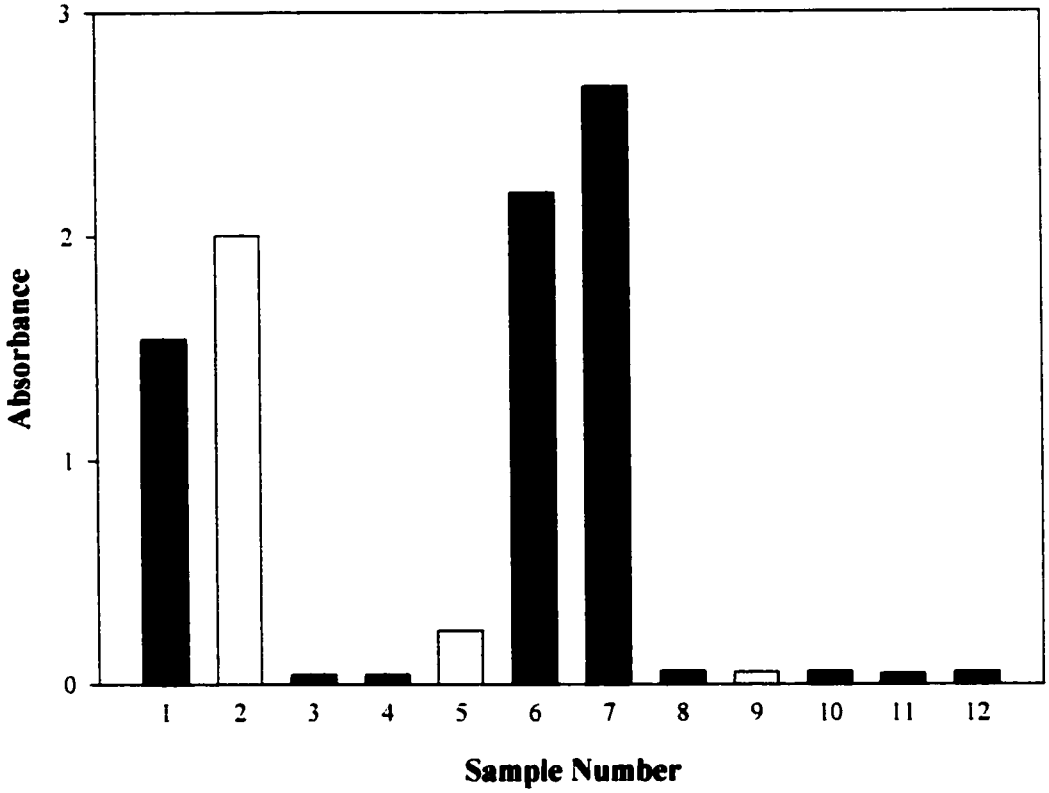
As expected, the negative controls included in the Cell Death Detection ELISA^{Plus} Kit gave very low absorbance values (bars 3 and 4 in Figure 11). Conversely, positive controls provided with the kit gave significantly higher absorbance values when tested with the ELISA protocol (bars 1 and 2 in Figure 11). Culture supernatants from McCoy monolayers were also incorporated, in a control capacity, into the ELISA experiment. As outlined in the Materials and Methods, monolayers incubated with the apoptosis-inducing compound CAM served as a positive control. Figure 11 illustrates the large difference in absorbance values between the co-culture supernatant taken from a monolayer grown in media only (bar 5) and those incubated instead with CAM-containing media (bars 6 and 7). This confirmed that CAM induces DNA fragmentation in McCoy cells.

The last group of co-culture supernatant samples used in the ELISA experiment came from frozen aliquots preserved from earlier TUNEL experiments. If DNA fragmentation was occurring in the McCoy cell/*T. vaginalis* co-incubations, the ELISA assay should detect the resulting DNA/histone fragments released into the culture supernatant. As can be seen in Figure 11, control supernatants (McCoy cells grown in media alone (bar 8) and *T. vaginalis* cultured alone (bars 11 and 12)) showed very low absorbance levels, indicating a lack of DNA/histone fragments. Very low (approaching background) levels of absorbance for supernatants from the

Figure 11: Cell Death Detection ELISA

The Cell Death Detection ELISA^{PLUS} Kit (Roche Molecular Biochemicals) was used to test for apoptosis in McCoy cells co-cultured with *T. vaginalis* (Tv). Negative kit controls gave very low absorbance values (bars 3 and 4). Positive kit controls gave comparatively high absorbance levels (bars 1 and 2). McCoy cells incubated with media alone (bar 5) gave an absorbance comparable to the background level, while those that had been cultured with CAM (to induce apoptosis) (bars 6 and 7) provided absorbance values greater than the kit positive controls. This indicated that CAM-induced McCoy cell apoptosis was being detected by the sandwich ELISA assay. Supernatants from six-hour co-cultures of McCoy cells *with T. vaginalis* were also assayed for DNA/histone fragments; the resulting low absorbance values (bars 9 and 10), suggested that apoptosis did not occur in the McCoy cells cultured with *T. vaginalis*. As expected, no DNA/histone fragments were detected in supernatant from McCoy cell growth alone in media (bar 8) or from *T. vaginalis* cultured alone in media (bars 11 and 12).




Absorbance
3
2
1
0
1
2
3
4
5
6
7
8
9
10
11
12
Sample Number

McCoy/10^5 *T. vaginalis* (bar 9 in Figure 11) and the McCoy/10^6 *T. vaginalis* (bar 10 in Figure 11) co-incubations suggested that DNA/histone fragments were not present and therefore that apoptosis was not taking place. The ELISA findings further supported the data from the TUNEL assays and indicated that the lightly staining cells seen in previous histological analyses of co-cultures were not apoptotic McCoy cells.

Assessment of Adherent *T. vaginalis* via the TUNEL Assay

All signs pointed to the beige cells seen in the co-incubation experiments being *T. vaginalis*, but this could not be concluded until a *T. vaginalis* control was developed. To do this, healthy *T. vaginalis* (at varying concentrations) was transferred to separate chamber slides and allowed to incubate for six hours. All conditions were comparable to the usual six-hour co-incubation, except for the omission of the McCoy cells. Once the slides were stained with TUNEL, it was discovered that the *T. vaginalis* did stick quite well to the glass substrate of the chamber slides (Figure 12a). Their morphology there was also noted to be quite different from cytospun *T. vaginalis*. In fact, it was observed that the *T. vaginalis* allowed to attach to the glass during a six hour incubation period (Figure 12a) looked almost identical to the lightly staining beige cells with the spindle nuclei which had previously been identified as apoptotic McCoy cells (open arrow in Figure 12b). McCoy cells grown in the absence of *T. vaginalis* (Figure 12c) have a distinctly different morphology from the trichomonads.

Figure 12: Morphology of *T. vaginalis* Adherent to Glass Substrate (Alone and in Co-Culture with McCoy Cells)

(a) Healthy *T. vaginalis* was added to chamber slides and was then allowed to incubate for six hours. After fixation and staining with TUNEL, it became apparent that the parasite did indeed adhere well to the glass substrate. It was also noted that the trichomonads did not counterstain and took on a faint beige hue. **(b)** A co-culture of *T. vaginalis* and McCoy cells. The large blue cells are non-apoptotic McCoy cells (closed arrow), and by comparison with **(a)**, it is apparent that the second cell type in **(b)** is adherent *T. vaginalis* (open arrow). **(c)** McCoy cells incubated on their own display the expected morphology and staining following treatment with the TUNEL assay. No apoptosis is apparent and no lightly-stained beige cells can be found.


a
b
c