Binaural Speech Intelligibility Prediction and Nonlinear Hearing Devices

by

Nicolas N. Ellaham

A thesis
submitted to the Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of
Doctorate of Philosophy in Electrical and Computer Engineering

Ottawa-Carleton Institute for Electrical and Computer Engineering
School of Electrical Engineering and Computer Science (EECS)
University of Ottawa

© Nicolas N. Ellaham, Ottawa, Canada, 2014

Abstract

A new objective measurement system to predict speech intelligibility in binaural listening conditions is proposed for use with nonlinear hearing devices. Digital processing inside such devices often involves nonlinear operations such as clipping, compression, and noise reduction algorithms. Standard objective measures such as the Articulation Index (AI), the Speech Intelligibility Index (SII) and the Speech Transmission Index (STI) have been developed for monaural listening. Binaural extensions of these measures have been proposed in the literature, essentially consisting of a binaural pre-processing stage followed by monaural intelligibility prediction using the better ear or the binaurally enhanced signal.

In this work, a three-stage extension of the binaural SII approach is proposed that deals with nonlinear acoustic input signals. The reference-based model operates as follows: (1) a stage to deal with nonlinear processing based on a signal-separation model to recover estimates of speech, noise and distortion signals at the output of hearing devices; (2) a binaural processing stage using the Equalization-Cancellation (EC) model; and (3) a stage for intelligibility prediction using the SII or the short-time Extended SII (ESII).

Multiple versions of the model have been developed and tested for use with hearing devices. A software simulator is used to perform hearing-device processing under various binaural listening conditions. Details of the modeling procedure are discussed along with an experimental framework for collecting subjective intelligibility data. In the absence of hearing-device processing, the model successfully predicts speech intelligibility in all spatial configurations considered. Varying levels of success were obtained using two simple distortion modeling approaches with different distortion mechanisms. Future refinements to the model are proposed based on the results discussed in this work.

To my wife and my parents,

Who have taught me that hearing truly takes place in the heart.

Acknowledgements

This thesis is the result of a little over six years of work. It would not have been possible without the guidance and support of some of the most important people to have crossed my path. I would like to take this opportunity to recognize them and express my gratitude.

First, I would like to thank Dr. Christian Giguère and Dr. Wail Gueaieb for supervising my doctoral studies . Their guidance and support were invaluable to my graduate life in general, and this thesis in specific. Over the years, Dr. Giguère has become a research mentor to me and role model who has inspired me to carve my path as an engineer in the field of audiology and hearing sciences. It has been an absolute privilege to work so closely with a researcher of his caliber on this thesis, as well as the numerous other projects he oversees at the Hearing Research Lab. His passion for science, his research ethics, and his patient pedagogical approach to explain complicated concepts are truly remarkable and contagious. It has also been a pleasure to work with Dr. Gueiaeb and witness his unmatched attention to details, as well as his wise academic insights that helped me greatly, especially during the earlier years of my academic life.

I am also very grateful for the members of the thesis advisory committee: Dr. Martin Bouchard, Dr. Hilmi Dajani, Dr. Rafik Goubran, and Dr. Vijay Parsa. Their constructive feedback and insightful comments during the proposal and final thesis defence have greatly contributed to the quality of this thesis. Many thanks also go to anonymous reviewers of various articles that arose from this thesis, whose feedback played a great role in shaping its current form.

Special thanks go to Nadine Godin and Catherine L'Heureux for collecting the subjective data reported in Chapter 6, to Veronique Vaillancourt for her help in setting up the experimental environment, and all the subjects who participated in this study.

My thanks also to the the Natural Sciences and Engineering Research Council of Canada, the Government of Ontario, and the University of Ottawa for their financial support for this work.

On a personal note, I am eternally indebted to the unwavering support of my loving parents Nasri and Agathe Laham, and my wonderful siblings Sami and Hiba whom I look up to. It is from them that I gather my passion for higher education and the pursuit of knowledge. Thank you for always being there for me. Mom, sharing our doctoral journeys together has been a source of inspiration. We made it!

This journey would have been unbearable without the support of many wonderful friends. Dr. Nizar Sakr, a tremendous source of encouragement and inspiration, a great coach who knows when and how much to gently push a friend, and a wise adviser who's ready to find the most sensitive approach to challenging situations; Flora Sakr, with her calm presence in the lab, always ready to find the silver lining and positive twist to life; Hanna Farah, for 5 years of being the most light-hearted and relaxing roommate anyone can come home to find (it was a fun journey bro); Wissam, Maya, Karen, Lara, Bassam, George, Nibal, Robert, Joe, Kostis and so many others whose names I would not be able to list, but whose friendship I treasure dearly.

A special thank you as well to my supportive relatives, especially my proud uncle Elie, who is always seeking to introduce me with embarrassing cheer-leading praise: "meet my nephew, the PhD"; my in-laws, the Saikaly's, Walid, Sabah, Corine, George, and Ely... The gatherings around your table provided so much more than food; and the supporting community at St. Elias Cathedral, my spiritual home where I recharge my batteries.

Last but certainly not least, I am forever grateful for my wonderful wife without whom I would not have crossed the finish line. Your love and encouragement, ever present throughout the good times and the most stressful of times, have carried me through to the end, especially when health concerns arose. A teacher by profession, and in every aspect of your life, your steadfast willingness to share my burden is my foremost teacher of endurance. I love you.

"And glory be to God for all things." (St. John Chrysostom, 5th c. A.D.)

Contents

List of Figures **xiii**

List of Tables **xvii**

List of Acronyms **xix**

1 Introduction **1**

1.1 Research background 1

1.1.1 Motivation 1

1.1.2 Research problem 2

1.1.3 Methodology and approach 3

1.2 Contributions 5

1.3 Thesis outline 7

2 Psychoacoustic Principles **8**

Introduction 8

2.1 Overview of hearing 8

2.1.1 Sound generation and propagation 9

2.1.2 Anatomy and function of the ear 9

2.1.2.1 The outer ear 9

2.1.2.2 The middle ear 9

2.1.2.3 The inner ear 10

2.1.3 The auditory nervous system . . . 13
2.2 Characteristics of normal hearing . . . 13
2.2.1 Hearing thresholds . . . 14
2.2.2 Auditory filters and frequency selectivity . . . 15
2.2.3 Loudness perception . . . 16
2.2.4 Temporal processing . . . 17
2.2.5 Binaural hearing . . . 18
2.3 Hearing loss . . . 20
2.3.1 General characteristics of hearing loss . . . 20
2.3.1.1 Types of hearing loss . . . 20
2.3.1.2 Severity and configuration . . . 21
2.3.2 Assessment of hearing loss . . . 22
2.3.2.1 Pure-tone audiometry . . . 22
2.3.2.2 Speech audiometry . . . 23
2.3.2.3 Most comfortable and uncomfortable levels . . . 24
2.3.2.4 Other audiologic tests . . . 24
2.3.3 Supra-threshold deficits . . . 25
2.3.3.1 Loss of frequency selectivity . . . 25
2.3.3.2 Reduced dynamic range . . . 25
2.3.3.3 Loss of temporal resolution . . . 27
2.4 Hearing aids . . . 27
2.4.1 Historical evolution . . . 27
2.4.2 Types of hearing aids . . . 28
2.4.3 Digital hearing aid technologies . . . 30
2.4.3.1 Amplification and signal enhancement strategies . . . 30
2.4.3.2 Artificial intelligence . . . 32

3 Literature Review **34**

Introduction . . . 34

3.1 Speech intelligibility . . . 34

3.1.1 Subjective tests . . . 35

3.1.1.1 The Hearing In Noise Test . . . 36

3.1.1.2 The Modified Rhyme Test . . . 37

3.1.2 Objective prediction models . . . 38

3.1.2.1 The Speech Intelligibility Index . . . 39

3.1.2.1.1 Extensions of the SII . . . 40

3.1.2.2 The Speech Transmission Index . . . 40

3.1.2.2.1 Versions of the STI . . . 41

3.1.2.2.2 Speech-based STI extensions . . . 42

3.1.2.3 Comparing objective predictions to subjective scores . . . 43

3.2 Modeling nonlinear distortion in hearing devices . . . 44

3.2.1 Signal and distortion estimation methods . . . 44

3.2.1.1 The coherence function . . . 44

3.2.1.2 The phase-inversion method . . . 46

3.2.1.3 A Hilbert-transform distortion measure . . . 47

3.2.2 Intelligibility models for nonlinear processing . . . 49

3.2.2.1 The coherence SII . . . 49

3.2.2.2 A coherence-based STI model . . . 50

3.2.2.3 Other models of speech intelligibility . . . 51

3.3 Binaural hearing . . . 53

3.3.1 The “cocktail party phenomenon” . . . 54

3.3.2 Spatial release from masking: subjective studies . . . 55

3.3.3 Spatial release from masking: objective models . . . 57

3.3.3.1 Equalization-cancellation theory . . . 57

3.3.3.2 Models of binaural speech intelligibility . . . 58

3.4 Summary 61

4 Modeling and Measurement Approach 63

Introduction 63

4.1 Proposed Binaural Intelligibility Model 64

4.1.1 Stage 1: Signal separation and distortion estimation 65

4.1.2 Stage 2: Binaural processing 68

4.1.2.1 Gammatone analysis filterbank 68

4.1.2.2 Internal masking noise 69

4.1.2.3 Equalization-cancellation processing 69

4.1.2.4 Processing errors 70

4.1.3 Stage 3: Speech Intelligibility 71

4.1.3.1 Intelligibility prediction 71

4.1.3.2 Intelligibility mapping 71

4.1.4 Short-time extension 72

4.1.4.1 Additional HINT signals 72

4.1.4.2 Block processing 74

4.1.5 Summary and verification of model implementations 75

4.1.5.1 Model requirements and assumptions 75

4.1.5.2 Monaural verification of modeling components 76

4.1.5.3 Alternative model implementations 79

4.2 Hearing-device processing 80

4.2.1 The hearing aid simulator toolbox (HASTB) 80

4.2.2 The hearing device simulator toolbox (HDSTB) 82

4.2.2.1 Noise reduction 82

4.2.2.2 Compression 84

4.2.2.3 Clipping 85

4.2.3 Binaural extension 87

4.3 The Matlab speech testing environment . . . 89
4.4 Validation approach . . . 90
4.4.1 Monaural validation . . . 91
4.4.1.1 Rationale . . . 91
4.4.1.2 Simulated conditions . . . 91
4.4.2 Binaural validation . . . 92
4.4.2.1 Rationale . . . 92
4.4.2.2 Experimental conditions . . . 92
4.5 Conclusion . . . 94

5 Monaural Evaluation of the Phase-Inversion Method 95

Introduction . . . 95
5.1 Evaluation overview . . . 96
5.1.1 Rationale . . . 96
5.1.2 Simulation setup . . . 98
5.2 Results . . . 99
5.2.1 Recovered speech and noise spectra . . . 100
5.2.1.1 No hearing aid processing . . . 100
5.2.1.2 Linear and compressive processing . . . 101
5.2.1.3 Peak clipping . . . 102
5.2.2 Speech-to-distortion and noise-to-distortion ratios . . . 104
5.2.3 SII predictions . . . 106
5.2.3.1 No hearing aid processing . . . 106
5.2.3.2 Effects of hearing loss . . . 106
5.2.3.3 Isolating effects of nonlinearities . . . 107
5.2.3.4 Coherence and window length considerations . . . 109
5.2.3.5 SII adjustment for phase-inversion method . . . 109
5.3 Discussion . . . 111

5.4 Conclusion . . . 114

6 Subjective Measurement of Speech Reception Thresholds 115

Introduction . . . 115

6.1 Methods . . . 116

6.1.1 Experimental setup . . . 116

6.1.1.1 Hardware and software equipment . . . 116

6.1.1.2 SRT measurement procedure . . . 117

6.1.2 Experimental conditions . . . 118

6.1.2.1 Experiment A: Spatial release from masking . . . 118

6.1.2.2 Experiment B: Clipping algorithms . . . 119

6.1.2.3 Experiment C: Compression algorithms . . . 121

6.1.2.4 Experiment D: Noise reduction algorithms . . . 122

6.1.2.5 Counterbalancing . . . 123

6.1.3 Subjects . . . 123

6.2 Results . . . 124

6.3 Discussion . . . 125

6.3.1 Spatial listening . . . 125

6.3.1.1 Comparison with HINT norms . . . 127

6.3.1.2 Comparison with other SRM data . . . 127

6.3.2 Continuous vs. intermittent masking . . . 129

6.3.3 Effects of distortion . . . 130

6.3.3.1 Clipping algorithms . . . 130

6.3.3.2 Compression algorithms . . . 131

6.3.3.3 Noise reduction algorithms . . . 134

6.4 Conclusion . . . 137

7 Binaural Validation of the Proposed Model 139

Introduction . . . 139

7.1 Overview . . . 140
7.2 Results . . . 142
7.2.1 Validation and fitting without distortion . . . 142
7.2.1.1 Target SII reference . . . 142
7.2.1.2 Short-time window selection . . . 143
7.2.2 Distortion modeling . . . 144
7.2.2.1 Clipping algorithms . . . 145
7.2.2.2 Compression algorithms . . . 146
7.2.2.3 Noise reduction algorithms . . . 146
7.3 Discussion . . . 147
7.3.1 Summary . . . 147
7.3.1.1 Modeling spatial effects . . . 147
7.3.1.2 Modeling masker-type effects . . . 147
7.3.1.3 Modeling distortion effects . . . 148
7.3.2 Model refinements . . . 150
7.3.2.1 Stage 1: Distortion modeling . . . 150
7.3.2.2 Stage 2: Binaural hearing . . . 152
7.3.2.3 Stage 3: Intelligibility prediction . . . 153
7.4 Conclusion . . . 153

8 Conclusion **155**

Introduction . . . 155
8.1 Summary of contributions . . . 155
8.2 Thesis limitations and future works . . . 157
8.2.1 Distortion modeling . . . 157
8.2.2 Hearing-loss modeling . . . 158
8.2.3 Other modeling considerations . . . 159
8.3 Outlook . . . 160

A Analysis-Synthesis Filterbanks **161**

A.1 Gammatone analysis filterbank 161

A.1.1 30-band Gammatone analysis filterbank 162

A.1.2 21-band Gammatone analysis filterbank 162

A.1.3 Subband decomposition summary 165

A.2 Gammatone synthesis 165

B Experimental Results **168**

List of Figures

2.1 Anatomy of the human ear showing the outer ear, middle ear, and inner ear. (From [139], [148]) 10
2.2 Free-field to eardrum pressure transformation for different values of sound-source azimuth. 11
2.3 (a) Cross section of the cochlea showing the basilar membrane, Reissner's membrane and the organ of Corti. (b) Details of the organ of Corti. . . . 12
2.4 Absolute threshold of hearing as a function of frequency. 14
2.5 Equal-loudness contours for various loudness levels. 17
2.6 Illustration of the path of sound from a distance source. 19
2.7 Audiogram of a hypothetical hearing-impaired subject. 23
2.8 Shapes of the auditory filter centered at 1 kHz estimates for the normal and impaired ears. 26
2.9 Different types of hearing aids. 29

3.1 Plot of the analytic function in equation 3.5 fitted to subjective data for the SII with the HINT. 44
3.2 Examples of spatial configurations commonly used in the study of SRM. 54
3.3 Mean SRT obtained with frontal speech presentation as a function of noise azimuth. 56
3.4 Illustration of the EC model for a target-masker stimulus pair. 58
3.5 Structure of the EC model. 59

4.1 Diagram of the proposed three-stage binaural speech intelligibility measure. 65
4.2 Stage 1: Signal separation with distortion estimation. 66
4.3 EC-stage subband implementation according to [206]. 70
4.4 Comparing signals SSNOISE, SSINT16 and SSMOD01. 73
4.5 SII and ESII prediction using SSNOISE and SSMOD01 under continuous and intermittent masking (using rectangular windows). 77
4.6 SII and ESII prediction using SSNOISE and SSMOD01 under continuous and intermittent masking (using Hanning windows). 78
4.7 SII and ESII prediction using the standard and filterbank implementations. 79
4.8 Schematic diagram of the monaural hearing aid simulator. 81
4.9 Schematic diagram of the extended hearing device simulator. 82
4.10 Input/output function of the AGC amplifier in the hearing device simulator. 84
4.11 Illustration of peak and center clipping operations. 86
4.12 Graphical user interface for the Matlab speech testing environment. . . . 90

5.1 Recovered spectra obtained under additive noise conditions for a normal hearing subject with no hearing aid processing. 100
5.2 Recovered spectra obtained under linear and compressive processing for subject HI02. 102
5.3 Recovered spectra obtained under linear and peak clipping processing for subject HI02. 103
5.4 SII-weighted speech-to-distortion and noise-to-distortion ratios computed from the phase-inversion method for subject HI02 at soft, normal and loud speech levels. 105
5.5 SII predictions as a function of SNR for all combinations of HI subjects and hearing aids from tables 5.2 and 5.1 at normal speech level (65 dB SPL). 107

5.6 SII predictions showing the effects of nonlinear processing. Subject HI02 is simulated using soft (50 dB SPL), normal (65 dB SPL) and loud (80 dB SPL) speech presentation levels. 108
5.7 Effect of SII adjustment. 110

6.1 Subjective SRTs from Experiments A. 125
6.2 Subjective SRTs from Experiments B, C and D. 126
6.3 Comparison of SRM results with data from different studies (left panel). Wideband ILD for KEMAR-based HRTF as a function of noise azimuth (right panel). 128
6.4 Unprocessed speech, noise, and mixed signals, as well as signals processed through AGC algorithms. 133
6.5 Unprocessed speech, noise, and mixed signals, as well as signals processed through noise reduction algorithms. 136

7.1 Binaural validation strategy: comparing subjective measurements and objective predictions of speech intelligibility. 141
7.2 Measured and predicted SRTs for different spatial configurations in the absence of a hearing device. 143
7.3 Measured and predicted SRTs under continuous and intermittent masking in the absence of a hearing device. 144
7.4 Measured and predicted SRTs under peak and center clipping conditions. 145
7.5 Measured and predicted SRTs under compression and NRA processing conditions. 146

A.1 Impulse and magnitude response of a fourth-order gammatone filter centered at 1 kHz. 162
A.2 Magnitude response of the gammatone filterbank used in [12]. 163
A.3 Magnitude response of the 21-band gammatone filterbank. 164

A.4 PSD of the 30-band ERB-spaced filterbanks' output to white, pink and speech-shaped noise. 166
A.5 PSD of the 21-band critical-band filterbanks' output to white, pink and speech-shaped noise. 167

List of Tables

3.1 Example sentences from the HINT. 37
3.2 Two example word lists from the MRT. 38
3.3 Summary of objective measures evaluated in [197] 52

4.1 List of NRAs included in the extended hearing device simulator. 83

5.1 Hearing thresholds at standard audiometric frequencies and uncomfortable levels for the simulated subjects (in dB HL). 98
5.2 Hearing aid systems used during the evaluation. 99

6.1 List of simulated hearing devices. 117
6.2 Listening conditions tested in Experiment A. 119
6.3 Listening conditions tested in Experiment B. 120
6.4 Listening conditions tested in Experiment C. 121
6.5 Listening conditions tested in Experiment D. 122
6.6 Comparing SRT and SRM for baseline conditions across experiments with normative data for the HINT. 127
6.7 Comparing the projected percentage increase/decrease in speech intelligibility obtained with peak and center clipping. 130
6.8 Comparing the projected percentage increase/decrease in speech intelligibility obtained with AGCi and AGCo. 131

6.9 Comparing the projected percentage increase/decrease in speech intelligibility obtained with NRA01 and NRA02. 135

A.1 Summary of differences between the subband decompositions used for the EC/SII processing combination. 165

B.1 Subjective SRTs measured in Experiment A 168

B.2 Subjective SRTs measured in Experiments B, C and D 169

List of Acronyms

SNR_{env} envelope signal-to-noise ratio.

AGC Automatic Gain Control.

AGCi input-controlled AGC.

AGCo output-controlled AGC.

AI Articulation Index.

BILD binaural intelligibility level difference.

BKB Bamford-Kowal-Bench.

BKB-SIN BKB speech-in-noise.

BM basilar membrane.

BMLD binaural masking level difference.

BSII Binaural SII.

BSIM Binaural Speech Intelligibility Model.

BTE Behind-the-ear.

CB critical band.

CI confidence interval.

CIC Completely-in-the-canal.

CSII Coherence SII.

CVC consonant-vowel-consonant.

dB decibels.

dB HL hearing level.

dB SL sensation level.

dB SPL sound pressure level.

DSL Desired Sensation Level.

EC Equalization-Cancellation.

ERB equivalent rectangular bandwidth.

ESII Extended SII.

FB filterbank.

HASPI Hearing Aid Speech Perception Index.

HASQI Hearing Aid Speech Quality Index.

HASTB hearing-aid simulator toolbox.

HDSTB hearing-device simulator toolbox.

HI hearing-impaired.

HINT hearing in noise test.

HRTF head-related transfer function.

HT hearing threshold.

IBM ideal binary mask.

IHC inner hair cell.

ILD interaural level difference.

ITC In-the-canal.

ITD interaural time difference.

ITE In-the-ear.

ITFS ideal time-frequency segregation.

KEMAR Knowles Electronic Manikin for Acoustic Research.

MAE mean absolute error.

MAF minimum audible field.

MAP minimum audible pressure.

MCL most comfortable loudness.

MMSE minimum mean-square error.

MRT modified rhyme test.

MSC magnitude-squared coherence.

MTF modulation transfer function.

NAL National Acoustic Laboratory.

NH normal-hearing.

NRA Noise-reduction algorithm.

NzDR noise-to-distortion ratio.

OHC outer hair cell.

PSD power spectral density.

RMS root-mean-square.

RMSE root-mean-square error.

SD standard deviation.

SDR signal-to-distortion ratio.

SDT speech-detection threshold.

sEPSM speech-based envelope power spectrum model.

SII Speech Intelligibility Index.

SIN speech in noise.

SNR signal-to-noise ratio.

SpDR speech-to-distortion ratio.

SPIN speech perception in noise.

SRM spatial release from masking.

SRMR Speech-to-Reverberation Modulation Ratio.

SRT speech-reception threshold.

SSI synthetic-sentence identification.

STI Speech Transmission Index.

STOI short-time objective intelligibility.

STSA short-time spectral amplitude.

TFS temporal fine structure.

UCL uncomfortable loudness.

VAD voice activity detector.

WIN words-in-noise.

Chapter 1

Introduction

1.1 Research background

1.1.1 Motivation

Hearing impairment is the full or partial loss of the ability to detect, discriminate, or recognize sounds due to a physiological or functional abnormality of the ear. According to The Hearing Foundation of Canada, hearing impairment is the fastest growing and third leading chronic disability in Canada after arthritis and hypertension, affecting more than one million adults in Canada [199]. With the widespread use (or misuse) of iPods and MP3 devices, and especially in-the-canal earphones, hearing loss is no longer an aging problem, but an illness which affects the younger generation as well. In fact, a recent study in the US suggests that types of hearing loss, traditionally considered typical of a 50- or 60-year-old, now affect 20% of adolescents [179].

Today, hearing aids remain the primary technological solution to help individuals with hearing impairment. Hearing aids help improve the ability to understand speech in quiet and sometimes in the presence of background noise. However, quantifying the benefits of a hearing aid system is a difficult challenge. The ability to understand speech is affected by several factors including hearing loss, background noise, room reverberation, and

other distortions. Speech intelligibility is typically assessed using two complementary approaches: (1) subjective measurement of intelligibility scores based on listening tests, and (2) objective prediction of speech intelligibility using models. Subjective listening tests provide the highest level of validity but are time consuming and limited to a few conditions per subject due to time constraints and/or limited amount of normalized test material. In situations where large amounts of data on speech intelligibility are required to assess the potential benefits for different design strategies, such as during the hearing aid algorithm development process, objective models are indispensable.

Simple metrics like the global signal-to-noise ratio (SNR) provide a rough indication of the effect of hearing aids. Standard objective measures of speech intelligibility, such as the Articulation Index (AI, ANSI S3.5–1969 [3]), the Speech Intelligibility Index (SII, ANSI S3.5–1997 (R2012) [4]), or the Speech Transmission Index (STI, IEC 60268-16 [103]) are designed to work in several bands (octave, third-octave, or critical bands) to tackle the frequency-dependent characteristics of the human ear and background noise. These measures are derived for monaural listening situations in mostly linear processing conditions. In recent years, extensions of these measures have been developed to deal with nonlinear effects common in hearing aids or with binaural listening conditions. However, to date, no objective measure exists which deals with both conditions.

1.1.2 Research problem

The goal of this doctoral thesis is to address precisely this problem: developing an objective measure to predict speech intelligibility when using nonlinear binaural hearing devices. The challenge is to isolate the effects of nonlinear processing in the hearing device from the effects of binaural cues contained within the signals at each ear. Nonlinear distortions may result from hardware limitations (*e.g.*, transducers, amplifiers, and analog-to-digital or digital-to-analog converters) or signal processing strategies (*e.g.*, output limiting, compression, noise reduction algorithms) in the hearing aid. Such distortions have mixed effects on sound quality and clarity, and may reduce speech intel-

ligibility [198]. Conversely, binaural cues between the signals at the left and right ears (the interaural time and level differences) are known to improve speech intelligibility in the presence of noise under a variety of conditions.

On one hand, the coherence function has been traditionally used for evaluating nonlinear hearing devices in a monaural setting [47, 118, 167], and has become a standard procedure to estimate the speech and noise spectra at the output of such devices [2]. Furthermore, its use in speech intelligibility prediction has been shown to produce good results if the coherence is computed separately over three regions (low, mid and high levels) [124]. On the other hand, binaural objective measures have been proposed combining a binaural processing stage with monaural intelligibility prediction using the better-ear or binaurally-enhanced signal [12, 202]. Such measures have been developed and tested under linear processing conditions, and can also handle listening situations which include room reverberation. The coherence function, however, is not well suited for extending such binaural measures to handle nonlinear processing conditions commonly used in hearing aids. First, the underlying binaural model requires access to the temporal dimension of the separated speech and noise signals at the output of each device which can not be recovered with the coherence estimation. Furthermore, the inclusion of a three-level coherence approach, as suggested in [124], would require fundamental changes to the way binaural processing is modeled. Therefore, developing such a binaural extension requires an alternative signal-separation technique which does not suffer from these drawbacks.

1.1.3 Methodology and approach

In this doctoral thesis we propose a new objective measure based on the SII to predict speech intelligibility under binaural listening conditions, while accounting for nonlinear processing of signals in hearing devices. The proposed measure is reference-based and operates as follows:

- STAGE 1: Speech and noise separation performed based on the phase-inversion

technique proposed by Hagerman and Olofsson [87]. It is performed through a dual presentation of speech and noise mixtures, with the phase of the noise inverted in the second mixture. In order to account for nonlinear distortions inherent to the hearing device, a new distortion estimator derived from the measure of Olofsson and Hansen [154] is also included in this stage.

- STAGE 2: Binaural pre-processing using Equalization-Cancellation (EC) theory to model binaural release from masking [43]. The output of this stage is a monaural set of signals (speech and noise) which incorporate the added benefit of binaural processing. Several alternative implementations of EC processing (different analysis/synthesis filterbanks for frequency decomposition, Wiener filtering, and neural-network-based error modeling) are investigated.

- STAGE 3: Intelligibility prediction based on the signals at the output of the EC model using the standard SII (ANSI S3.5–1997 (R2012) [4]), or the short-time Extended SII (ESII [170]).

The measure receives the speech and noise signals at the input of the left and right hearing devices (four signals in total) as well as the listener's hearing thresholds, and outputs an intelligibility index that can be mapped to correlate with subjective intelligibility score. Validation of the model is carried out in two phases: monaural and binaural. This will be based on theoretical validation as well as measurements from subjective listening tests published in the literature or performed in-house at the Hearing Sciences Laboratory at the University of Ottawa. The reference-based model is applicable to a wide range of nonlinear devices including hearing aids, electronic hearing protectors and communication headsets. The use of hearing aids represents a particular subset of these devices to which we dedicate a significant portion of this thesis. In our validation efforts, the nonlinear processing systems will be addressed individually, regardless of which type of device they implement.

1.2 Contributions

The research contributions which will be discussed in the remainder of this thesis, along with the publications which have arisen to date, can be summarized as follows:

A three-stage model is proposed to predict speech intelligibility under binaural listening and nonlinear processing conditions. Earlier versions of the model have been discussed in the following publications:

- N.N. Ellaham, C. Giguère, and W. Gueaieb. Objective characterization of binaural speech intelligibility with hearing aids. In *Proc. Int. Hear. Aid Res. Conf.*, Lake Tahoe, CA, August 2010

- N.N. Ellaham, C. Giguère, and W. Gueaieb. Prediction of binaural speech intelligibility when using non-linear hearing aids. In *J. Acoust. Soc. Am.*, volume 123 of *Proc. 159th Meeting of the Acoustical Society of America*, page 1848, Baltimore, MD, April 2010

- N.N. Ellaham, C. Giguère, and W. Gueaieb. Binaural objective intelligibility measurement and hearing aids. In *Canadian Acoustics*, 37(3), pages 136–137, Niagara-on-the-lake, ON, October 2009

A novel method is derived to estimate speech and noise distortion introduced by nonlinear signal processing. These distortion estimators are integrated into the first stage of the model, and have been the subject of the following publications:

- N.N. Ellaham, C. Giguère, and W. Gueaieb. Evaluation of the phase-inversion signal separation method when using nonlinear hearing aids. *IEEE Trans. Audio, Speech, and Language Processing*, 21(4):879–888, April 2013

- N.N. Ellaham, C. Giguère, and W. Gueaieb. Evaluation of two speech and noise estimation methods for the assessment of nonlinear hearing aids. In *Proc. 36th*

IEEE International Conference on Acoustics, Speech and Signal Processing, Prague, Czech Republic, May 2011

Two optimizations of the EC process used in the second stage of the model have been implemented during the course of this doctoral work and will be briefly discussed in the thesis. The first uses Wiener filters to speed up EC processing, and has been presented at a workshop of the IEEE signal processing society. The second optimization uses neural networks to model the artificial EC processing errors.

- N.N. Ellaham, C. Giguère, and W. Gueaieb. A Wiener-based implementation of equalization-cancellation pre-processing for binaural speech intelligibility prediction. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October 2009

The Matlab speech testing environment developed during this work has proven to be a powerful tool with tremendous benefits to our research group, and has been the subject of a recent publication.

- N.N. Ellaham, C. Giguère, and W. Gueaieb. A new research environment for speech testing using hearing-device processing algorithms. In *Canadian Acoustics*, 42(3), pages 89–90, Winnipeg, MB, October 2014

Additional contributions, not yet published, will also be discussed in the thesis. The experiments carried out to collect subjective intelligibility scores (Chapter 6) have provided new original data on the interaction of spatial sound configurations, stationary and intermittent masking, and different nonlinear signal processing algorithms. Finally, the model's validation in a binaural experimental setting (Chapter 7) complements the central contribution of this thesis and lays the foundation for future research efforts.

1.3 Thesis outline

The remainder of this thesis is organized as follows. Chapter 2 provides a general overview of basic notions of psychoacoustics pertaining to hearing, hearing loss, and hearing aids, followed by a review of the relevant literature in Chapter 3. Chapter 4 gives an overview of the proposed objective binaural intelligibility model and provides details of the methodology employed to implement and validate the model. The model validation is planned in two steps: monaural and binaural. The monaural step is a simulation study which is presented in Chapter 5. The binaural step consists of collecting subjective data under different listening conditions, and comparing them to model predictions. We report on the experimental data collection in Chapter 6, and perform the model validation in Chapter 7. Finally, chapter 8 presents concluding remarks summarizing the contributions of this doctoral work, its limitations, and its future prospects.

Chapter 2

Psychoacoustic Principles

Introduction

This doctoral work is concerned with modeling the intelligibility of speech in binaural listening situations when using nonlinear hearing aids. A good grasp of certain fundamental notions of psychoacoustics is necessary in order to understand the modeling of speech intelligibility and the types of nonlinear processing performed in hearing devices. This chapter is dedicated to provide a brief overview of the basic principles and characteristics of hearing, hearing loss, and hearing aids. For a more in-depth discussion of these topics, the reader is referred to more comprehensive texts [148–150].

2.1 Overview of hearing

The ear is the organ responsible for the perception of sounds, an elaborate task which involves complex physical and biochemical processes. This section starts with a brief explanation of the physics of sound, which is necessary for a good understanding of hearing. A description of the structure and function of the ear and the auditory nervous system follows.

2.1.1 Sound generation and propagation

Sound can be thought of as mechanical vibrations traveling in a material medium, most commonly air. It is typically generated by a vibrating object, such as the human vocal chords or the membrane of an audio speaker. As the sound source vibrates it causes a displacement of the particles of the surrounding medium, which is the basis of sound propagation. In air, for example, it is the movement of air molecules that forms waves of alternating pressure, causing local regions of compression and rarefaction. As a particle is displaced from its equilibrium position, adjacent particles are also displaced, causing these local regions (and hence the sound wave) to travel through the medium.

2.1.2 Anatomy and function of the ear

Figure 2.1 shows the general anatomy of the peripheral auditory system, revealing three principal components: the outer ear, the middle ear, and the inner ear.

2.1.2.1 The outer ear

The outer ear consists of the externally visible part (auricle or pinna) and the auditory canal. It acts as a funnel that conducts incoming sound waves to the eardrum. Owing to the peculiar shape of the pinna, the outer ear alters the frequency content of the incoming waveform, especially at high frequencies. Figure 2.2 shows the sound pressure transformation that occurs in the outer ear, at four different sound-source azimuths as reported by Shaw in [180]. These curves illustrate how the outer ear plays an important role in sound localization.

2.1.2.2 The middle ear

The middle ear is an air-filled cavity containing the eardrum and a chain of three ossicles: the malleus, incus and stapes. These ossicles conduct sound vibrations from the eardrum to the the inner ear. Functionally, the middle ear serves at least three functions. First,

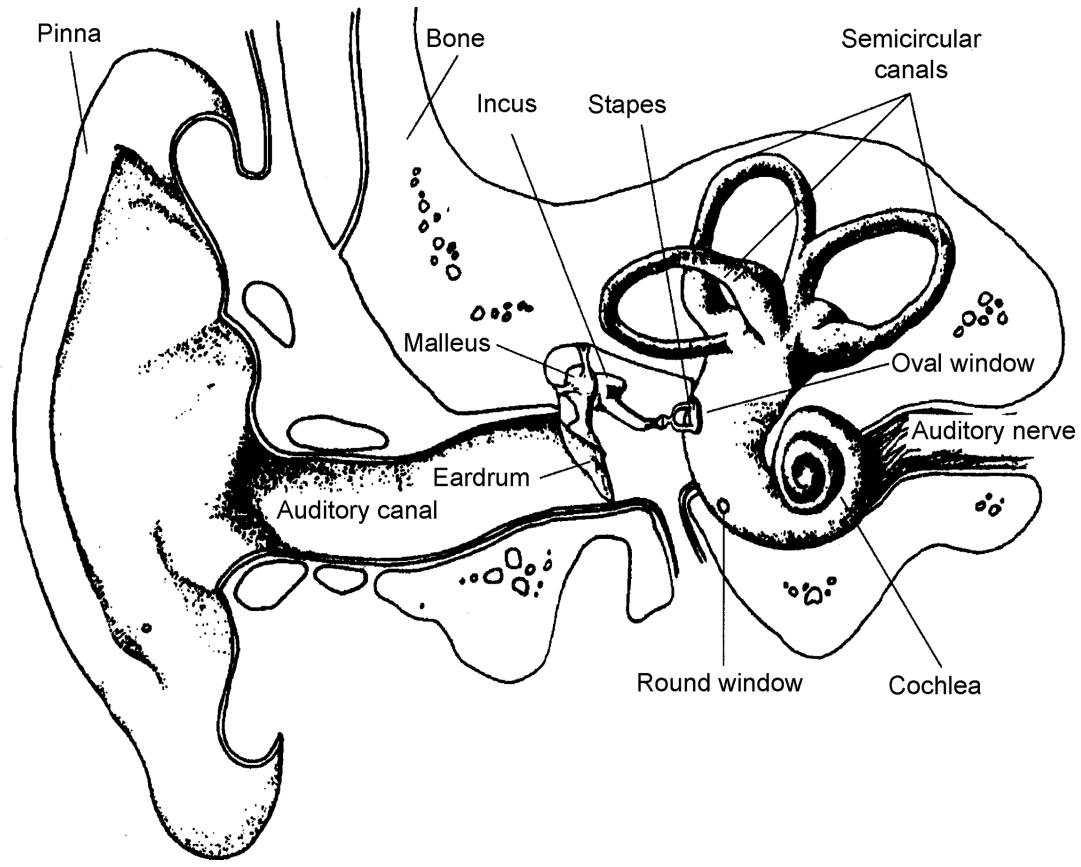


Figure 2.1: *Anatomy of the human ear showing the outer ear, middle ear, and inner ear. (From [139], [148])*

it acts as an acoustical impedance-matching mechanism in order to ensure an efficient transfer of mechanical vibrations from air (in the ear canal) to liquid (the fluid within the cochlea of the inner ear). Its second function is to reduce the transmission of bone-conducted sound to the cochlea in order to prevent the masking of incoming sounds by internal vibrations of the skull [7]. Its third function, known as the auditory reflex, is accomplished by small muscles attached to the ossicles which contract in response to intense sounds to reduce sound transmission through the middle ear.

2.1.2.3 The inner ear

The inner ear starts at the oval window and consists of the spiral-shaped cochlea and the vestibular apparatus (semicircular canals and vestibule). A cross section of the cochlea

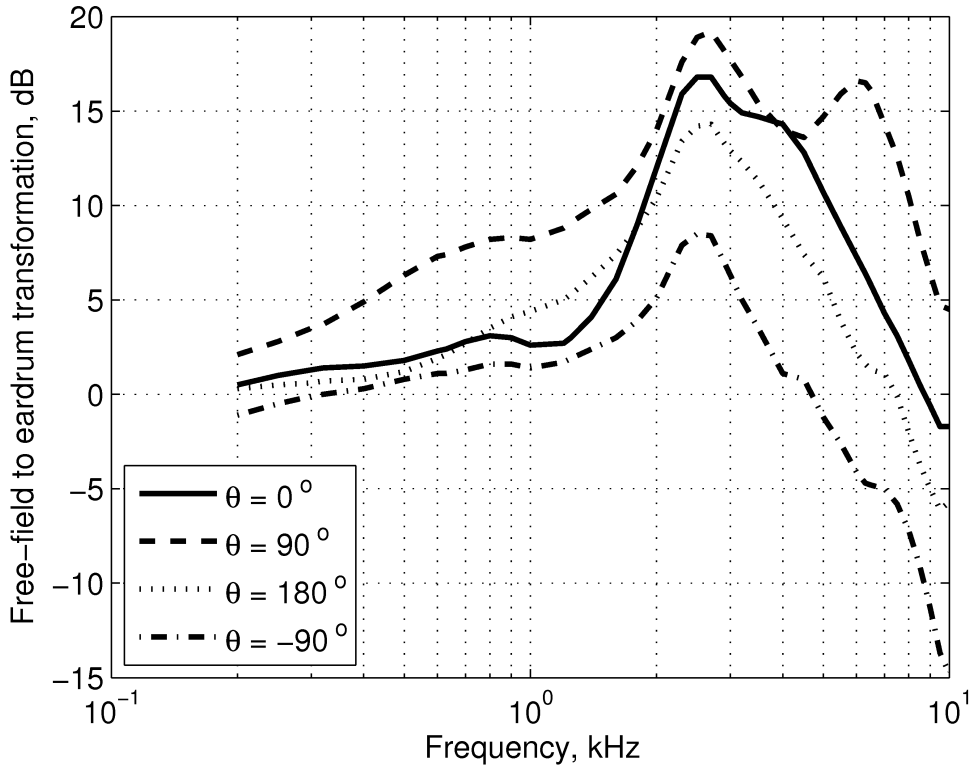
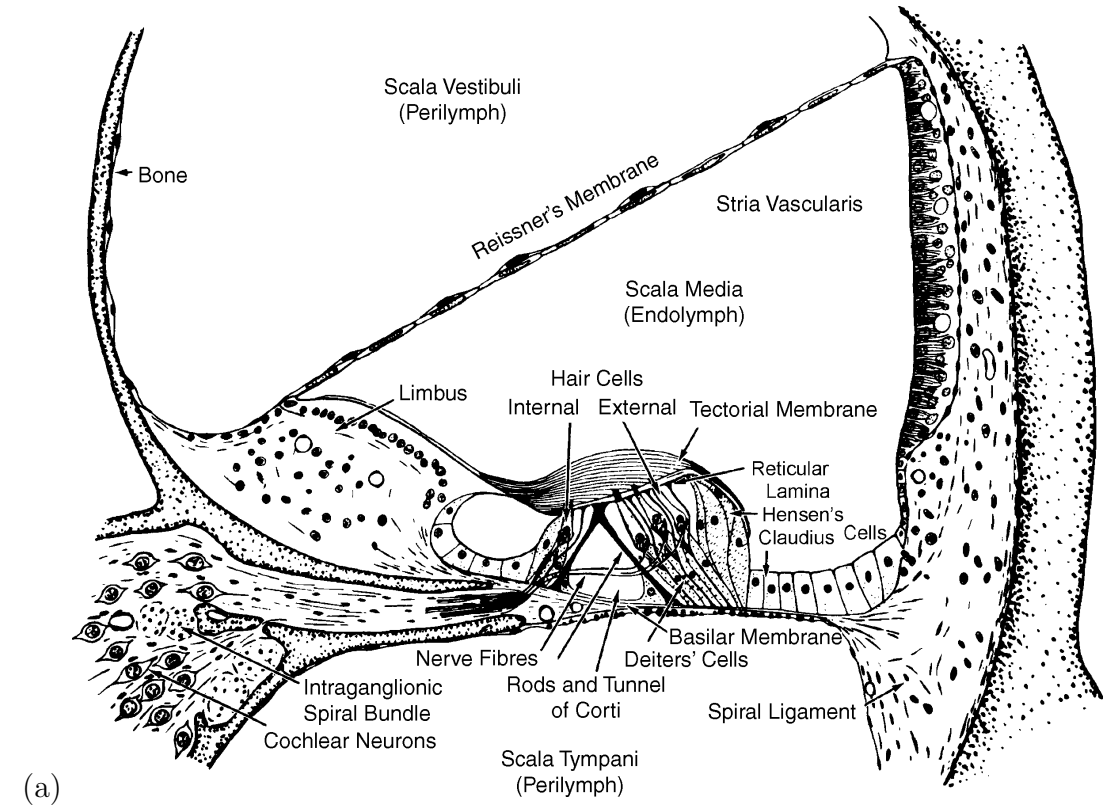


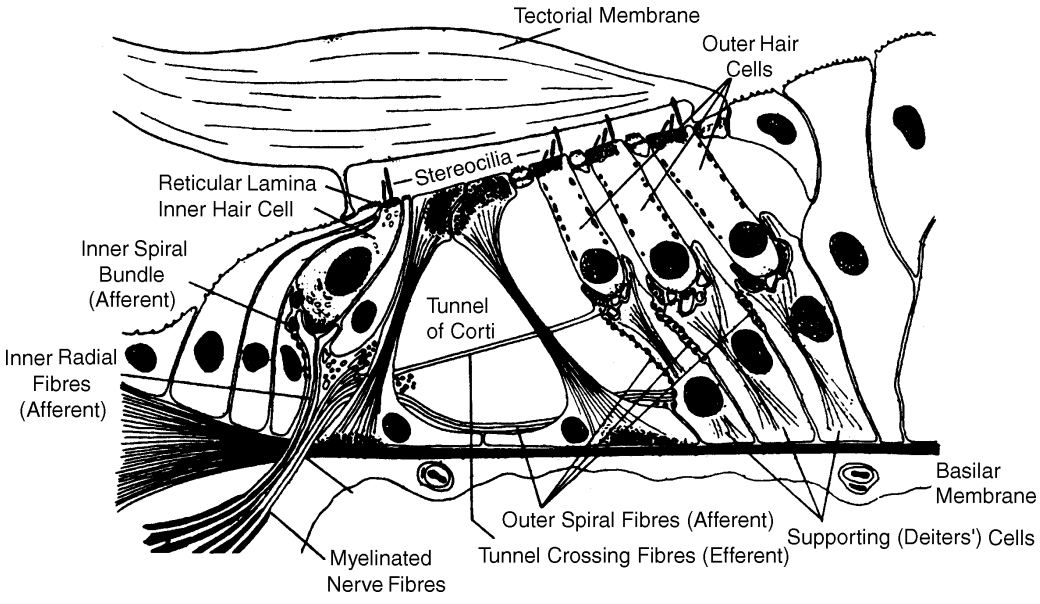
Figure 2.2: *Free-field to eardrum pressure transformation for different values of sound-source azimuth θ. These curves have been fitted to data covering 100 subjects from various studies over a period of 40 years (Data taken from [180])*

(Figure 2.3a) reveals the fluid-filled basilar membrane (BM) whose mechanical properties play an important role in sound perception: it is narrow and stiff towards the base of the cochlea and progressively becomes wider and less stiff from base to apex. As a result, the BM responds differently along its length to different sound frequencies: at the base it is more sensitive to higher frequencies, while lower frequencies produce a maximal response towards the apex.

Running parallel to the BM along the length of the cochlea, is the tectorial membrane. Enclosed between these two membranes are hair cells, which form part of a structure known as the organ of Corti (Figure 2.3b). The organ of Corti is the site of sound transduction in the inner ear: In response to the movement of the footplate of the stapes through the oval window, the BM is set in motion, which causes the movement



(a)



(b)

Figure 2.3: *(a) Cross section of the cochlea showing the basilar membrane, Reissner's membrane and the organ of Corti. (From [37], [148]). (b) Details of the organ of Corti. (From [175], [148])*

of stereocilia at the tip of the hair cells. The motion of these mechanoreceptors is accompanied by a complex mechano-chemical process which generates electrical signals in the neurons forming the cochlear nerve. These signals, also called action potentials, travel along the cochlear nerve to higher centers in the nervous system.

2.1.3 The auditory nervous system

It is beyond the scope of this thesis to provide but the briefest description of some aspects of auditory processing in the (very complex) higher centers of the brain. Along the length of the cochlea, auditory nerve fibers synapse with the hair cells of the inner ear. *Afferent* never fibers transmit sound signals from the inner hair cells (IHCs) towards the brainstem. *Efferent* never fibers from the brainstem synapse with the outer hair cells (OHCs) and are part of an active mechanism which regulates the functioning of the cochlea. This active mechanism depends on the integrity of the OHCs and is thought to be responsible for the high sensitivity and sharp frequency tuning of the BM, and the nonlinear input-output function of the BM in response to different sound levels.

Finally, fibers from both ears often converge on the same neuron in the brainstem. These interneurons are influenced by differences in intensities and arrival times of the input from both ears. These differences, along with cues from the pinna and head movement, are used to determine the location of a sound source. This will be discussed further in Section 2.2.5.

2.2 Characteristics of normal hearing

We now develop an understanding of what characterizes "normal" hearing. This section introduces the notion of hearing thresholds (HTs) and the approach of viewing the ear as an auditory filterbank. This is followed by a discussion of loudness perception, temporal processing in the auditory system, and binaural hearing.

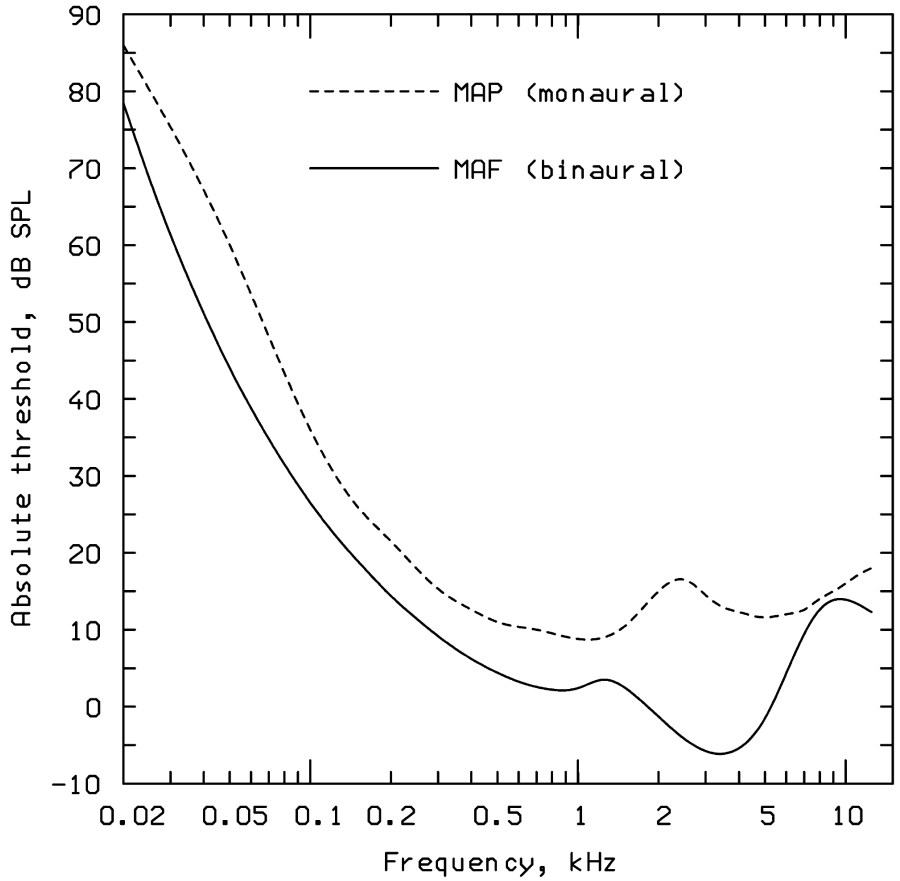


Figure 2.4: *Absolute threshold of hearing as a function of frequency. MAP-threshold data taken from [128]. MAF-thresholds taken from ISO 389–7, 1996 [106]. (From [150])*

2.2.1 Hearing thresholds

The human ear is sensitive to sounds in the range between 20 Hz and 20 kHz, and is most sensitive to the middle frequencies (1000–5000 Hz). The absolute threshold of hearing is the minimum level for a sound stimulus to be detected 50% of the time in the absence of other sounds. Measurements of absolute thresholds vary with the measurement procedure and stimulus length. Figure 2.4 shows estimates of the absolute thresholds measured for normal-hearing listeners using long-duration tones (greater than 200 ms). The plots show the average thresholds for many young listeners. An individual "normal-hearing" listener may have thresholds 20 decibels (dB) above or below the mean

at any given frequency. The data for the minimum audible pressure (MAP) is obtained from monaural measurements using a probe microphone placed as close as possible to the listener's eardrum, with the sound usually delivered through headphones. The data for the minimum audible field (MAF) represents a binaural listening condition with the sound coming from the front, and is measured using a microphone at the center of the listener's head (in the absence of the listener), with the sound delivered through loudspeakers. Binaural thresholds are, on average, roughly 2 dB lower than those measured monaurally. Note that the two curves differ at high frequencies, due to the influence of the head, pinna, and auditory canal on incoming sounds.

2.2.2 Auditory filters and frequency selectivity

The peripheral auditory system behaves as if it contains a bank of bandpass filters with overlapping passbands, known as the auditory filters [69]. This behaviour is primarily due to the mechanical properties of the BM, since each point on the membrane responds maximally to a limited range of frequencies. At moderate sound levels, the shape of auditory filters is symmetric and consists of a rounded top with steep sloping edges. At high levels, the shape of the filters becomes broader especially at the low-frequency side.

The auditory filters characterize frequency selectivity at different center frequencies. The effective bandwidth of these filters (known as the critical band (CB)) increases from low to high center frequencies. A common measure of the CB is the equivalent rectangular bandwidth (ERB), which is defined as the bandwidth of a rectangular filter which has the same peak transmission as the auditory filter and passes the same total power for a white-noise input [150]. For normal-hearing subjects this value is denoted ERB_N, and can be calculated as a function of center frequency f, as follows [81]:

$$ERB_N = 24.7(4.37f + 1) \tag{2.1}$$

where the value of ERB_N is in Hz, while f is specified in kHz.

The CB concept is closely related to auditory masking: when listening to a target

signal in the presence of background noise (the masker), the auditory filter centered at the frequency of the target removes the noise components which lie outside of the CB. The threshold of the signal is then determined by the signal-to-masker ratio at the output of the auditory filter or an adjacent filter with the best signal-to-masker ratio. Therefore the larger the CB, the more the signal is masked. For complex signals, the outputs from several filters are used in sound perception.

2.2.3 Loudness perception

The human auditory system is said to have a wide dynamic range in that it responds to a remarkable range of sound intensities defined on the logarithmic dB scale: the most intense sound we can hear without damaging our ears is about 120 dB above (*i.e.*, 10^{12} times more than) the faintest sound we can detect. More remarkable is that, within this range, we can detect small changes in sound level of about 0.3–2 dB. It is no wonder that much attention has been dedicated to the study of the perception of loudness.

Loudness is a subjective attribute of the intensity of an auditory sensation. The loudness level of a sound, defined in units of phons, is the sound pressure level of a 1000-Hz pure tone (in dB) which a listener judges to be equivalent in loudness to the given sound. The equal-loudness contours have a shape that is similar to the absolute thresholds curve, and tend to become flatter at high loudness levels (*c.f.* Figure 2.5). The rate of growth of loudness levels with increasing level is not the same across all frequencies. For example, for low frequencies the rate of growth is greater than for middle frequencies.

The wide dynamic range of the auditory system, and its performance in intensity discrimination (the detection of changes in intensity) seem to involve several mechanisms related to the excitation of neurons in the auditory nerve. These mechanisms include: (1) changes in firing rates, (2) the spread of excitations over center frequencies above and below the frequency of stimulation, and (3) phase locking, that is the tendency of nerve firings to occur in phase with the stimulating waveform on the BM. Moreover, intensity

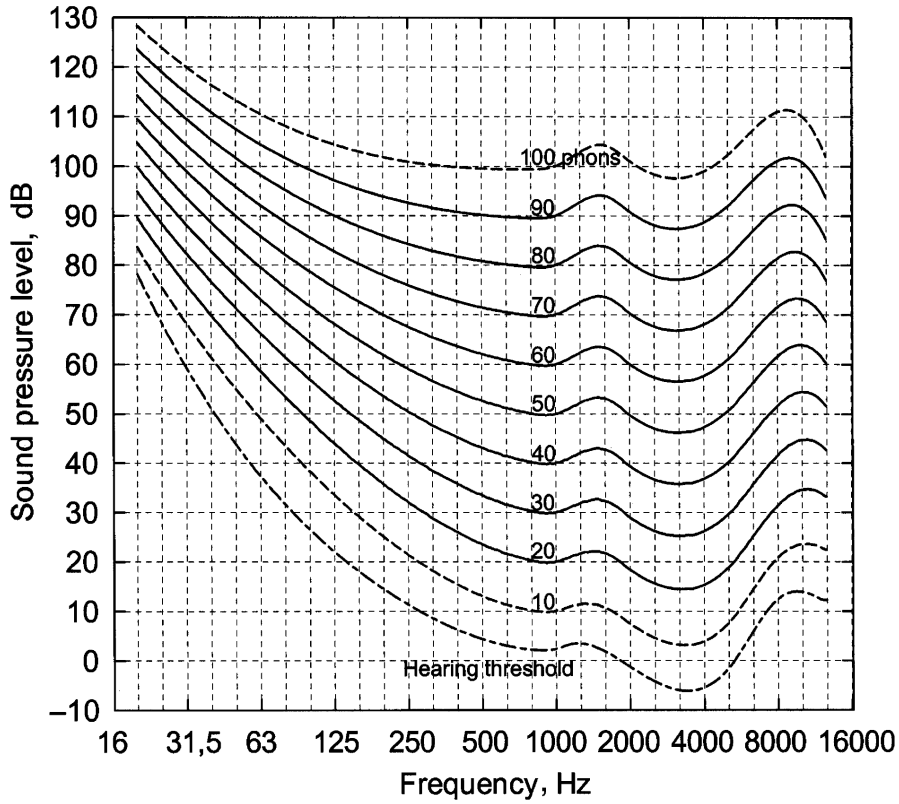


Figure 2.5: *Equal-loudness contours for various loudness levels. The bottom curve is the absolute threshold (MAF) curve. For all these curves, sounds were presented binaurally in free field, with the sound coming from the frontal direction. (From [105]. Permission to reproduce Figure A.1 from ISO 226:2003 was provided by Standards Council of Canada. No further reproduction is permitted without prior written approval from Standards Council of Canada.)*

discrimination seems to be limited, not by the information on the auditory nerve, but by the use of this information by higher centers in the brain.

2.2.4 Temporal processing

The previous sections were concerned with the long-term frequency representation of an incoming waveform. Most sounds, however, are time-varying (*e.g.*, speech, music), and much of the information is carried out within the temporal fluctuations of the signal. In considering the temporal dimension of sound, one must distinguish between the rapid fluctuations in sound pressure, known as the "fine structure", and the overall outline

of these amplitude fluctuations, known as the "envelope". Temporal resolution, the ability to detect changes in sound over time, is normally concerned with changes in the signal envelope. Several indicators of temporal resolution are studied in this field such as detecting a gap in a sound signal, detecting tones in temporally modulated noise, release from masking in fluctuating noise, forward masking, and temporal integration [169].

2.2.5 Binaural hearing

Binaural hearing refers to our ability to hear with two ears. The signals arriving at the left and right ears contain cues (binaural cues) which play an important role in our ability to locate the direction of sound sources in the horizontal plane. For example, when the sound source is frontal the signals at the two ears are almost identical and the binaural cues are negligible. When the sound source is located to one side of the head (*e.g.*, to the right), the sound reaching the farther ear (the left ear) is slightly delayed in time, and is less intense than the sound reaching the closer ear (the right ear). The corresponding binaural cues are commonly termed as the interaural time difference (ITD) and interaural level difference (ILD). ITD cues are reflective of the path difference between the two ears as illustrated in Figure 2.6 for a distant sound source. ILD cues are primarily due to a "head shadow" effect that is somewhat similar to the shadow produced by a beam of light. ITD and ILD cues are not equally important at all frequencies: owing to the physical nature of sound, ITDs are more effective at low frequencies, while ILDs are more useful at high frequencies.

Therefore, ITDs and ILDs are related to the location of a sound source in space relative to the listener. In an experimental setting, interaural differences can be introduced by physically moving a sound source in space or, when performing measurements with earphones, by using head-related transfer functions (HRTFs) measured from real subjects or on a human head-and-torso simulator, such as the Knowles Electronic Manikin for Acoustic Research (KEMAR) [25]. As we will see later in this thesis in Section 3.3.1, these binaural cues play an instrumental role in the auditory system's ability to improve

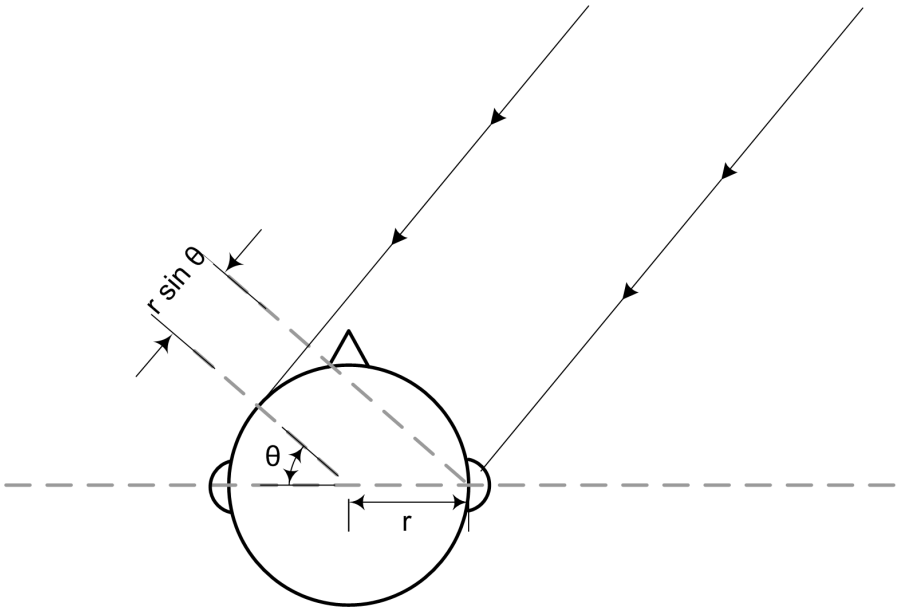


Figure 2.6: *Illustration of the path of sound from a distance source at azimuth θ (in radians). The ITDs can be geometrically related to the path difference $d = r\theta + r\sin\theta$.*

the detection of sounds in the presence of background noise. Finally, we introduce a conventional notation commonly used to code experimental stimulus conditions in different spatial configurations: the symbols S and N are used to designate the speech and noise signals, each followed by a suffix to denote the relative phase between the two ears:

- S_0N_0 means that both speech and noise are in phase at the two ears,
- S_0N_π means the speech is in phase, but the noise is phase-inverted,
- S_m or N_m indicates a monaural presentation of speech or noise respectively,
- N_d indicates a diffuse noise presentation,
- N_u indicates that the noises presented at both ears are uncorrelated,

The use of this notation will be frequently employed in this thesis.

2.3 Hearing loss

This section is dedicated to defining basic hearing loss concepts, and discussing its perceptual consequences. We start with a general overview of the characteristics of hearing loss, followed by a synopsis of hearing assessment procedures. The section concludes with a discussion of those attributes of cochlear hearing loss known as supra-threshold deficits, namely loss of frequency selectivity, reduced dynamic range, and loss of temporal resolution.

2.3.1 General characteristics of hearing loss

2.3.1.1 Types of hearing loss

Three types of hearing loss can be distinguished reflecting the point in the auditory system where the impairment occurs. When the loss is due to a reduced transmission of sounds through the outer or middle ear, it is known as *conductive hearing loss*. This can be caused by a deposit of cerumen (wax) in the auditory canal, damage to the eardrum or the middle-ear ossicular chain, or the presence of fluid in the middle-ear cavity as a result of an infection (otitis media). When the impairment is caused by damaged to the structures of the cochlea, it is termed as *cochlear hearing loss*. This could be due to prolonged exposure to loud sound levels, genetic factors, infections, allergies, metabolic disturbances, or as a side effect of some ototoxic drugs. When the loss occurs as a result of damage to neural structures (in the auditory nerve or the auditory cortex), we speak of *retrocochlear hearing loss*. This type of hearing impairment is most often caused by a benign tumor that exerts pressure on the auditory nerve. Moreover, the term *sensorineural hearing loss* is used to refer to an impairment that affects either cochlear or neural structures. It is also common to speak of a *mixed hearing loss* as one that has conductive and sensorineural components [149].

Perceptually, conductive hearing loss is characterized by a (frequency-weighted) at-

tenuation of incoming sounds. This is well predicted by the elevation in the subject's pure-tone thresholds (*c.f.* Section 2.3.2.1). This type of hearing loss is usually treated with drugs (to cure an infection) or surgery. The effects of sensorineural hearing loss go beyond the ability to detect weak sounds. A common complaint among subjects with this type of loss is the reduced ability to understand speech, especially in the presence of background noise. For many subjects, even when sounds are amplified above their detection thresholds, the perceived sound is unnatural, muffled, or distorted. Most sensorineural hearing losses cannot be treated with surgery, and require hearing aids running sophisticated algorithms to alleviate these effects.

2.3.1.2 Severity and configuration

Aside from determining the type of hearing loss, an audiologist's assessment aims to evaluate certain quantitative and qualitative attributes of the loss. The severity of the loss is determined based on the subject's pure-tone audiogram (*c.f.* Section 2.3.2.1). Using the average of the pure-tone thresholds between 500 and 4000 Hz specified in hearing level (dB HL), the following classes of hearing loss severity are defined:

- Normal hearing: between 0 and 20 dB HL,
- Mild hearing loss: between 20 and 40 dB HL,
- Moderate hearing loss: between 40 and 60 dB HL,
- Severe hearing loss: between 60 and 80 dB HL, and
- Profound hearing loss: above 80 dB HL.

Qualitatively, the audiologist may describe a hearing loss in terms of whether it is unilateral or bilateral, and in the latter case whether it is symmetrical or asymmetrical. The audiogram may also show whether the loss is flat at all frequencies, sloping or precipitous, or in general terms if it is predominantly a high-frequency versus low-frequency

loss. Finally, when monitoring the hearing-impaired subject, an audiologist determines whether the loss is sudden or progressive, and whether it is stable or fluctuating (getting better or worse) over time.

2.3.2 Assessment of hearing loss

2.3.2.1 Pure-tone audiometry

A clinical hearing assessment starts with a physical examination of the patient's outer ear (checking for malformation in the pinna or blockage of the ear canal) and eardrum (examining for perforations or signs of infection or the presence of fluid in the middle-ear cavity). Following the physical examination, the first audiological test is performed using *pure-tone audiometry* to determine the type, severity, and configuration of the patient's hearing loss. With the subject sitting comfortably in a sound-proof booth, pure tone signals are presented (through headphones, earphones or a bone-conduction oscillator) in order to determine the softest tones which the patient is able to detect (the pure-tone thresholds). Although humans are sensitive to a much larger frequency range, pure tones in the range 250–8000 Hz are used, as this represents most of the speech spectrum. The "audiogram" is the resulting chart of hearing sensitivity representing the pure-tone thresholds specified in units of dB HL as a function of frequency. An example of an audiogram depicting a high-frequency sloping sensorineural hearing loss in the right ear is shown in Figure 2.7. In this case, a pure-tone threshold of 50 dB HL at 2000 Hz means that this subject's threshold is 50 dB higher than the reference otologically "normal" population at that frequency. The average pure-tone thresholds for a normal subject, which corresponds to the MAP thresholds curve (*c.f.* Section 2.2.1), is always shown on the audiogram as a straight line at 0 dB HL.

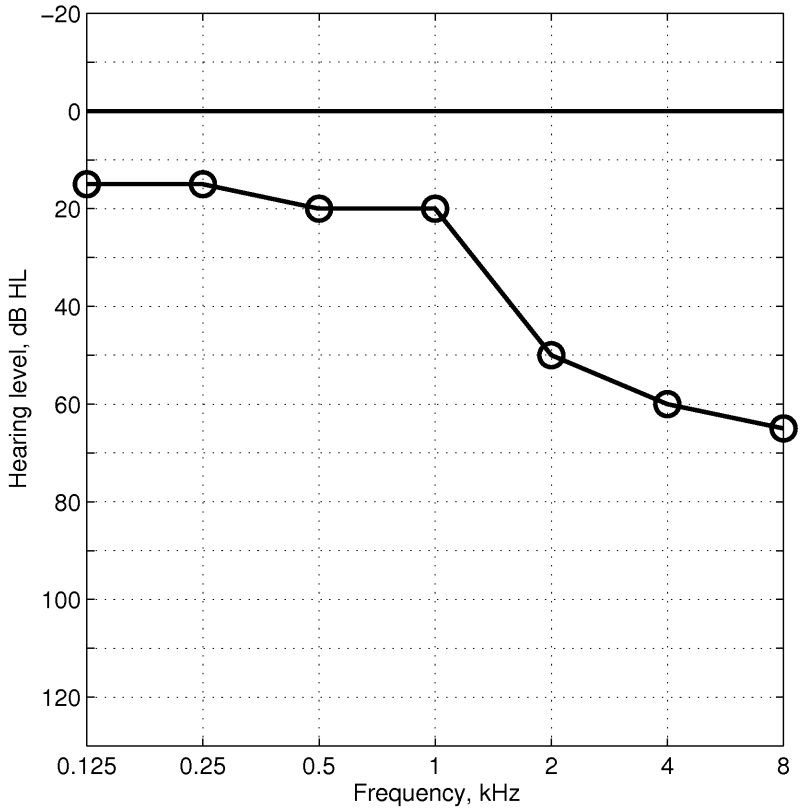


Figure 2.7: *Audiogram of a hypothetical hearing-impaired subject (right ear).*

2.3.2.2 Speech audiometry

The audiogram offers only a limited assessment of a subject's hearing. A wide variety of tests, which use speech as sound material, is available to audiologists to evaluate a patient's ability to understand speech as it is presented in everyday situations. These tests fall under the term *speech audiometry*, and are used to determine one of two measures: the speech-detection threshold (SDT), that is the lowest level at which the presence of speech can be detected a specified percent of the time (usually 50%); and the speech-reception threshold (SRT), that is the lowest level at which speech can be correctly identified - for example by repeating words - a specified percent of the time (usually 50%). The SRT in quiet is particularily interesting as it is highly correlated with

the pure-tone thresholds between 500 and 2000 Hz and can, thus, be used to validate a patient's audiogram. Speech tests can also be carried in the presence of noise. A review of speech audiometry tests will be presented later in this thesis in Section 3.1.1.

2.3.2.3 Most comfortable and uncomfortable levels

Two important measures, which are determined during a hearing assessment using speech as sound input, are the most comfortable loudness (MCL) and uncomfortable loudness (UCL) levels. The MCL is obtained by averaging two measurements using a bracketing technique: first the speech sound is presented at a level just above the SRT and the intensity is increased until the subject indicates that the sound is most comfortably loud; the second iteration starts at a level above the MCL found in the initial step, then the intensity level is decreased until the subject indicates having reached the MCL. The MCL is useful in determining the ideal amplification for a hearing aid candidate. During the UCL measurement, subjects are instructed to indicate when the speech level becomes uncomfortably loud. The UCL is important for determining a subject's dynamic range for speech, and the maximum tolerable amplification.

2.3.2.4 Other audiologic tests

In addition to pure-tone and speech audiometry, a number of audiologic tests can be used to complement the hearing assessment. Acoustic immittance tests help diagnose abnormalities in the ear canal (static acoustic impedance), eardrum mobility (tympanometry) and middle-ear muscle reflex (acoustic reflex). Otoacoustic emissions (OAE) can be used to diagnose damage of cochlear origin. These are inaudible sounds emitted by the normal cochlea when stimulated by sound, and can be measured using a small probe inserted into the auditory canal. Finally, tests based on auditory evoked potentials, such as the auditory brainstem response (ABR), are used to monitor brain activity in response to a sound stimulus. Such tests are often used with infants or adults that may be suffering from a neurological disorder.

2.3.3 Supra-threshold deficits

2.3.3.1 Loss of frequency selectivity

Auditory filters and frequency selectivity in normal-hearing subjects have been discussed in Section 2.2.2. The fine frequency resolution of the normal auditory system depends on an active mechanism mediated by the OHCs on the cochlea (*c.f.* Section 2.1.3). For subjects with damage to the OHCs (a common form of cochlear hearing impairment) the auditory filters tend to become broader than for normal-hearing subjects (*c.f.* Figure 2.8). Perceptually, the broadening of the auditory filters results in (1) increased masking of signals by interfering noise (*c.f.* Section 2.2.2) which causes difficulty in following a conversation in noisy situations, and (2) a reduced ability to identify the spectral composition of complex sounds (*e.g.*, speech, music) resulting in difficulty in distinguishing between different vowel sounds or musical instruments.

2.3.3.2 Reduced dynamic range

Hearing impairment of cochlear origin causes a subject's absolute thresholds to increase but the level at which sounds become uncomfortably loud (the UCL) often remains in the normal or near-normal range. *Loudness recruitment* is the mechanism that is responsible for this reduced dynamic range. It is defined as an unusually rapid growth in loudness level with an increase in stimulus level. Loudness recruitment can also be explained as a steepening of the input-output function of the BM, a reduction of compressive nonlinearity produced by OHC damage. Perceptually, the consequences of loudness recruitment are manifested as a reduced ability to hear very faint sounds, while high-level sounds are just as loud as for normal listeners. Moreover comfortable speech is not as easily intelligible.

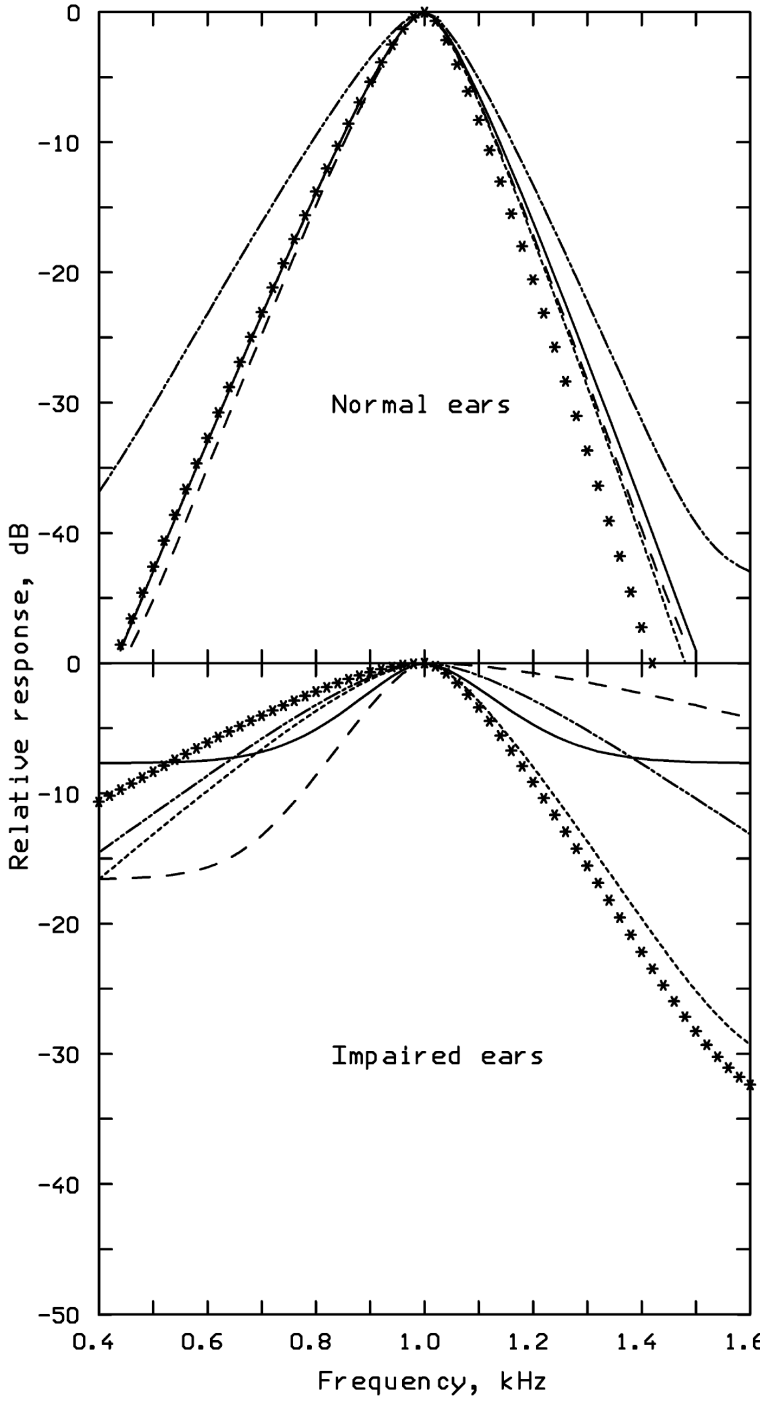


Figure 2.8: *Shapes of the auditory filter centered at 1 kHz estimates for the normal (TOP) and impaired (BOTTOM) ears of five subjects with unilateral cochlear hearing loss. (From [80])*

2.3.3.3 Loss of temporal resolution

Reed et al. provided a comprehensive review of temporal resolution for listeners with cochlear hearing loss [169]. Studies show that certain temporal processing abilities are degraded in hearing-impaired individuals compared to normal-hearing listeners at equal sound pressure level (dB SPL) (*e.g.*, gap detection, and release from masking in temporally fluctuating noise), or at equal sensation level (dB SL) (*e.g.*, forward masking). When isolating the effects of reduced audibility from supra-threshold deficits, performance for both effects was found to match for some tasks (*e.g.*, gap detection, gap-duration discrimination, and detection of tones in temporally modulated noise), while subjects with supra-threshold deficits under-performed in the task of temporal integration [169].

2.4 Hearing aids

"The purpose of a hearing aid is to amplify sound to overcome a hearing loss" [120]. This, however, is no simple task, owing to the complicated nature of the deficits associated with hearing loss as seen in the previous section. In this section we present a brief overview of developments in hearing aid technology and the most common types of hearing aids available in today's market. The section ends with a discussion of processing strategies and artificial intelligence typically used in digital hearing aids [120, 207].

2.4.1 Historical evolution

The evolution in hearing aid technology can be classified in five generations. The first is the acoustic (*i.e.*, mechanical) generation using ear trumpets and horns of different shapes and sizes to amplify sounds, in a manner similar to the cupped hand. The second generation initiated at the beginning of the twentieth century when elements of telephone technology (such as the use of carbon microphones) were integrated into hearing aids, adding to the amplification provided by acoustical resonances. In the third generation,

greater acoustical amplification was achieved by means of vacuum tubes. The 1940s witnessed the invention of the transistor which replaced vacuum tubes and made possible hearing aid designs small enough to be worn on the head. Finally, owing to advances in integrated-circuit technology in the 1990s and its use in electronic amplifiers, the fifth (and current) generation of hearing instruments has witnessed the gradual replacement of many analog components by digital functionality. Thanks to scientific leaps in digital signal processing, artificial intelligence, wireless communications, and binaural hearing technology, today's hearing aids are context-aware devices with highly efficient noise reduction, feedback cancellation, and directional processing algorithms, which are also capable of sensing and adapting to varying acoustic environments and learning a user's preferred settings.

2.4.2 Types of hearing aids

Advances in hearing aid technology have led to an ongoing process of miniaturization of these devices since the generation of trumpets and horns. Today's hearing aids come in a variety of different styles and sizes. While most users would rather choose the smallest (least visible) device, the choice of hearing aid type is constrained by other factors such as the amount of amplification needed, and manual dexterity. Figure 2.9 shows the most common styles of hearing aids on the market today. We present below a brief overview of the basic types in order of size (from largest to smallest).

- **Behind-the-ear (BTE)** aids house the microphone, amplifier circuitry, receiver (*i.e.*, speaker), and filtering controls in a small case that sits behind the pinna. The more recent BTE hearing aid cases are small enough that they are relatively invisible from the front of the aid wearer. The output sound is conducted from the receiver through a transparent plastic tube supported in the ear canal by a custom earmold that fits the wearer's ear. Such aids are appropriate for mild to profound hearing losses. BTE aids do not benefit from pinna cues, but the larger casing

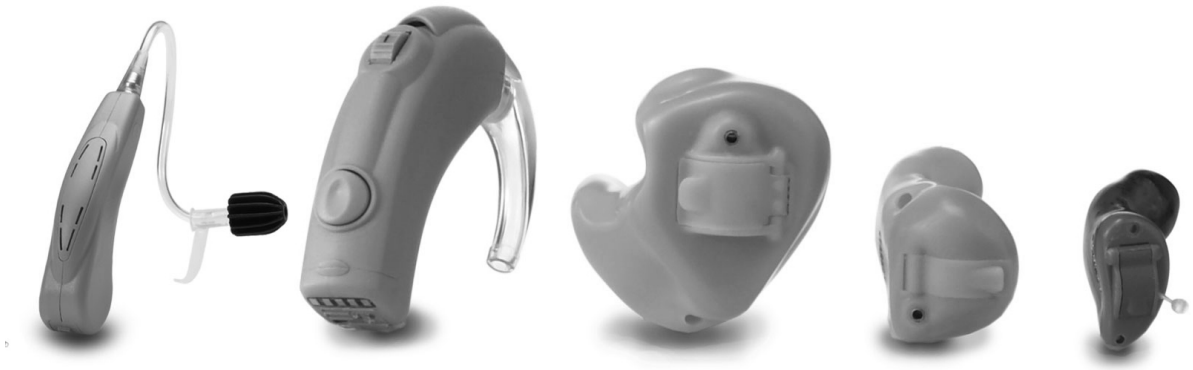


Figure 2.9: *Different types of hearing aids. (Courtesy of Hear & Speak Clinic)*

leaves plenty of room for a higher-quality battery and receiver, and the possibility of using directional microphones that assist in noise reduction.

- **In-the-ear (ITE)** aids fit completely in the outer ear and are generally designed for mild to severe hearing losses. These types of hearing aids come in two designs: (1) a custom design in which all electronic components are housed in a plastic shell produced from an impression of the ear, and (2) a modular design which consists of a standard housing attached to a custom earmold, with the receiver installed into the earmold. The controls and microphone of an ITE hearing aid are located in the concha, thereby taking advantage of pinna cues. The receiver is placed as deep as possible into the ear canal, which makes it susceptible to clogging (by earwax) or damage (from ear drainage).

- **In-the-canal (ITC)** aids are very small and are custom built to fit inside the ear canal, the outer surface of the aid protruding only slightly. The small size of these devices makes them not suitable for smaller ears. It also imposes restrictions on component size (*e.g.*, battery, receiver), which limits its maximum output power, making them appropriate for patients with mild to moderately-severe hearing losses. A particular type of ITC hearings aids is the Completely-in-the-canal (CIC) aid. These aids are made very small to fit deep into the ear canal thereby making them completely invisible. They are, however, not easy to handle and

require a small transparent cable to be attached to the housing to permit removal.

In addition, body hearing instruments, the oldest form of electronic devices, continue to be in use although their market share is progressively shrinking. These devices house the electronics in a case worn on the user's body (*e.g.*, on a belt, or in a pocket) and deliver the aided sound through an earphone inserted in the outer ear.

2.4.3 Digital hearing aid technologies

2.4.3.1 Amplification and signal enhancement strategies

Today's digital hearing aids are programmable devices that make use of several different strategies to improve the quality of sounds reaching the user's ear. Aside from a volume control to apply an overal gain to boost the sound level, most hearing aids today are equipped with many of the common features listed below:

- **Prescription fitting**: There are different methods for predicting the target gain required to correct for a specific hearing loss. These are known as prescription fitting formulas, and are based on audiometric data from hundreds of individuals with a particular hearing loss. Several linear and nonlinear formulas are available, such as the prescription developed at the National Acoustic Laboratory (NAL) in Australia, or the Desired Sensation Level (DSL) method developed at the National Centre for Audiology at the University of Western Ontario.

- **Automatic Gain Control (AGC)**: AGC amplifiers function as linear amplifiers when sounds are below a specified signal level (the kneepoint), and reduce amplification (compression) when sounds reach or exceed that level. Two classes of AGC amplifiers can be differentiated:

  - **Input-controlled AGC (AGCi)**: The AGC circuitry is placed before the volume control, such that compression is activated by the level of the input

sound. An AGCi may be used to restore normal loudness perception in subjects with reduced dynamic range due to cochlear hearing impairment (*c.f.* Section 2.3.3.2).

  - **Output-controlled AGC (AGCo)**: The AGC circuitry is placed after the volume control such that compression is activated by the level of the output sound. Such amplifiers are largely used for output limiting.

- **Multi-channel processing**: Most hearing aids process signals by splitting them into a number of frequency bands before applying linear or nonlinear amplification. It is common, for example, for AGC amplifiers to perform multi-channel compression with adjustable parameters in each band.

- **Microphone arrays**: Today's hearing aids may use directional microphones or microphone arrays instead of a single omni-directional microphone. Beamforming refers to a signal processing technique in which the magnitude and phase of signals from multiple microphones can be modified to achieve a better spatial selectivity. Fixed or adaptive beamformers can be used to optimally suppress (noise) signals from different directions.

- **Feedback cancellation**: Acoustic feedback is an important limitation in hearing aid fitting. It occurs when a fraction of the output sound from the hearing aid receiver leaks back into the microphone causing an annoying audible whistle in the ear. Feedback cancellation is performed by adding an adaptive compensation filter to rapidly detect the onset of feedback and suppress its build-up pattern as fast as possible (ideally before feedback becomes audible).

- **Noise cancellation**: There exist a large family of algorithms knows as Noise-reduction algorithms (NRAs) that are used to enhance the quality of speech or listening comfort. NRAs are nonlinear adaptive systems that rely on a speech detection mechanism (also known as a voice activity detector (VAD)), a temporal

envelope follower, or a statistical estimation procedure, in order to separate speech from noise.

- **Binaural processing**: Motivated by the natural binaural processing in the auditory system (*c.f.* Section 2.2.5), binaural algorithms, when a subject is fitted with two hearing aids, exploit the differences between the signals at the two ears (ITDs and ILDs in different frequency bands) to compensate for certain aspects of cochlear hearing loss. This technology was fueled by advancements in wireless communications (*e.g.*, bluetooth) and has led to significant improvements especially in addressing the "cocktail party" problem.

- **Output limiting**: Output limiting prevents the maximum sound at the output of the hearing aid from reaching a discomfort level (the UCL) and possibly causing damage to the wearer's ear. The easiest way to implement this is by peak clipping where the peaks of a signal that exceed a certain voltage are simply "clipped". This type of output limiting introduces virtually no delay, but results in a large amount of signal distortions. "Soft peak clipping" mechanisms may be implemented to reduce harmonic distortion. Output limiting may also be performed using an AGCo amplifier.

2.4.3.2 Artificial intelligence

Today, intelligence is the observable trend in the evolution of hearing instruments. In a typical day the aid wearer moves between diverse listening environments such as the office, the car, the cafeteria, or the street. The type of noise exposure, directionality, and room reverberation in all these circumstances are different. Moreover listening activities vary from conversing, to talking on the phone, to listening to music, or simply being in a quiet environment. Hearing aids are typically loaded with programs designed for specific purposes or auditory scenes, which can be turned on and off by the user to achieve the most appropriate compensation. An intelligent hearing device would eventually eliminate

the need for user intervention, and not only chose a suitable program for a given listening environment, but also fine-tune its settings for specific needs (settings which the aid wearer would normally not have access to). To this end, an intelligent hearing device would need to include some or all of the following capabilities:

- **Auditory scene adaptation**: This refers to the ability to recognize and adapt to changes in the listening environment and the nature of sound sources. Auditory scene adaptation is typically achieved through the use of classifiers trained to recognize a given number of preset environments and, on the long run, form and adapt to new environment categories as they are repeatedly encountered.

- **Adaptive signal enhancement and hearing compensation**: This gives the hearing device the capability to optimize the signal quality using any, or a combination of, the strategies discussed in the previous section as would be most suitable to a particular listening environment. An intelligent device may for example extend the concept of level-dependent compression to multiple frequency bands when appropriate.

- **Adaptive learning functionality**: The smart hearing aid may "decide" which settings are hypothetically optimal for the user in a specific listening environment. However, the aid wearer may have preferences which differ from these "optimal" settings, and may ultimately choose to change the controls. The smart device, then, must be equipped with an adaptive system, typically designed using common artificial-intelligence tools (such as neural networks, or fuzzy logic), which can be trained to learn and adapt to the user's preferences over time.

Such intelligent features have already become an integral part of many of today's hearing devices, and are setting forth the next milestones for the future of the hearing aid industry.

Chapter 3

Literature Review

Introduction

The aim of this doctoral thesis is to produce an objective measure to predict binaural speech intelligibility involving signals subject to nonlinear distortions common to hearing aids and other wearable devices. This chapter presents a survey of the literature pertaining to three relevant research areas: speech intelligibility, models of nonlinear distortion, and binaural hearing.

3.1 Speech intelligibility

It is important to distinguish between subjective tests used to assess speech intelligibility and objective measures used for prediction. Subjective tests can be used under various conditions and are ultimately the gold standard for evaluation or validation purposes. They are, however, time consuming and can be affected by attention level, fatigue, and other factors. Therefore, they are often used for a limited set of conditions per subject. In situations where large amounts of data on speech intelligibility are required, objective models are indispensable. In this section, we review these two aspects of speech intelligibility, while highlighting relevant subjective tests and objective models.

3.1.1 Subjective tests

A variety of speech audiometry tests are available to assess a subject's ability to understand speech in quiet or in the presence of background noise. These tests are classified as nonsense-syllable, word, or sentence tests, and each has advantages and disadvantages. For example, some would favor sentence tests arguing that they adequately reflect day-to-day communications, unlike isolated word tests which eliminate contextual information. A test generally consists of a large database of speech material (syllables, words, or sentences) arranged into different lists. The purpose of this arrangement is to present the speech material to listeners in a way to prevent learning or memorizing, especially when multiple conditions need to be tested. When designing a listening test, several factors are considered such as representation of all fundamental speech sounds (phonemes), equal difficulty of identification across different lists, and control of contextual information [141]. A few listening tests available in the English language are presented below:

- The everyday sentence test, developed in the 1950s [90].
- The synthetic-sentence identification (SSI) test, developed in the late 1960s [113].
- The modified rhyme test (MRT) introduced in the early to mid 1960s [94].
- The speech perception in noise (SPIN) test, developed in the late 1970s and revised in the 1980s [14].
- The hearing in noise test (HINT), developed in the early 1990s [153].
- The speech in noise (SIN) test, developed in the early 1990s [63], and the shorter version (QuickSIN) developed in 2004 [130].
- The words-in-noise (WIN) test, developed in the early 2000s [210, 211].
- The BKB speech-in-noise (BKB-SIN) test developed in the early to mid 2000s [64] based on sentence material from the Bamford-Kowal-Bench (BKB) test [10].

While clinical use of older tests is limited, the QuickSIN and WIN tests are most often used for discriminating between normal-hearing and hearing-impaired subjects. Owing to the semantic context in the speech material, BKB-SIN and HINT are easier tests for subjects, and are often used with young children, adults, or subjects with a high degree of hearing loss. The SPIN test is also used to assess the benefit of contextual cues in speech intelligibility. The MRT is widely used to assess speech intelligibility under communication systems, and has been included, along with two other tests, as a standard test material for such measurements in ANSI S3.2–2009 [1].

3.1.1.1 The Hearing In Noise Test

The HINT was developed at the Hearing Aid Research Laboratory at the House Ear Institute [153]. It was commercially made available in the early 1990s on compact disk and later in software/hardware systems. The test consists of sentences spoken by a professional male voice actor speaking General American English, an unaccented dialect free of noticeable regional influences [204]. The speech material was based on the BKB sentence list [10], which were revised to remove British idioms and rewritten in American English. The resulting sentences were normalized for sentence length, naturalness, difficulty, and phonemic distribution to ensure adequate representation of natural speech for native American English speakers. The sentences were originally grouped in 25 lists of 10 sentences, and later re-arranged in 12 lists of 20 sentences. Table 3.1 shows the first list of sentences from the original test. The test also includes 3 additional lists of training material (12 sentences each), and a speech-shaped noise with a long-term spectrum matching the average spectrum of all sentences. The original HINT was designed for binaural testing in the sound field, with the speech presented through speakers in quiet or with noise coming from any of three directions (front, left, or right). Later versions of the HINT allow stimulus presentation through headphones, where the masker location is simulated with HRTF. The HINT has grown in popularity over the years, and is now available in multiple languages [185].

Table 3.1: *Example sentences (List 1) from the HINT [153].*

1	(A/the) boy fell from (a/the) window.
2	(A/the) wife helped her husband.
3	Big dogs can be dangerous.
4	Her shoes (are/were) very dirty.
5	(A/the) player lost (a/the) shoe.
6	Somebody stole the money.
7	(A/the) fire (is/was) very hot.
8	She's drinking from her own cup.
9	(A/the) picture came from (a/the) book.
10	(A/the) car (is/was) going to fast.
Note:	*Allowed variations in response are shown in parenthesis with the words used in the original recording underlined.*

The purpose of the HINT is to measure the SRT for sentences for normal-hearing and hearing-impaired subjects in quiet or in background noise. In quiet, the SRT is defined as the speech presentation level at which listeners can repeat all keywords in sentences 50% of the time. When testing in background noise, the SRT is the SNR at which the keywords are identified correctly with 50% accuracy. To compute the SRT, the HINT uses an adaptive *up-down procedure*, which is known to converge towards the 50% point [39, 136]. Sentences from a list are presented at an initial presentation or SNR level. The level is decreased each time the listener correctly repeats all the words presented until the words can no longer be identified. The level is then increased until the listener can identify the words correctly. This procedure is repeated and the SRT is calculated as an average of speech or SNR levels over a specified number of sentences, while ignoring a number of initial sentences to reduce bias due to the starting level.

3.1.1.2 The Modified Rhyme Test

The MRT was developed as a simplified testing procedure for evaluating speech communication systems [94]. The format of the test is largely based on an earlier Rhyme Test [65], however, with relaxed constraints: word lists are not phonetically balanced, and strict orthographic restrictions have been loosened. The test material consists of

Table 3.2: *Two example word lists from the MRT [94].*

List #	Words					
1	bat	bad	back	bass	ban	bath
26	led	shed	red	bed	fed	wed

300 American English monosyllabic words which, except for a few exceptions, are of the form consonant-vowel-consonant (CVC). The stimulus words are arranged into 50 lists of 6 words. Within each list, the words consist of the same vowel-based nucleus and differ only by either the initial or final phoneme element. Table 3.2 shows two example lists of words from the MRT.

The purpose of the MRT is to measure the discrimination of initial and final consonant sounds. Answer forms with a closed set of alternative words are used to record the listener's responses: as the stimulus word is presented, the 6 alternatives are shown to the listener to identify which word was spoken. The stimulus word is normally presented within a carrier sentence spoken at approximately the same level for all words without emphasis (*e.g.*, "You will mark ____ now"). Listener responses can be scored according to the number of words heard correctly or incorrectly. Alternatively, the responses may be scored in terms of the frequency of particular consonant confusions [1].

3.1.2 Objective prediction models

Objective measures of speech intelligibility date back to the mid-20th century, when the AI was developed by engineers at AT&T research laboratories (later Bell Telephone Laboratories) to predict the impact of changes in telephone circuits on speech understanding [70, 71]. The measure, which was standardized in ANSI S3.5–1969 [3], takes as inputs the speech, noise and hearing threshold levels, and produces at the output an index which is monotonically related to the intelligibility of speech. The fundamental concepts used in the calculation of the AI have been extended in the development of several other measures, two of which are summarized next.

3.1.2.1 The Speech Intelligibility Index

A major revision of the AI standard in 1997 (ANSI S3.5–1997 (R2012) [4]) formally replaced the name Articulation Index by the Speech Intelligibility Index, or SII. The revision essentially provides a generalized framework for determining the input variables and the reference point for measurements (*e.g.*, free-field or eardrum). Other changes in the standard include new data which provide corrections for spread of masking and high presentation levels, as well as changes to the relative importance of different frequencies to speech intelligibility. A brief overview of the calculation is summarized below.

The SII is calculated as a weighted sum of audible cues available to the listener over a chosen number of frequency bands. The standard specifies four computation procedures which differ mostly in the number and size of the frequency bands used. In decreasing order of accuracy, these procedures are:

- The critical band procedure (21 bands)
- The one-third octave band procedure (18 bands)
- The equally contributing critical band procedure (17 bands)
- The octave band procedure (6 bands)

Mathematically, the computation of the SII can be written as:

$$SII = \sum_{i=1}^{n} I_i A_i \tag{3.1}$$

where n is the chosen number of frequency bands, I_i the relative importance of the frequency band i (also known as the Band Importance Function), and A_i is the audibility function which represents the proportion of the speech cues that are audible in that band. The values I_i depend on the speech stimulus and are summarized in different tables in the standard. The audibility function A_i is based on the SNR in a given band, and takes into account the subject's hearing thresholds and other factors that can negatively affect

speech intelligibility, namely the spread of masking and high presentation levels. The result is an index between 0 and 1 that is "highly correlated with the intelligibility of speech under a variety of adverse listening conditions" [3].

3.1.2.1.1 Extensions of the SII

The SII is designed to predict intelligibility under monaural conditions with linear processing and stationary additive noise. The standard also includes provisions to account for room reverberation (clauses 5.2 and 5.3 of ANSI S3.5–1997 (R2012)) using a procedure based on the modulation transfer function methodology developed for the STI (*c.f.* section 3.1.2.2). Several studies have shown, however, that the SII fails to predict speech intelligibility in the presence of fluctuating noise [68, 97, 205], since its computation is based on the long-term average spectra of the speech and noise signals. An extension to the SII, which uses a segmental approach to predict intelligibility in such conditions, has been proposed by Rhebergen and Versfeld [170]. Their extended measure (the ESII) was shown to correlate reasonably well with subjectively measured intelligibility scores for the majority of conditions with fluctuating noise [170, 171]. Other extensions of the SII have been proposed to deal with nonlinear processing or binaural listening conditions. These will be discussed in subsequent sections in this chapter.

3.1.2.2 The Speech Transmission Index

Another extension of the AI measure is the Speech Transmission Index (STI), commonly used in the evaluation of communication systems. The STI was originally introduced by Steeneken and Houtgast for evaluating VHF-radio systems [95, 96] and further refined in [188], then revised again by the same authors in 1999 in what is known as the STIr metric [189]. It was standardized in IEC 60268-16 [103] originally in 1998, and then revised to reflect the changes in the STIr in 2003. A comprehensive compilation of the development of the STI, as well as extensions, improvements and applications can be found in [194]. A brief overview of the computational procedure is presented next.

The STI follows the AI concept in that it is calculated as a weighted sum of contributions from different frequency bands. It uses artificial modulated test signals designed to model the fluctuations of the envelope of speech in connected discourse. Test signals are constructed for octave bands in the range 125 to 8000 Hz and using different modulation frequencies. A modulation transfer function (MTF) is measured by comparing the modulation of the envelope spectrum at the input and output of a communication channel. The computation also includes adjustments for spread of masking and auditory thresholds. From the MTF, an *effective SNR* (which includes the effects of distortions and noise) is computed for each octave band i, and modulation frequency k. Next, a transfer index $TI_{i,k}$ is derived from each value, and a modulation transfer index MTI_i is obtained for each octave band by averaging over all modulation frequencies.

$$MTI_i = \frac{1}{14}\sum_{k=1}^{14} TI_{i,k} \tag{3.2}$$

A weighted sum over all bands, taking into account redundancy from adjacent bands, produces the final STI value:

$$STI = \sum_{i=1}^{7} \alpha_i MTI_i - \sum_{i=1}^{6} \beta_i \sqrt{MTI_i \cdot MTI_{i+1}} \tag{3.3}$$

where the α_i and β_i coefficients are provided for male and female talkers in [190, 191].

3.1.2.2.1 Versions of the STI

The "full" STI procedure is a robust measure that can account for the effects of noise, linear filtering, and time-domain distortions such as room reverberation, echoes, and automatic gain control. Moreover, special provisions are included in the procedure to model harmonic distortion and inter-modulation components produced by nonlinear operations such as peak clipping and quantization. Nonetheless, the procedure requires the measurement of a large matrix of values resulting in long computation times. Over the years, several versions have been designed to reduce the amount of computations but which result in reduced applicability. The different versions of the STI are listed below:

- **STI-14**: the original version using 7 octave bands and 14 modulation frequencies. It can be used with all types of communication systems, except vocoders.
- **STI-3**: a reduced version which uses all 7 octave bands and 3 modulation frequencies. Its applicability is reduced to communication channels which do not suffer from time-domain degradations.
- **STITEL**: a reduced version which uses simultaneous modulation and parallel processing adapted for telecommunication systems that are free of time-domain and nonlinear distortions.
- **STIPA**: another reduced version which uses simultaneous modulation and parallel processing adapted specifically to public address systems. This version can be used with systems with time-domain degradations, but has limited coverage of nonlinear distortions.
- **RASTI**: a reduced version which uses only 2 octave bands and 4–5 modulation frequencies which can be used as a screening approach for person-to-person communications in a room acoustical environment. This version can handle time-domain degradations but fails to account for nonlinear distortions or band-pass filtering.

3.1.2.2.2 Speech-based STI extensions

The STI procedure relies on the use of artificial signals in order to measure modulation transfer. Many applications require a measure which operates directly on speech signals. Most early attempts at using a speech-based STI measure to predict speech intelligibility were without success [92, 93, 142, 143, 162]. Some methods, however, showed promising results and deserved further investigation [162]. Goldsworthy and Greenberg proposed modified versions of 4 speech-based STI methods which seemed to correlate well with the traditional STI under reverberation conditions [82]. Furthermore, intermediate modulation metrics for these methods showed "qualitatively reasonable results" when tested with 2 noise reduction algorithms (using spectral subtraction and envelope thresholding).

In [161], Payton and Shrestha investigated the use of two of these metrics with shorter speech segments. The two methods were shown to "accurately track short-term fluctuations in STI" under conditions of noise and reverberation. More recently, Brammer et al. proposed a speech-based STI model employing the coherence and cross-covariance functions [19, 20]. This will be briefly described in Section 3.2.2.2.

3.1.2.3 Comparing objective predictions to subjective scores

Subjective test results, such as percentage correct scores, are a monotonic function of intelligibility indices such as the AI, SII and STI. The exact function is not unique, and depends on the speech material used and the proficiency of talkers and listeners. For a given test, an analytic function is usually chosen and fitted to speech intelligibility scores for a large group of subjects. The following function has been suggested to relate the STI to a predicted intelligibility score based on a group of normal-hearing listeners [191]:

$$\text{predicted score} = \left(A \cdot e^{B \cdot STI} + C\right) \times 100\% \tag{3.4}$$

where the parameters A, B, and C are defined in tables for male and female speech for different phoneme-groups and CVC-words scores. A different relationship with boundary conditions was used by Beutelmann and Brand in their binaural extension of the SII [12]:

$$P(SII) = \frac{m}{a + e^{-b \cdot SII}} + c,\ P(0) = 0,\ P(1) = 1 \tag{3.5}$$

We can fit the parameters to subjective intelligibility data for HINT using the following constraints: (1) the boundary conditions, (2) the reference SRT for 50% intelligibility of -2.6 dB [204] which corresponds to a SII value of 0.375, and (3) a slope of 10.6% per dB increase in SNR [185]. This yields the following parameter values: $a = 0.007272$, $b = 13.17$, $m = 0.007327$, and $c = -0.007553$. This relationship, plotted in Figure 3.1, is not a unique mapping. In fact, the reference SII value of 0.375 depends on the speech material and language used, and has been found to vary by as much as ± 0.1 around the mean value for the ESII with various real-life background noises [172].

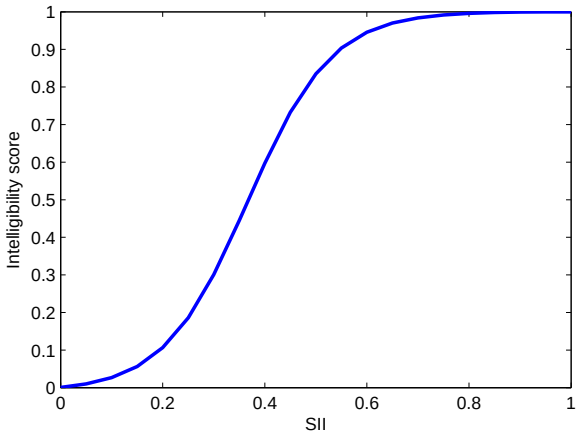


Figure 3.1: *Plot of the analytic function in equation 3.5 fitted to subjective data for the SII with the HINT [185, 204].*

3.2 Modeling nonlinear distortion in hearing devices

Hearing aids and other wearable hearing devices may contain nonlinear signal processing circuitry such as the ones found in AGC, NRAs, and output limiting systems. Such devices have signal-adaptive characteristics which produce distortion at the output. For example, rapid changes in the compressor's gain in a fast-attack AGC may produce side-effects, such as spectral splatter, that can be perceptually audible. Distortions may also be introduced by imperfect transducers, analog-to-digital converters and other circuitry. Developing measurement techniques to obtain meaningful performance measures for such devices is a growing research area. In this, section we review some of the methods and speech intelligibility models designed to address nonlinear distortions in hearing devices.

3.2.1 Signal and distortion estimation methods

3.2.1.1 The coherence function

Coherence measurements using speech and broadband signals to evaluate nonlinear hearing aids [46–48, 118, 119] were standardized in ANSI S3.42–1992 Part 1 (R2012) [2]. The magnitude-squared coherence (MSC) function [28], is a frequency-domain measure that

reflects the linear relationship between two signals. For two stationary random processes $x(t)$ and $y(t)$, the MSC is computed according to equation (3.6):

$$|\gamma(f)|^2 = \frac{|S_{xy}(f)|^2}{S_{xx}(f) \cdot S_{yy}(f)} \tag{3.6}$$

where $S_{xy}(f)$ is the cross-spectral density between $x(t)$ and $y(t)$, and $S_{xx}(f)$ and $S_{yy}(f)$ are the auto-spectral densities. The MSC represents the fraction of $y(t)$ that is linearly dependent on $x(t)$: the coherent portion. If $x(t)$ and $y(t)$ are the input and output of a system, $|\gamma(f)|^2 = 1$ indicates that the system is linear. The MSC is decreased by noise and nonlinear distortion. The fraction $1 - |\gamma(f)|^2$ is the noncoherent portion of $y(t)$, and the system's signal-to-distortion ratio (SDR) is defined as:

$$SDR(f) = \frac{|\gamma(f)|^2}{1 - |\gamma(f)|^2} \tag{3.7}$$

Mixtures of speech, $u(t)$, and noise, $v(t)$, signals are often used in hearing aid analysis. By computing the MSC between the system's output, $y(t)$, and the input speech signal, $u(t)$, the coherent and noncoherent portions of $y(t)$ will provide an estimate of the power spectra of the output speech and noise signals, $u'(t)$ and $v'(t)$ respectively:

$$\widehat{S}_{u'u'}(f) = |\gamma(f)|^2 \cdot S_{yy}(f) \tag{3.8}$$

$$\widehat{S}_{v'v'}(f) = (1 - |\gamma(f)|^2) \cdot S_{yy}(f) \tag{3.9}$$

In practice, the MSC is estimated using fast Fourier transform techniques. Signals are divided into M overlapped windowed segments and the cross-spectrum and autospectra are computed for each segment and averaged across segments, as follows:

$$|\hat{\gamma}(f)|^2 = \frac{\sum_{m=0}^{M-1} |X_m(f) Y_m^*(f)|^2}{\sum_{m=0}^{M-1} |X_m(f)|^2 \cdot \sum_{m=0}^{M-1} |Y_m(f)|^2} \tag{3.10}$$

This estimation is subject to variance and bias [2, 118]. Variance is inversely proportional to the number of frames, and can be reduced with smoothing techniques like the third-octave running average. The sources of bias are the number of frames used in averaging

and the system's overall and group delay. Averaging bias is greatest when the coherence is near zero and is inversely proportional to the number of frames. The overall delay can be removed by ensuring frames are aligned [118]. The group-delay bias is dominant for coherence values close to one, and is inversely proportional to the frame length. The significance of the bias can be tested by comparing estimates with different frame lengths [2]. Choosing the optimal frame length is a trade-off between these sources of bias and variance. Other unbiasing techniques are discussed in [118] and [119].

3.2.1.2 The phase-inversion method

In a study evaluating SNR improvements due to NRA and AGC hearing aid systems, Hagerman and Olofsson proposed a method to obtain estimates of the speech and noise signals at the hearing aid output [87]. Their approach is based on two measurements using speech and noise mixtures, the second with the noise phase reversed (the phase-inversion method). The recovered estimates are used to derive the system's equivalent speech and noise transfer functions, from which a frequency-weighted SNR is derived. Owing to its inherent mathematical simplicity, this approach has been employed by several groups in assessing hearing aid performance [34, 75, 152, 155, 182, 186, 187, 212].

This signal separation is performed by presenting the speech and noise mixture to the hearing aid twice; the second time with the phase of the noise inverted. The input-output pairs (a_{in}/a_{out} and b_{in}/b_{out}) can be modeled by:

$$a_{in}(t) = u(t) + v(t), \quad a_{out}(t) = u'(t) + v'(t) + e_1(t) \tag{3.11}$$

$$b_{in}(t) = u(t) - v(t), \quad b_{out}(t) = u'(t) - v'(t) + e_2(t) \tag{3.12}$$

where $e_1(t)$ and $e_2(t)$ are additive errors that model nonlinear distortions in the device. Estimates of $u'(t)$ and $v'(t)$ are recovered using equations (3.13) and (3.14) respectively:

$$c_1(t) = \tfrac{1}{2}[a_{out}(t) + b_{out}(t)] = u'(t) + \tfrac{1}{2}[e_1(t) + e_2(t)] \tag{3.13}$$

$$d_1(t) = \tfrac{1}{2}[a_{out}(t) - b_{out}(t)] = v'(t) + \tfrac{1}{2}[e_1(t) - e_2(t)] \tag{3.14}$$

If the hearing aid is linear, $e_1(t)$ and $e_2(t)$ would vanish and the recovered signals would be the same as though the speech and noise signals were processed independently. For nonlinear devices, the recovered estimates are accurate only if the error signals are sufficiently small. In [87], the authors confirmed this by comparing plots of the distortion measure described in Section 3.2.1.3 with nonlinear mechanisms turned on and off. Most recently, Suelzle employed the phase-inversion method to evaluate and compare different hearing aid systems equipped with NRAs [193]. Using the recovered signals at the output of each hearing device, the authors developed a test framework which includes the SII, two speech quality measures (the Hearing Aid Speech Quality Index (HASQI) [126] and a modified version of the Speech-to-Reverberation Modulation Ratio (SRMR) [66]), as well as an estimate of the NRA's attack time.

3.2.1.3 A Hilbert-transform distortion measure

Olofsson and Hansen derived an objective measure to quantify nonlinear distortions produced in hearing aids [154]. Distortion may be characterized based on technical aspects such as harmonic or intermodulation distortion. In this work, the authors focused on perceptual distortion. The measure is based on analyzing pairs of signals related by the Hilbert transform as they are passed through the hearing aid.

The Hilbert transform [156] is a time-domain operator which, for a signal $x(t)$, assigns the function $\mathcal{H}[x(t)]$, denoted $\tilde{x}(t)$. It is often conveniently defined as a linear filtering operation in the frequency domain: $\widetilde{X}(f) = H(f) \cdot X(f)$, where $X(f)$ and $\widetilde{X}(f)$ are the Fourier transforms of $x(t)$ and $\tilde{x}(t)$ respectively, and $H(f)$ is the filter defined as:

$$H(f) = \begin{cases} -j & f > 0 \\ 0 & f = 0 \\ +j & f < 0 \end{cases} \tag{3.15}$$

This definition illustrates an important property of the Hilbert pair $(x(t), \tilde{x}(t))$: they have the same amplitude spectrum, with a phase shift of $\pm\pi/2$ at negative and positive

frequencies. This is critical for the analytic signal $x_a(t) = x(t) + j\tilde{x}(t)$, whose two-sided Fourier transform contains only the non-negative frequency components of $X(f)$:

$$X_a(f) = X(f) + jH(f)X(f) = \begin{cases} 2X(f) & f > 0 \\ X(0) & f = 0 \\ 0 & f < 0 \end{cases} \tag{3.16}$$

Another property is that $\mathcal{H}[-\tilde{x}(t)] = x(t)$. Thus the pair $(x(t), \tilde{x}(t))$ have the same Hilbert envelope (magnitude of the analytic signal). Moreover, the convolution property, mathematically expressed as

$$\mathcal{H}[x(t) * g(t)] = \tilde{x}(t) * g(t) = x(t) * \tilde{g}(t) \tag{3.17}$$

implies that when signals which form a Hilbert pair are passed through a linear system, the output signals will also form a Hilbert pair. Finally, the Hilbert product theorem [8] states that under certain constraints,

$$\mathcal{H}[a(t)x(t)] = a(t)\tilde{x}(t). \tag{3.18}$$

A sufficient condition for the product theorem to hold is for $a(t)$ to be a low-pass signal and $x(t)$ a high-pass signal with non-overlapping spectra (*i.e.* $A(f) = 0$ for $|f| > f_0 > 0$ and $X(f) = 0$ for $|f| < f_0$).

Consider an AGC system whose gain signal, $a(t)$, depends on the input signal envelope. Since $x(t)$ and $\tilde{x}(t)$ have the same envelope, equation (3.18) implies that the Hilbert-transform relationship between the corresponding output signals could describe the system's nonlinear distortion. Thus, the distortion measure in [154] is based on two measurements using $x_1(t)$ and $x_2(t) = \tilde{x}_1(t)$ as input signals. The output signals, $y_1(t)$ and $y_2(t)$ respectively, are combined to form the signal $v(t) = y_1(t) + jy_2(t)$, for which a two-sided power spectral density (PSD) is computed. If the hearing aid is linear, the output signals would form a Hilbert pair, and the PSD of $v(t)$ would be zero for $f < 0$. For a nonlinear device, the authors define the portion of the PSD over negative frequen-

cies, $S^{-}(f) = S_{vv}(-f)$, as the product of distortion, and that corresponding to positive frequencies, $S^{+}(f) = S_{vv}(f)$, as the linear response. Thus, they propose a new definition for the SDR given by:

$$SDR(f) = \frac{S^{+}(f)}{S^{-}(f)} \tag{3.19}$$

When A-weighted to produce a single value, the SDR defined in equation (3.19) was found to correlate well with subjective scores of distortion when testing with different types of AGC hearing aids [154].

3.2.2 Intelligibility models for nonlinear processing

3.2.2.1 The coherence SII

Kates and Arehart used spectral estimates recovered with the coherence function using equations (3.8) and (3.9) with the SII to predict speech intelligibility under conditions of additive noise, peak clipping and center clipping [124]. They argued that these distortion mechanisms affect signals at different levels in dissimilar ways: for example center clipping tends to remove large fractions of the consonant information (lower-level sounds) while peak clipping tends to affect vowels, which are generally associated with higher-levels sounds. They concluded that a straightforward application of the coherence function to the SII computation to produce a single-level index (the coherence-based SII or $CSII$) "cannot accurately predict speech intelligibility for all forms of distortion and noise". They proposed a three-level approach to deal with these forms of distortion. Their measure is based on partitioning the signal envelope into three amplitude regions (low, medium and high), and computing the $CSII$ in each region independently. A weighted sum of the $CSII$ values in each region produces a single value, which can be transformed into an intelligibility index using a monotonic function:

$$c = -3.47 + 1.84CSII_{Low} + 9.99CSII_{Mid} + 0.0CSII_{High} \tag{3.20}$$

$$I_3 = 1/(1 + e^{-c}) \tag{3.21}$$

The I_3 index[1]was found to produce accurate speech intelligibility predictions for normal-hearing and hearing-impaired subjects under the noise and distortion conditions tested. The procedure was later extended to produce a speech quality measure in [125]. The Q_3 index is essentially computed in the same way as in equation (3.20) using a different weighting which better describes how the low-, mid- and high-level $CSII$ affect speech quality:

$$Q_3 = 4.42CSII_{Low} + 3.74CSII_{Mid} + 2.87CSII_{High} \tag{3.22}$$

3.2.2.2 A coherence-based STI model

The coherence function was also used in a STI-based intelligibility model which operates on speech signals proposed by Brammer et al. [19, 20]. First, the MSC between the original and corrupted speech signals is computed for each octave band i, and modulation frequency k. This is used to correct the estimate of modulation reduction under the influence of noise and distortion (transfer index $TI_{i,f}$ in Equation (3.2)) as follows:

$$MTI_i = \frac{1}{14}\sum_{k=1}^{14} TI_{i,k} \cdot \gamma_{i,k}^2 \tag{3.23}$$

The authors also introduced the normalized cross-covariance function between intensity modulations in three adjacent octave bands, $\rho_{i,j}(j = i+1, i+2)$ to correct for inter-band redundancy. The new intelligibility measure is expressed as:

$$STI_{\rho-speech} = \sum_{i=1}^{7} \alpha_i MTI_i - \frac{1}{6}\sum_{i=1}^{6}\sum_{j=i+1}^{i+2} \sqrt{\rho_{i,j}^2 \cdot MTI_i \cdot MTI_j} \tag{3.24}$$

where α_i are the coefficients used in Equation (3.3). The $STI_{\rho-speech}$ predictions were found to successfully account for additive noise and center clipping distortion when compared with subjective scores obtained with the MRT (*c.f.* Section 3.1.1.2).

[1]Although the authors originally used this terminology to denote their measure, it has become popularly named the Coherence SII (CSII) in the literature (*e.g.*, [127, 197]).

3.2.2.3 Other models of speech intelligibility

Several objective measures, not necessarily based on the SII or STI, have been developed to predict speech intelligibility under different monaural nonlinear processing conditions. The overall rationale is that signal degradations due to noise and distortion affect the signal's temporal fine structure (TFS) as well as the signal envelope. A generally accepted paradigm in distortion modeling has been to make use of some type of correlation measure between the processed and pure signals in the TFS or envelope domain. The CSII [124] successfully predicted speech intelligibility for three distortion mechanisms based strictly on the TFS (*c.f.* Section 3.2.2.1). Other models, based strictly on the signal envelope, have produced better results under some conditions.

Christiansen et al. developed a measure [29] using a peripheral auditory model [36], which transforms the input stimulus into an internal representation of auditory nerve activity before extracting the signal envelope. The intelligibility measure computes an envelope correlation metric between the model outputs for a reference and distorted speech signals. Inspired by the approach in [124], signals are processed in 20-ms overlapping segments, which are classified in three levels (low, medium, and high). The correlation metric is weighted across the levels followed by a logistic function transformation to obtain the final intelligibility prediction [29]. The measure produced good predictions for noisy speech processed using ideal time-frequency segregation (ITFS) [24], a spectro-temporal signal processing technique based on the ideal binary mask (IBM) [98] used in auditory scene analysis. Taal et al. argued that longer segments were needed to capture some low temporal modulations which are important to speech intelligibility [196]. They developed the short-time objective intelligibility (STOI) measure specifically to handle ITFS-processed speech by averaging the envelope correlation over 384-ms segments. This resulted in improved intelligibility prediction for noisy speech processed through ITFS and two single-channel NRAs ([60, 62]). In a concurrent study [197], the authors evaluated 17 objective measures of speech intelligibility under the same process-

Table 3.3: *Summary of objective measures evaluated in [197] and their performance (correlation coefficients) for speech intelligibility prediction with speech processed with ITFS or two NRAs. Three of these methods presented (CSII, CSTI, and FWS2) used a modified frequency weighting according to new band-importance functions proposed in [144].*

Objective measure	Abbr.	Ref	ρ (ITFS)	ρ (NRA)
Magnitude spectral correlation coefficient	MCC	[197]	0.93	0.93
Log spectral correlation coefficient	LCC	[197]	0.88	0.90
Dau auditory model	DAU	[29, 36]	0.89	0.84
Normalized subband envelope correlation	NSEC	[16]	0.89	0.75
Coherence SII	CSII	[124]	0.45	0.92
Normalized covariance based STI	CSTI	[82]	0.80	0.87
Perceptual evaluation of speech quality	PESQ	[9]	0.41	
Log likelihood ratio	LLR	[83]	0.35	
Itakura saito distance	IS	[107]	0.33	
Cepstral distance	CEP	[83]	0.19	
Segmental SNR	SSNR	[38]	0.58	
Magnitude spectral distance	MSD		0.88	0.90
Log spectral distance	LSD		0.30	
Frequency weighted SSNR	FWS1	[200]	0.67	
Normalized frequency weighted SSNR	FWS2	[101]	0.69	0.89
Weighted spectral slope metric	WSS	[132]	0.60	
Van de Par auditory model	PAR	[201]	0.52	

ing conditions. A list of these measures along with the correlation coefficients between their predictions and subjective scores is given in Table 3.3. They concluded that good performing measures for ITFS-processed speech were also good intelligibility predictors for speech processed with single-channel noise reduction, but not vice-versa.

Jørgensen and Dau developed the speech-based envelope power spectrum model (sEPSM) to predict the intelligibility of noisy speech based on the estimation of an envelope signal-to-noise ratio (SNR_{env}) [114]. The model uses an auditory filterbank and a modulation filterbank to extract the noise and noisy-processed envelope in different bands. Thus, the estimation of SNR_{env} is based on changes of modulation in the noise envelope and not the speech envelope. The model, which completely ignores the signal TFS, produced good speech intelligibility predictions for stationary masking conditions

involving room reverberation and spectral subtraction. A multi-resolution extension of the model was presented in [115] which successfully predicted intelligibility under fluctuating masking conditions. The extended model uses a temporal segmentation approach with modulation-filter dependent segment durations.

Most recently, Kates and Arehart proposed the Hearing Aid Speech Perception Index (HASPI), an intelligibility measure which incorporates information from both the TFS and envelope domains [127]. It is based on a model of auditory periphery [122] which includes complex modeling of outer and inner hair-cell damage. The model outputs two internal signal representations: an auditory output based on basilar membrane motion and a compressed envelope output. HASPI compares the model outputs for a processed test signal and a reference (unprocessed) signal. The comparison includes an envelope correlation metric as well as a three-level auditory correlation metric (*i.e.*,based on TFS) which are weighed to produce the final measure. The HASPI measure gives accurate intelligibility predictions for speech degraded by noise and a wide range of processing including peak and center clipping, frequency compression, IBM noise suppression, as well as vocoded speech.

3.3 Binaural hearing

In Section 2.2.5, we introduced the notion of binaural hearing along with the interaural time and level differences (ITD and ILD respectively). The auditory system makes use of these binaural cues embedded in the signals arriving at the left and right ears in order to improve our ability to understand speech in background noise. This is, in fact, an active research area aiming to better understand the "cocktail party phenomenon", a term used to describe the complicated task of extracting information from a specific talker when others are speaking at the same time. In this section, we present a brief overview of this phenomenon and highlight relevant studies dealing with subjective measurement or objective modeling of speech intelligibility in a binaural listening environment.

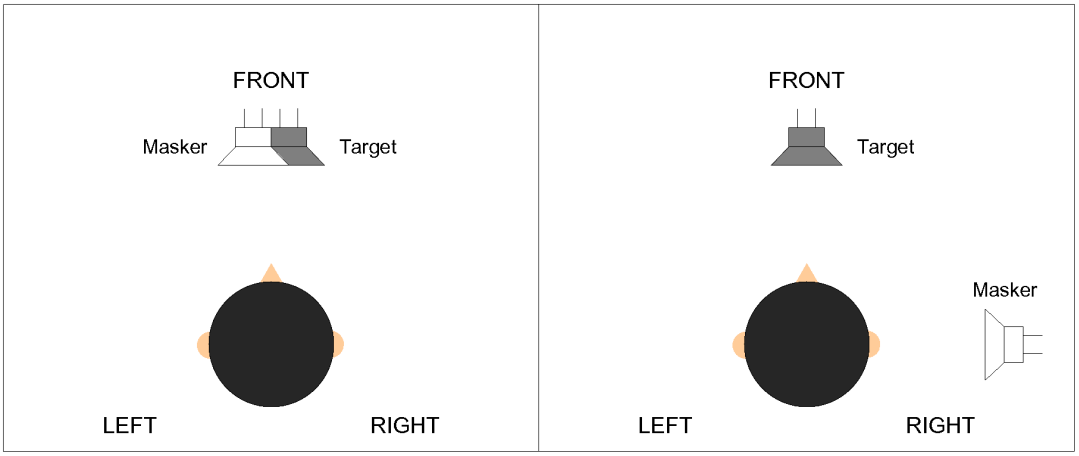


Figure 3.2: *Examples of spatial configurations commonly used in the study of SRM: target and masker co-located in front of the listener (left panel); target source in front, with masker to the right of the listener (right panel).*

3.3.1 The "cocktail party phenomenon"

Introduced 60 years ago [49], the cocktail party problem has since been the subject of a number of studies for normal-hearing and hearing-impaired populations. In order to understand how the auditory system is able to focus on a speech signal (the "target") in the presence of competing noise (the "masker"), much research has been dedicated to studying SRM: the improvement in speech understanding which occurs when the target and masker are spatially separated compared to when they are co-located. When sources are spatially separated (*e.g.*, right panel of Figure 3.2), the difference in binaural cues for the target and masker helps reduce (release) the masker interference compared to the co-located condition (*e.g.*, left panel of Figure 3.2). This principle introduces the notion of binaural masking level difference (BMLD) frequently used in studies of SRM where ITD and ILD cues are experimentally manipulated. The BMLD is defined as the difference between the threshold of a pure tone or narrowband noise target signal when the target and masker have different phase and/or level relationships at the two ears (separated sources), and the threshold when these relationships are the same (co-located sources). The BMLD can be thought of as a simpler version of the binaural intelligibility level

difference (BILD) defined as the difference in speech intelligibility thresholds between the separated and co-located conditions.

A recent review covering a wide range of SRM studies in adults, children and special populations can be found in [140]. In the literature, the term "energetic masking" is often used to refer to the spectral and temporal interference of signals due to the signal energy which is accounted for by the peripheral auditory system. Conversely, the term "informational masking" is used to refer to confusability which arises when the target and masker contain similar content or have similar fundamental frequencies. Finally, in addition to masking and binaural processing, the cocktail party problem also involves non-sensory "top-down" processes (such as attention, cognition, memory and emotion) which, along with informational masking, are beyond the scope of this thesis work.

3.3.2 Spatial release from masking: subjective studies

Benefits due to binaural cues have been reported in various experimental studies of speech intelligibility. Under earphone listening, Levitt and Rabiner reported that reversing the phase of the speech signal in one ear while maintaining the noise maskers in phase at the two ears ($S_\pi N_0$ condition) produces a decrease of about 6 dB in SRT for 50% intelligibility [137]. Plomp and Mimpen introduced natural interaural differences by moving a real noise source from a frontal position to the side while keeping the speech source in front [166]. The measured SRT as a function of noise azimuth is shown in Figure 3.3 (closed circles). An increase in SRT of about 9-10 dB is obtained for a noise azimuth of 90°, with the largest release from masking (10–11 dB) occuring when the noise is located behind the interaural axis (around 110–120°).

Bronkhorst and Plomp reproduced those measurements using maskers with simulated ITD and ILD cues obtained from sound-field manikin recordings and reproduced through earphones (Figure 3.3, open circles) [22]. Similar to [166], they found a binaural improvement in SRT of 10 dB when virtually moving a noise source from the frontal to the lateral position. This represents a binaural benefit of about 2.5 and 13.2 dB in SRT compared

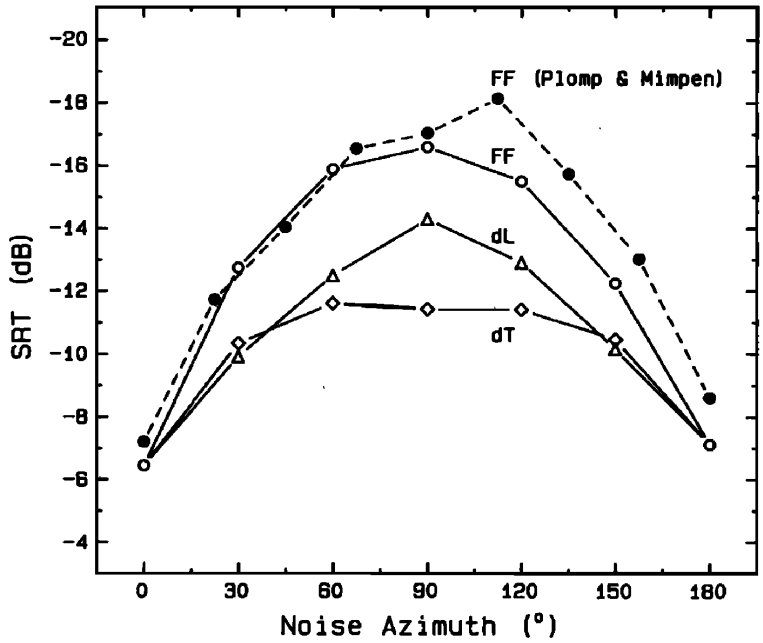


Figure 3.3: *Mean SRT obtained with frontal speech presentation as a function of noise azimuth. The data in closed circles is adapted from free field measurements reported by Plomp and Mimpen [166]. The data in open markers are the results of measurements by Bronkhorst and Plomp using a masker containing both ILD and ITD cues (simulated free-field condition, FF), or with ITD cues alone (dT) or ILD cues alone (dL) [22].*

to monaural listening through the ear with the best and worst SNR respectively. The authors further investigated the effects of ITD and ILD cues separately (Figure 3.3, open diamonds and open triangles respectively) and found that these effects were not additive: the binaural gain was 5 dB for maskers with ITD cues alone, and 7 dB for maskers with ILD cues alone. Finally, the authors extended this work to hearing-impaired listeners and found that their mean SRT was 2.5 dB higher than for normal-hearing listeners when the speech and masker were both presented from the front [23]. However, when the masker was moved to the lateral position, hearing-impaired listeners experienced 2.6–5.1 dB less binaural gain than normal-hearing listeners. Further measurements with maskers containing only ITD or ILD cues suggest that this decrease in binaural gain is due to an inability to take full advantage of ILD cues.

3.3.3 Spatial release from masking: objective models

Several models of binaural hearing exist in the literature, which attempt to model how the auditory system benefits from binaural cues. These models generally fall in two categories: physiological models simulate neural activity at the cell level in the auditory nervous system, while psychoacoustic models provide a mathematical modeling of auditory pathways in order to predict psychological behaviour such as speech localization, detection or reception. In this thesis, we are interested in psychoacoustic models, and their application to the task of speech reception (or more precisely intelligibility prediction). Binaural models include the early coincidence model by Jeffress [108], the geometric vector model [109–111], Durlach's "equalization-cancellation" EC model [33, 42–45], as well as models based on interaural cross correlation [40, 146, 157, 177]. An in-depth review of the topic can be found in [18]. Next, we present the theoretical background behind the EC model which is of paramount importance to this work.

3.3.3.1 Equalization-cancellation theory

Originally introduced by Durlach [42], EC theory models masking release when listening to a target speech signal in the presence of a noise masker. The theory suggests that the auditory system transforms the total signal at one ear relative to the total signal at the other ear such that the masking components are equalized (equalization process), and then subtracts one from the other (cancellation process). This is illustrated in Figure 3.4 for a given target-masker pair. It is assumed that the only differences between the signals presented are the ITDs and ILDs. Therefore, the equalization transform consists of applying gain/attenuation and time-delay factors. If the transform is perfect, and provided that interaural differences for the target signals are different from those for the maskers, the latter will ideally cancel out due to destructive interference. If these differences are identical, the EC processing will cancel out both target and masking components. Otherwise, the residual signal will theoretically have an improved SNR.

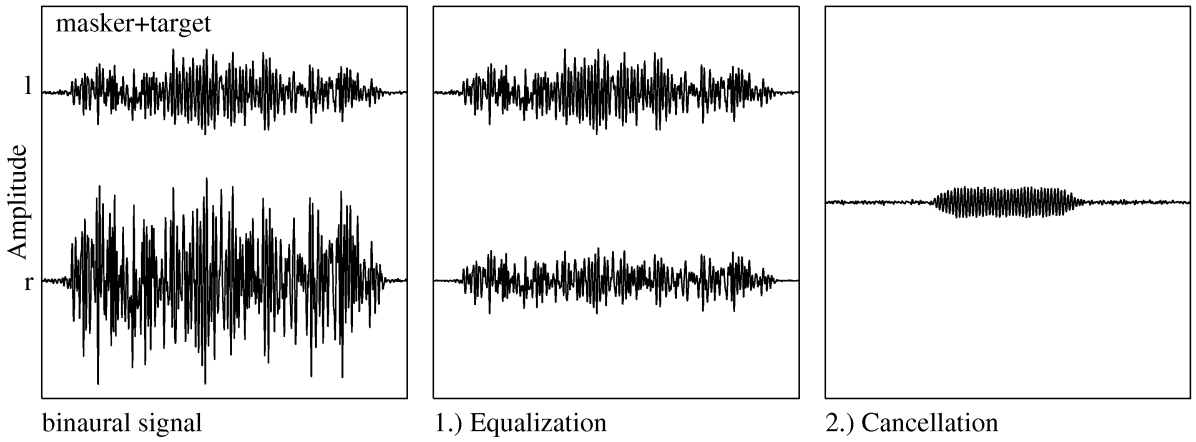


Figure 3.4: *Illustration of the EC model for a target-masker stimulus pair at the left (l) and right (r) ears. The target is a 500 Hz pure tone presented diotically. The masker is a narrowband noise centered at 500 Hz and presented with an ILD of 10 dB. (Adapted from [18])*

The structure of the EC model is shown in Figure 3.5. At the front end, the signals at the left and right ears are processed through a bank of band-pass filters, which simulate the human auditory filters (*c.f.* Section 2.2.2). The EC process described above is carried out in independent stages in each band. At the back end, a decision device operating according to signal-detection theory [84, 85] consists of a selection mechanism at the output of each EC stage which selects the signal with the best SNR (left, right or EC-processed) in each band. Finally, the goal of the model is not to achieve the best technically-possible masker cancellation, but to simulate the ability of the human auditory system to perform the speech detection task. To that end, the EC model introduces noise in the form of independent Gaussian gain and delay variables (δ_l, δ_r, ϵ_l, and ϵ_r) that model human inaccuracies. The variances of these artificial errors have been optimized to improve the accuracy of the EC model in predicting BMLDs.

3.3.3.2 Models of binaural speech intelligibility

Unlike binaural models of speech detection which use an elaborate model of peripheral preprocessing (outer/middle ear, basilar membrane and hair cell behaviour) to convert signals into an internal representation to be processed by an EC-like mechanism [21, 215],

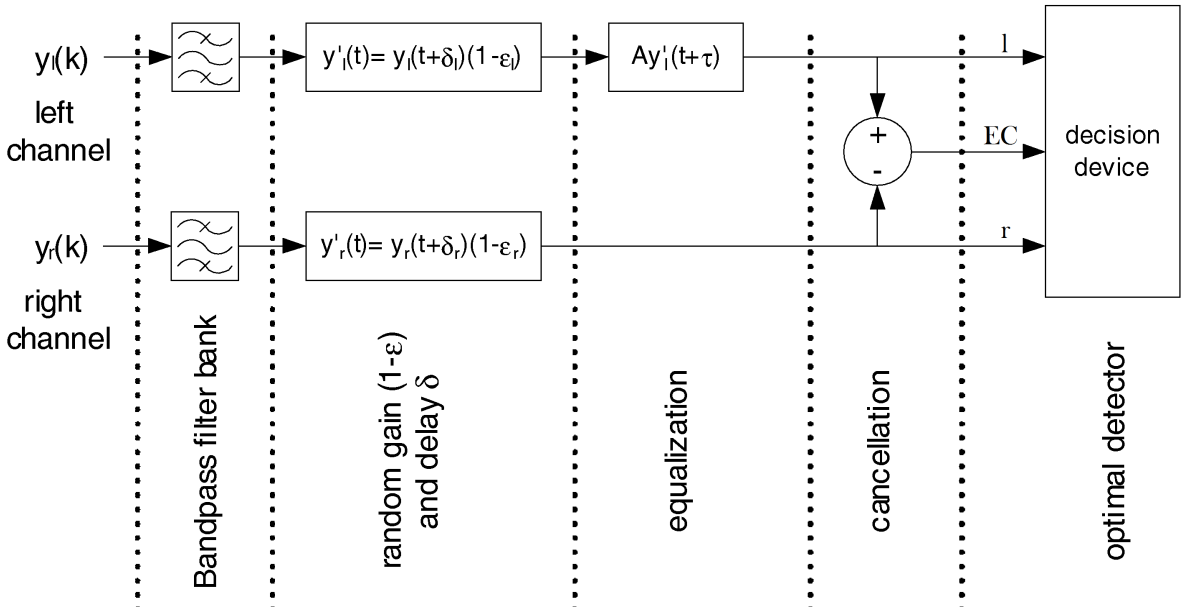


Figure 3.5: *Structure of the EC model. (Adapted from [18]).*

models of binaural speech intelligibility are kept more simple. These models typically consist of a model of binaural interaction, such as the EC model, acting as a processing unit for monaural intelligibility models such as the AI, SII or STI. In essence, the benefit of binaural interaction is expressed in terms of a reduction in masking (an improvement in SNR) in different frequency bands which positively affects intelligibility predictions. In [138], Levitt and Rabiner proposed a model based on the AI, which includes a frequency-dependent reduction in masking noise level derived from data on binaural tonal unmasking from Durlach. vom Hövel proposed a model based on the EC model and a modified version of the AI, which incorporates a coarse estimate of the effect of reverberation [206]. Zurek proposed another AI-based binaural model in which release from masking was calculated based on an equation from the binaural interaction model of Colburn [31, 32].

In [12], Beutelmann and Brand proposed a binaural intelligibility measure using the EC model and the SII, to predict intelligibility for normal-hearing and hearing-impaired subjects under different spatial masker configuration and conditions of reverberations.

Their model, which maintained as closely as possible the approach followed in [206], uses internal uncorrelated masking noise signals to model hearing thresholds at the two ears prior to the binaural processing. EC processing is performed using a modified implementation of the artificial errors suggested by vom Hövel. The model uses a Gammatone filterbank [91] to split the input signals into 30 frequency bands which model the auditory filters. This implementation allows the signals to be resynthesized at the output of the independent EC stages in each band. The SII is computed from the resynthesized signals and speech reception thresholds are obtained using a SII-to-intelligibility mapping which was fitted to measurement data. The details of the model will be elaborately explained in the next chapter in Section 4.1.2. Overall, the model predictions yielded a high correlation coefficient (0.95) when compared with subjective measurements, however release from masking was overestimated for mild hearing losses. The model was revised in [13] to improve computational speed using an analytical expression of binaural unmasking for the implementation of the EC process. The model revision, further reduces complexity by removing the signal resynthesis stage and adapting the standard SII computation to the 30 bands of the analysis filterbank. The revised Binaural Speech Intelligibility Model (BSIM) is further extended for modulated maskers using short-time block processing similar to the approach suggested by Rhebergen and Versfeld [170]. The resulting measure is denoted stBSIM.

Finally, we mention the binaural intelligibility model of Van Wijngaarden and Drullman [202]. The model is similar to the previous binaural schemes in that it also uses a binaural interaction model which acts as a processing unit for a monaural measure, the STI. However, a concerted effort is made to reduce the complexity of the model in order for the implementation to be a feasible extension to current STI measuring devices. Compared to the other binaural models, the current model represents a gross simplification. For example, instead of using a much narrower filterbank to model auditory filters, the STI is computed over 7 octave bands. The model operates on two-channel recordings of the artificial test signals of the STI (*c.f.* Section 3.1.2.2) as follows: First,

the STI is computed on the left and right ears independently using standard equipment. Next, binaural unmasking is accounted for in two ways:

- For the octave bands centered at 125 and 250 Hz and at 4000 and 8000 Hz, a simple better-ear mechanism is implemented by taking the best signal (highest value for the MTF) between the left ear and the right ear. This approach accounts for the effect of ILDs [50].

- For octave bands centered at 500, 1000 and 2000 Hz, in which the most useful binaural interactions occur, interaural correlograms are calculated in short frames, and the MTF is computed as a function of internal delay and frequency band. This approach models the effects of ITDs.

The MTFs computed in each frequency band are combined to calculate the overall STI. The binaural STI was validated under 39 dichotic listening conditions. The relationship between the binaural STI predictions and subjective CVC word scores was found to closely match the STI reference curve for monaural listening.

3.4 Summary

This chapter presented a review of the literature dealing with speech intelligibility measurement and prediction, modeling of nonlinear distortion, and binaural hearing. On the one hand, various models based on the coherence function or some type of correlation measure have been proposed for speech intelligibility modeling under monaural listening situations with nonlinear processing conditions. A common thread in these measures is the use of an auditory model or some form of preprocessing that transforms auditory signals into internal representations. On the other hand, binaural models based on the SII or STI have also been proposed using a better-ear listening approach, or a combination of binaural processing and better-ear listening; however, these methods only work under linear conditions. The current research addresses the need for a binaural model predict-

ing speech intelligibility under nonlinear processing conditions commonly encountered in hearing devices such as hearing aids, electronic hearing protectors, and communication headsets. Combining the benefits of all the monaural coherence- or correlation-based methods above presents the difficult technical challenge of incorporating the internal signal representations into a binaural processing scheme. The phase-inversion signal-separation method and the Hilbert-based distortion estimation approach may provide a useful alternative to perform distortion modeling in a binaural context. The following chapters will present our proposed solution based on the SII, and its validation using subjective SRT data measured with an in-house speech testing environment using the HINT.

Chapter 4

Modeling and Measurement Approach

Introduction

The main challenge in this work is to derive a binaural measure of speech intelligibility in which we isolate the effects of nonlinear distortions produced by hearing devices from the effects of binaural cues at each ear. As we have seen in the previous chapter, these factors tend to have competing effects on speech intelligibility: distortions may have adverse effects on sound quality and clarity, and may reduce speech intelligibility while binaural cues are known to improve the intelligibility of speech in noise.

In this chapter, we propose a new model designed to address these challenges. The chapter is based, in part, on the publications listed in Section 1.2. The contributions discussed in the upcoming sections can be summarized as follows:

- A three-stage model is proposed to predict speech intelligibility under binaural listening and nonlinear processing conditions.
- A novel method is derived to estimate speech and noise distortion introduced by nonlinear signal processing in the first stage of the model.

- A new testing environment is implemented in Matlab to acquire speech perception data under processing and listening conditions not possible with the commercial HINT software.

- Subjective measurement of SRT are performed with the new testing environment using sentences from the HINT in order to validate the proposed model.

In the remainder of this chapter we discuss the details of the model (Section 4.1), as well as an in-house hearing device simulator which has been thoroughly extended over the course of this work (Section 4.2). This simulator has been integrated into the new Matlab speech testing environment described in Section 4.3. In Section 4.4, we propose a two-step approach to validate the model followed by concluding remarks in Section 4.5.

4.1 Proposed Binaural Intelligibility Model

Figure 4.1 shows a block diagram illustrating our three-stage objective measure of binaural speech intelligibility based on the SII. The system's inputs are the individual speech and noise signals received at the left and right ears (four signals), as well as the subject's hearing thresholds (HT). The signals may be recorded at each ear (*e.g.*, using a microphone mounted on the hearing device), or obtained by convolution of monaural speech and noise signals with HRTFs for the left and right ears. The system's output is the Binaural SII (BSII), an index which provides an objective prediction of speech intelligibility in binaural listening and nonlinear processing conditions. The reference-based model illustrated in Figure 4.1 operates as follows:

- STAGE 1: This stage is performed separately at the left and right ears (denoted by the subscripts L and R respectively in Figure 4.1). The 2x2 Hilbert mixer provides 4 sets of input mixtures, which are processed independently through the hearing device. The phase-inversion method [87] is used to perform speech and noise signal separation with distortion estimation to account for nonlinear processing.

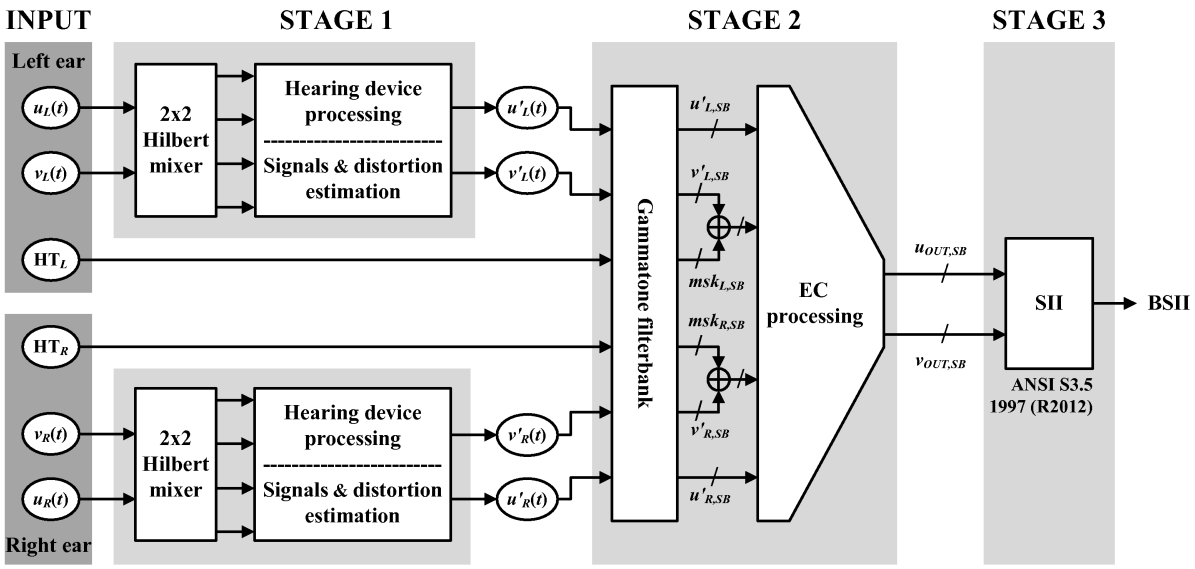


Figure 4.1: *Diagram of the proposed three-stage binaural speech intelligibility measure. The different components in the diagram are explained in detail in subsequent sections.*

- STAGE 2: This stage models the benefits of binaural listening based on the EC theory [43]. It includes a Gammatone filterbank (FB) to process signals into subbands (denoted by the subscript SB in Figure 4.1). Hearing loss is modeled using internal masking signals at each ear. EC processing is performed independently in each subband producing a set of speech and noise signals at the output.

- STAGE 3: This stage performs intelligibility prediction based on the signals at the output of the EC model (denoted by the subscript OUT in Figure 4.1) using the monaural SII (ANSI S3.5–1997 (R2012) [4]).

The details of each stage are presented in the sections that follow. The proposed model is an extension of the binaural measure of Beutelmann and Brand [12], which has been modified to include revisions from [13] and account for hearing device processing.

4.1.1 Stage 1: Signal separation and distortion estimation

This stage is an extension of the phase-inversion signal-separation method proposed by Hagerman and Olofsson [87] (*c.f.* Section 3.2.1.2) and the Hilbert-transform based dis-

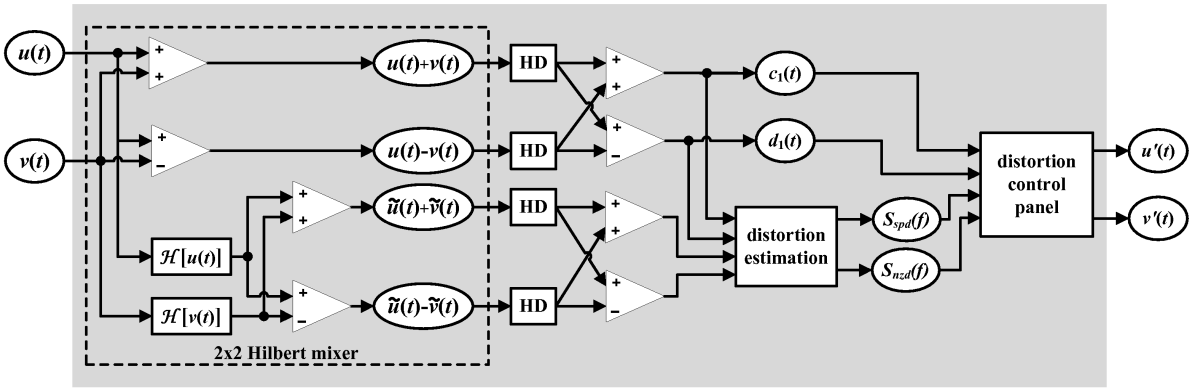


Figure 4.2: *Stage 1: Signal separation with distortion estimation. The diagram shows the details of the 2x2 Hilbert mixer responsible for the four speech and noise mixtures at the input to the hearing device (denoted by the HD block).*

tortion measure of Olofsson and Hansen [154] (*c.f.* Section 3.2.1.3). The phase-inversion method models nonlinear distortion as additive error signals, while the distortion measure has been shown to overestimate the SNR under additive noise conditions (*c.f.* Appendix 2 in [154]). We propose a modified distortion estimation by performing the phase-inversion approach twice as illustrated in the block diagram in Figure 4.2. A total of four measurements are performed using speech and noise mixtures: with and without phase-inversion as well as with and without Hilbert transformation. Given a speech signal $u(t)$ and a noise signal $v(t)$, and their respective Hilbert transform signals $\tilde{u}(t)$ and $\tilde{v}(t)$, we can express the input/output relationships for each of these measurements as follows:

$$a_{in}(t) = u(t) + v(t), \quad a_{out}(t) = u'(t) + v'(t) + e_1(t) \tag{4.1}$$

$$b_{in}(t) = u(t) - v(t), \quad b_{out}(t) = u'(t) - v'(t) + e_2(t) \tag{4.2}$$

$$\tilde{a}_{in}(t) = \tilde{u}(t) + \tilde{v}(t), \quad \bar{a}_{out}(t) = \tilde{u}'(t) + \tilde{v}'(t) + e_3(t) \tag{4.3}$$

$$\tilde{b}_{in}(t) = \tilde{u}(t) + \tilde{v}(t), \quad \bar{b}_{out}(t) = \tilde{u}'(t) - \tilde{v}'(t) + e_4(t) \tag{4.4}$$

where the prime notation ($'$) is used to denote the device's output to a given input signal, while the terms $e_1(t)$, $e_2(t)$, $e_3(t)$ and $e_4(t)$ are used to model nonlinear distortion in each measurement. The overbar notation is used for $\bar{a}_{out}(t)$ and $\bar{b}_{out}(t)$ emphasizing

that these signals are not the Hilbert transforms of $a_{out}(t)$ and $b_{out}(t)$. Next, we recover the following output signals from each measurement by addition and subtraction:

$$c_1(t) = \tfrac{1}{2}[a_{out}(t) + b_{out}(t)] = u'(t) + \tfrac{1}{2}[e_1(t) + e_2(t)] \tag{4.5}$$

$$d_1(t) = \tfrac{1}{2}[a_{out}(t) - b_{out}(t)] = v'(t) + \tfrac{1}{2}[e_1(t) - e_2(t)] \tag{4.6}$$

$$\bar{c}_2(t) = \tfrac{1}{2}[\bar{a}_{out}(t) + \bar{b}_{out}(t)] = \tilde{u}'(t) + \tfrac{1}{2}[e_3(t) + e_4(t)] \tag{4.7}$$

$$\bar{d}_2(t) = \tfrac{1}{2}[\bar{a}_{out}(t) - \bar{b}_{out}(t)] = \tilde{v}'(t) + \tfrac{1}{2}[e_3(t) - e_4(t)] \tag{4.8}$$

As we have seen in Section 3.2.1.2, signals $c_1(t)$ and $d_1(t)$ represent distorted estimates of the speech and noise signals at the output of the hearing device. The PSD of these signals can be expressed as:

$$S_{c_1c_1}(f) = S_{u'u'}(f) + S_{spd}(f) \tag{4.9}$$

$$S_{d_1d_1}(f) = S_{v'v'}(f) + S_{nzd}(f) \tag{4.10}$$

where $S_{spd}(f) = \frac{1}{4}[S_{e_1e_1}(f) + S_{e_2e_2}(f) + 2Re\{S_{e_1e_2}(f)\}]$ and $S_{nzd}(f) = \frac{1}{4}[S_{e_1e_1}(f) + S_{e_2e_2}(f) - 2Re\{S_{e_1e_2}(f)\}]$ represent the PSD of the speech and noise distortions respectively. Our objective is to quantify these distortion terms by investigating the Hilbert-transform relationships between the pairs of signals $\bar{a}_{out}(t)/a_{out}(t)$ and $\bar{b}_{out}(t)/b_{out}(t)$. To do so, we form the complex signals $g(t) = c_1(t) + j\bar{c}_2(t)$ and $h(t) = d_1(t) + j\bar{d}_2(t)$, and estimate their two-sided PSDs:

$$S_{gg}(f) = \begin{cases} 4S_{u'u'}(f) + \overline{S_{spd}}(f) & f > 0 \\ \overline{S_{spd}}(f) & f < 0 \end{cases} \tag{4.11}$$

$$S_{hh}(f) = \begin{cases} 4S_{v'v'}(f) + \overline{S_{nzd}}(f) & f > 0 \\ \overline{S_{nzd}}(f) & f < 0 \end{cases} \tag{4.12}$$

where $\overline{S_{spd}}(f)$ and $\overline{S_{nzd}}(f)$ can be expressed as combinations of the auto- and cross-spectral densities of $e_1(t)$, $e_2(t)$, $e_3(t)$ and $e_4(t)$. If we denote the portions of the PSD of signal $x(t)$ over positive and negative frequencies as $S_{xx}^{+}(f)$ and $S_{xx}^{-}(f)$ respectively, we

can estimate the speech and noise distortion terms in equations (4.9) and (4.10) as:

$$S_{spd}(f) = \frac{1}{4}[S^{-}_{gg}(f) - S^{+}_{gg}(f) + 4S_{c_1c_1}(f)] \tag{4.13}$$

$$S_{nzd}(f) = \frac{1}{4}[S^{-}_{hh}(f) - S^{+}_{hh}(f) + 4S_{d_1d_1}(f)] \tag{4.14}$$

These new estimators may be used as performance indicators for the phase-inversion speech and noise recovery process at each ear. A simulation-based study was conducted early during this doctoral work to research this perspective. The study, which will be presented in Chapter 5, also explored whether the new distortion estimators could be incorporated in a speech intelligibility prediction task. This is further developed in the "distortion control panel", the final block in Figure 4.2. The panel provides various options for distortion modeling, which will be explored in Chapter 7. Future revisions of the model would include an intelligent decision mechanism which selects and/or weighs the contribution of each distortion component to speech intelligibility.

4.1.2 Stage 2: Binaural processing

4.1.2.1 Gammatone analysis filterbank

Prior to EC processing, the signals recovered at the output of Stage 1 are split into a number of frequency bands modeling the behaviour of the peripheral auditory system (*c.f.* Section 2.2.2). This is accomplished using a 21-band Gammatone filterbank designed according to the specification of the SII critical-band procedure outlined in Table 1 of ANSI S3.5–1997 (R2012) [4]. 12^{th}-order filters using Hohmann's all-pole approximation of Gammatone filters [91] were required to achieve the desired specifications. Using this decomposition means that the SII computation in Stage 3 can be performed directly on the subband signals at the output of Stage 2. This represents a significant departure from the model of Beutelmann and Brand who used a 30-band filterbank that does not match the specifications of any of the SII procedures [12]. The differences between the 2 filterbank implementations is thoroughly documented in Appendix A.

4.1.2.2 Internal masking noise

A subject's hearing thresholds can be interpreted as a masked threshold produced by an internal noise acting as a masking sound [217]. Thus, in keeping with Beutelmann and Brand's model, the pure-tone audiogram is modeled by an internal masking noise. First, HT_L and HT_R are interpolated logarithmically over the center frequencies of the analysis filterbank. Then, Gaussian noise (uncorrelated between the left and right ears) is generated within each band with noise energy equal to the energy of a pure tone 4 dB above the hearing threshold at the band's center frequency [21, 217]. The resulting signals ($msk_{L,SB}$ and $msk_{R,SB}$) are added to the noise estimate recovered at each ear.

4.1.2.3 Equalization-cancellation processing

The overall architecture of the EC model presented in Figure 3.5 consists of EC stages carried out independently in each of the auditory filters. For the implementation of each EC stage, we have adopted the same approach as Beutelmann and Brand who used vom Hövel's modified multifrequency channel EC model [206] shown in Figure 4.3. This approach differs from Durlach's original model in the handling of artificial errors, which are included at the back end of the speech detection model instead of before the EC process as described earlier in Section 3.3.3.1. The equalization transform consists of applying delay and attenuation factors such that the maskers are equalized at both ears. The delay factor is applied to the left channel, while attenuation is applied to the channel with the higher noise energy. In the cancellation process the right channel is subtracted from the left. Initial estimates of the delay and attenuation factors are computed as the lag of the cross-correlation maximum and the root-mean-square (RMS) level difference between the maskers respectively. The optimal factors are derived iteratively using an unconstrained nonlinear optimization procedure (Matlab function `fminsearch`) such that the SNR after the cancellation process is maximized. The attenuation and delay factors are derived from the noise maskers first, then applied separately to both speech and noise

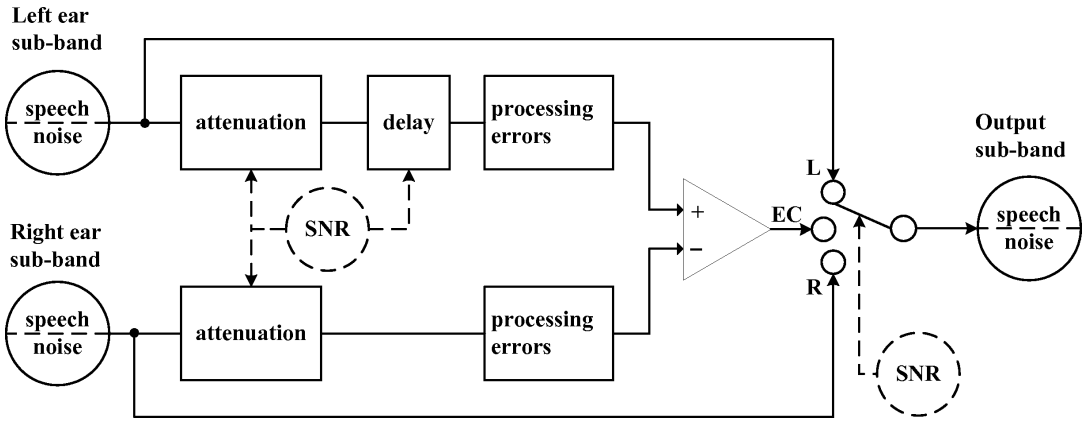


Figure 4.3: *EC-stage subband implementation according to [206]. The SNR is used both as a maximization criterion to find the attenuation and delay parameters, and as an output selector to chose the pair of signals that best contributes to speech detection. (Adapted from [12])*

components. Finally, in keeping with the original EC theory, a decision mechanism at the output of the EC stage in each band selects the pair of speech and noise signals (from the left, right or EC-processed channels) which provides the best SNR.

4.1.2.4 Processing errors

Artificial processing errors for modeling human inaccuracies are incorporated into the EC model using the underlying assumption that speech detection (including EC processing) is carried out in parallel by several processing units, with the final result obtained by averaging the outputs. In each processing unit, pairs of Gaussian-distributed gain (ϵ_l, ϵ_r) and delay (δ_l, δ_r) errors are introduced to the optimal gain (α) and delay (Δ) parameters computed during the EC process in each frequency band. The standard deviations σ_ϵ and σ_δ are defined according to:

$$\sigma_\epsilon = \sigma_{\epsilon 0}[1 + (|\alpha|/\alpha_0)^p] \tag{4.15}$$

$$\sigma_\delta = \sigma_{\delta 0}(1 + |\Delta|/\Delta_0) \tag{4.16}$$

with $\sigma_{\epsilon 0} = 1.5$ dB, $\alpha_0 = 13$ dB, $p = 1.6$, and $\sigma_{\delta 0} = 65\mu$s, $\Delta_0 = 1.6\mu$s. These parameters derived by vom Hövel were found to provide a better fit to subjective BMLD data from [15, 51, 134] than was possible with the original parameters from Durlach [43]. Therefore

a Monte Carlo method is used to model artificial processing errors: 25 sets of Gaussian-distributed random numbers, with the standard deviations defined in equations (4.15) and (4.16), are generated and added to the EC parameters in each frequency band. EC stages and subsequent processing steps are carried out for each set of parameters, resulting in a set of 25 predictions, which are averaged to produce the model's output.

4.1.3 Stage 3: Speech Intelligibility

4.1.3.1 Intelligibility prediction

The computation of the SII has been described in detail in Section 3.1.2.1. In [12], Beutelmann and Brand computed the SII based on the re-synthesized speech and noise signals at the output of the EC stage using the 18-band one-third octave-band procedure. The SII predictions are computed based on a Matlab implementation made available by Müsch [151]. In our model, the Matlab code for the SII has been adapted to the 21-band critical-band procedure, and the standard band-importance function is used (Table 1 in [4]). The SII is computed directly from the subband signals at the output of the EC model (*c.f.* Section 4.1.2.1). Since these signals are derived from inputs obtained by convolution with HRTFs, no freefield-to-eardrum level adjustment was necessary. Finally, since the hearing thresholds have been accounted for at an earlier stage (*c.f.* Section 4.1.2.2), the thresholds are set to 0 dB HL during the SII computation.

4.1.3.2 Intelligibility mapping

In order to compare the model predictions with subjective SRT data, the psychometric function described in equation (3.5) is used with the parameters derived to fit subjective scores for the HINT test (*c.f.* Section 3.1.2.3). The SII-to-SRT mapping is performed by iteratively calculating the SII until we reach the SNR that yields a target SII value (*e.g.*, 0.375, *c.f.* Section 3.1.2.3) with a tolerance of 0.03 corresponding to the SRT for 50% intelligibility. This is done using a "fast-search" procedure summarized below:

- Begin with a given SNR and a step size of 4 dB.
- Compute the BSII and start iterations:
  - If the BSII is above the target value, increase the SNR by the step size.
  - If the BSII is below the target value, decrease the SNR by the step size.
  - The step size is halved each time a reversal is encountered, or set to 1 dB if the target value is reached within a specified tolerance.
  - Compute a new BSII value at the new SNR and reiterate.
  - Iterations are stopped when the step size reaches 1 dB.
- Interpolate (or extrapolate) to obtain the SRT with a precision below 1 dB.

4.1.4 Short-time extension

The ESII measure has been proposed by Rhebergen and Versfeld in order to deal with situations involving fluctuating maskers [170] (*c.f.* Section 3.1.2.1.1). In this thesis, short-time processing is considered not only to deal with non-stationary maskers, but also to account for distortion fluctuations in time-adaptive processing systems (*e.g.*, fast-acting compression systems, and noise reduction algorithms). In this section, we discuss the necessary considerations for a short-time extension of the proposed measure. This includes additional signals based on the HINT material, implementation details for frame-by-frame processing, and their effect on intelligibility prediction.

4.1.4.1 Additional HINT signals

The speech-shaped noise (SSNOISE) signal included in the HINT material is the optimal masker. It is a continuous[1] signal with a long-term spectrum equivalent to that of an av-

[1]In the Engineering literature, it is common to distinguish between signals with constant statistical properties (stationary) and those with statistical properties that change over time (non-stationary). In hearing research, it is more common to distinguish between continuous and fluctuating signals on the basis of changes in the signal envelope.

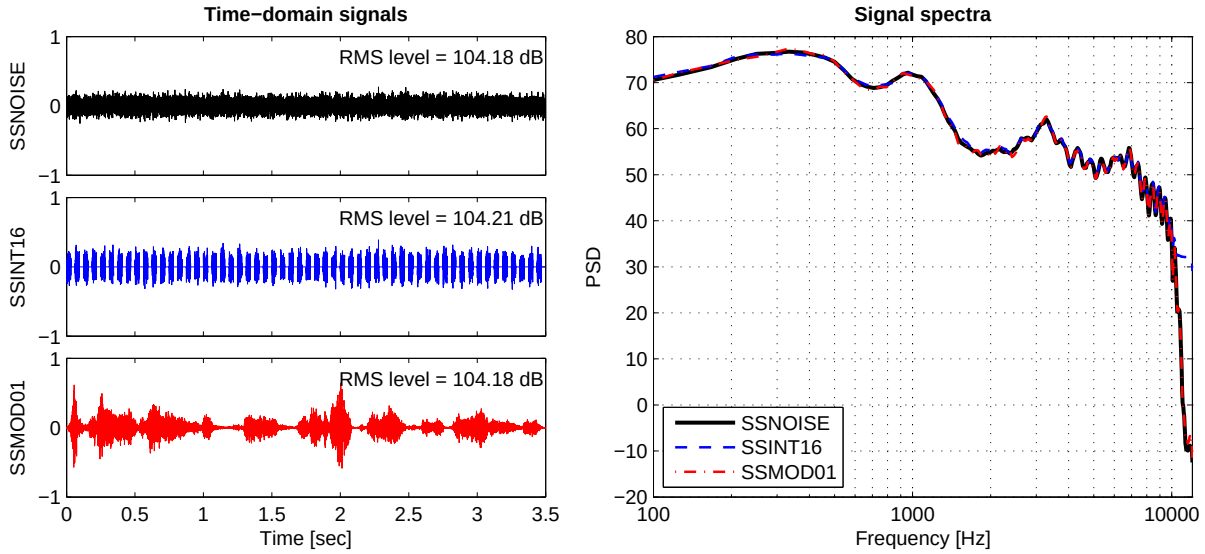


Figure 4.4: *Comparing signals SSNOISE, SSINT16 and SSMOD01. The time-domain plots also display the signals' RMS power (full-scale dB reference: 127.5 dB). The frequency-domain plots of all three signals overlap over the entire frequency range used in SII computations.*

erage HINT sentence list. However, real-life listening situations are generally easier than when sound is masked by continuous noise, since listeners are able to extract additional speech information during periods when the masker level is low. Non-stationary maskers such as intermittent, amplitude-modulated, or speech-modulated noise signals have been frequently used to simulate real-life background noises [172]. Given that intermittent maskers provide the most listening advantage [170, 171], we introduce the intermittent speech-shaped noise (SSINT16) when testing subjects and generating model predictions under non-stationary masking conditions. The SSINT16 masker has been generated by multiplying the SSNOISE signal by a 16 Hz square wave with a 50% duty cycle [174]. This rate of interruption has been shown to produce the largest difference in improvements between normal-hearing (NH) and hearing-impaired (HI) subjects [183]. Together, these two signals (SSNOISE and SSINT16) provide the best and worst-case masking scenarios allowing us to test the limits of the model's applications.

Next, we introduce the single-talker speech-modulated noise (SSMOD01) used as the speech signal when generating model predictions. Using this type of signal helps capture

fluctuating distortion effects typically applied by adaptive processing systems acting on speech signals. The SSMOD01 signal was generated by time-reversing the SSNOISE signal (to eliminate correlation with the masking signals) and multiplying it by the envelope of two concatenated HINT sentences: HINT134 and HINT163. These two sentences could be thought of as having "average difficulty" among the HINT sentences, since their RMS power and spectra deviated the least from the SSNOISE signal. The length of the resulting signal is 3.5 seconds, which falls within the range of 2 to 4 seconds investigated by Rhebergen and Versfeld [170] (*c.f.* Section 4.1.4.2). Time-domain and frequency-domain plots of these signals are shown in Figure 4.4.

4.1.4.2 Block processing

The rationale for computing the ESII [170] consists of dividing the masking signal into short windows and using a frame-based approach to compute the final index: within each frame the standard SII is computed using the long-term speech spectrum and the short-term spectrum of the masker. Rectangular windows are used to divide the signal with frequency-dependent window lengths ranging from 35 ms at the lowest frequency band to 9.4 ms at the highest. Thus, a new SII value is predicted every 9.4 ms and the ESII is computed by averaging over a duration of at least 2 seconds (Rhebergen and Versfeld [170] reported that this duration gives a between-sample standard deviation of 0.0056 for the SII, and that increasing it to 4 seconds reduces the standard deviation to 0.0030). In addition, the authors noted that predictions with a simpler model using fixed-length windows of 12 ms would fit their subjective SRT data, however not as well as the model with frequency-dependent windows. The model was later revised to include a forward-masking function in [171].

In our model, block processing is performed using 576-point Hanning windows with 50% overlap. At a sampling rate of 24 kHz, each frame corresponds to a 24 ms window. Since the equivalent rectangular duration for the Hanning window is half its length, the resulting effective duration of 12 ms matches the fixed-duration window suggested

in [170]. Beutelmann et al. found that using a similar window in their revised model was successful in predicting binaural SRTs in the presence of non-stationary maskers [13]. A second implementation using 12 ms non-overlapping rectangular windows (288 points) has also been included for comparison. A more advanced implementation which incorporates frequency-dependent windowing and a forward masking function as in [171] could be considered for future extensions.

4.1.5 Summary and verification of model implementations

Two implementations of the proposed model have been completed: a single-frame version and the short-time version. The details of these implementations and the design choices involved have been discussed in the previous sections. In this section, we provide additional information to finish documenting the model. First a summary of the model's assumptions and requirements is presented in Section 4.1.5.1. Next we document the possible effects on intelligibility prediction of some of the novelties introduced in the model in Section 4.1.5.2. A discussion of the phase-inversion approach and the associated distortion estimation procedure used in Stage 1 of the model has been published in [57, 58] and is presented in Chapter 5. Thus the discussion in this section will focus on the short-time extension of the model (*c.f.* Section 4.1.4) as well as the filterbank implementation used in Stage 2 (*c.f.* Section 4.1.2.1). Investigating these effects in a monaural context, allows us to verify these components independently of any binaural processing. Finally, alternative implementations of the model that have been investigated over the course of this doctoral work are briefly described in Section 4.1.5.3.

4.1.5.1 Model requirements and assumptions

The proposed model is designed to predict speech intelligibility under various binaural listening situations and nonlinear processing conditions. In doing so, it overcomes several limitations of the standard SII [4], on which it is based. However, the model still requires access to the separate speech and noise signals in order to generate the 4 signal mixtures

at the input to each hearing device. Four passes through the hearing device at each ear are required in order to estimate the speech, noise, and distortion components at the output. As such, the model is an offline tool for intelligibility prediction, and is not intended for a real-time application. Moreover, it is assumed that the device under test is time-invariant. For devices with time-varying signal processing algorithms, care must be taken to avoid memory effects between the different measurements required: this could range from allowing a sufficient amount of time for the system's response to a previous signal mixture to die out, to disabling features in the device that maintain a longer term memory (*e.g.*, feedback cancellation).

Furthermore, it is assumed that the perceptual byproducts of nonlinear processing can be modeled by additive uncorrelated noise signals. This assumption stems from Hagerman and Olofsson's phase-inversion model which was briefly validated in [87] and is further investigated in Chapter 5. Finally, the standard SII assumes hearing loss to be modeled by an internal masking noise consisting of the hearing thresholds and a reference internal noise (*c.f.* equation (10) in [4]). This assumption still holds in the proposed binaural model, however, the hearing thresholds are now modeled by internal masking noise at each ear independently at the input of the EC model. Supra-threshold deficits (*c.f.* Section 2.3.3) have not been accounted for in the model at this point.

4.1.5.2 Monaural verification of modeling components

Plots of SII and ESII predictions obtained using SSNOISE and SSMOD01 as speech under continuous and intemittent masking are shown for a wide range of SNRs in Figures 4.5 and 4.6. Several observations concerning the short-time extension of the model can be made from these plots. First, the ESII predictions obtained using SSNOISE and SSMOD01 as speech signal overlap over the whole SNR range as expected, since the two signals have the same long-term spectrum. The same overlap was also observed between the SII predictions; however, the SII values using SSMOD01 have been omitted from the plots for visual clarity.

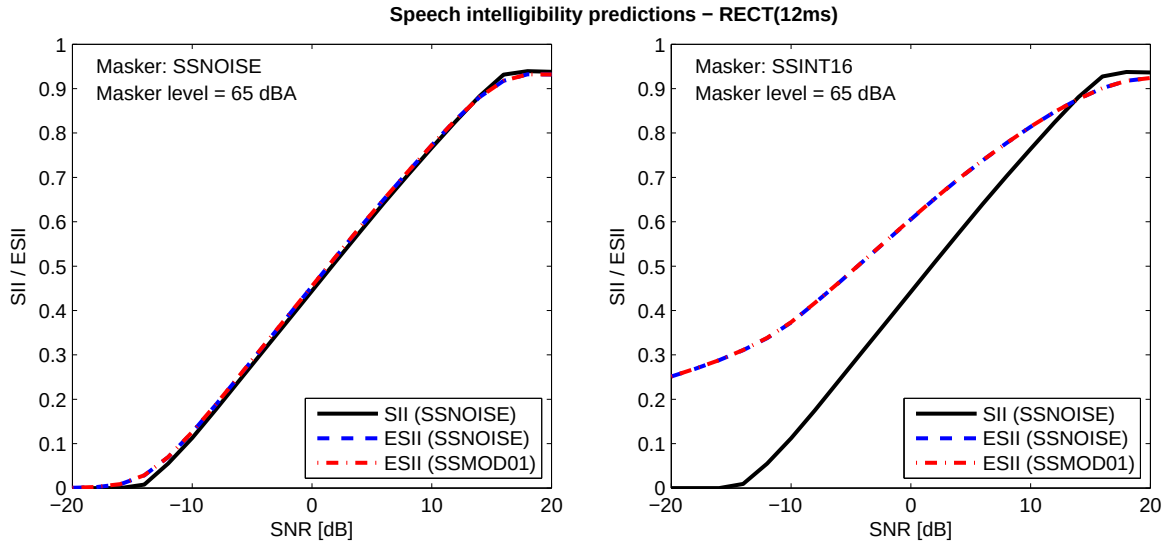


Figure 4.5: *SII and ESII prediction using SSNOISE and SSMOD01 under continuous (left panel) and intermittent (right panel) masking. ESII predictions use non-overlapping rectangular windows. ESII predictions with SSNOISE and SSMOD01 overlap over the entire range of SNR.*

Next, we compare the SII and ESII predictions in Figure 4.5. As expected, the ESII is equivalent to the SII under continuous masking (left panel), over the whole SNR range. Small deviations occur around ±15 dB as the SII procedure accounts for level distortion by limiting the difference between the speech and noise spectra. In the ESII computation, some segments will exceed the limit and others will not, resulting in the difference observed between the ESII and SII values. However, it is under intermittent masking (right panel) that the effect of block processing in the ESII is more clearly observed. The audibility of speech during the gaps in the SSINT16 masker leads to higher ESII values, except at high SNR where level distortion effects occur. The difference between ESII and SII predictions is greatest at very low SNR conditions where the speech is entirely masked by the noise when present, and audible during silent segments. The results shown in the right panel of Figure 4.5 are very similar to those reported by Rhebergen and Versfeld for a 10-Hz intermittent masker (*c.f.* FIG6 in [170]). Conversely, the SII predictions obtained with the SSNOISE and SSINT16 maskers are the same, since the SII only uses the long-term spectrum of the masker.

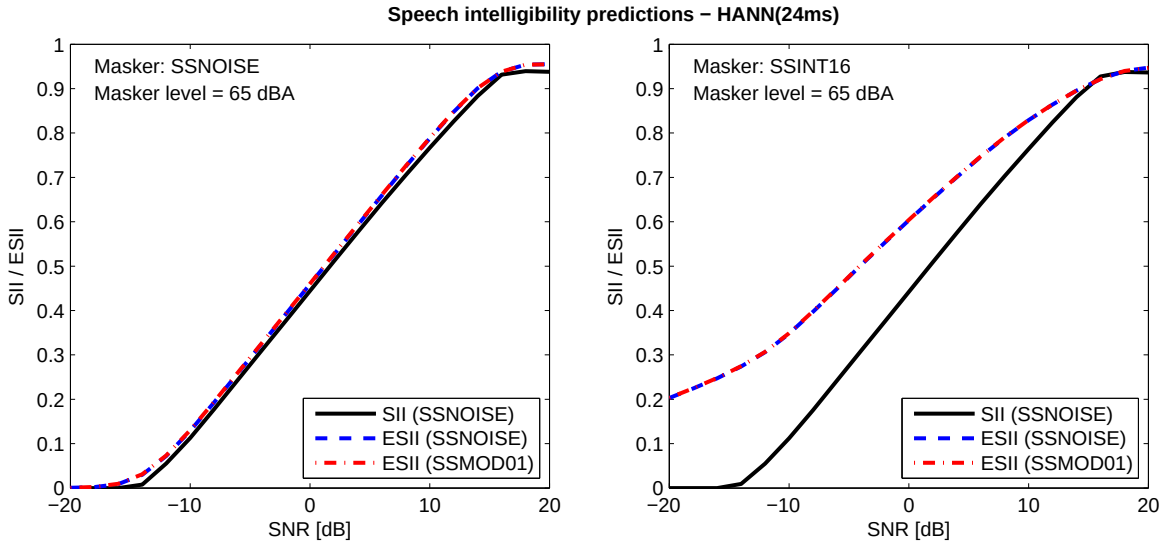


Figure 4.6: *SII and ESII prediction using SSNOISE and SSMOD01 under continuous (left panel) and intermittent (right panel) masking. ESII predictions use overlapping Hanning windows. ESII predictions with SSNOISE and SSMOD01 overlap over the entire range of SNR.*

Next, we consider the effect of the two framing procedures by comparing the corresponding plots in Figures 4.5 and 4.6. Computing the ESII with rectangular windows gives a closer match to the standard SII under continuous masking (SSNOISE), although the differences are very small: Hanning windows seem to produce slightly larger predictions at high SNR levels. For the intermittent masker (SSINT16), the plots for rectangular windows are more similar to those reported in [170], especially at very low and high SNR conditions. It is worth noting, however, that those predictions were correlated with subjective measurements only at the SRT. It remains to be seen which of the two windowing methods will produce predictions that better match our own subjective SRT measurements, especially when the binaural aspects of the model are taken into considerations. This will be investigated in Chapter 7.

Finally we consider the effect of filterbank implementations. Figure 4.7 shows a comparative plot of SII and ESII predictions computed using: (1) the critical-band SII/ESII subband decomposition, (2) the 21-band critical-band filterbank implementations, and (3) the 30-band filterbank from [12]. In the latter case, the filterbank was used to perform

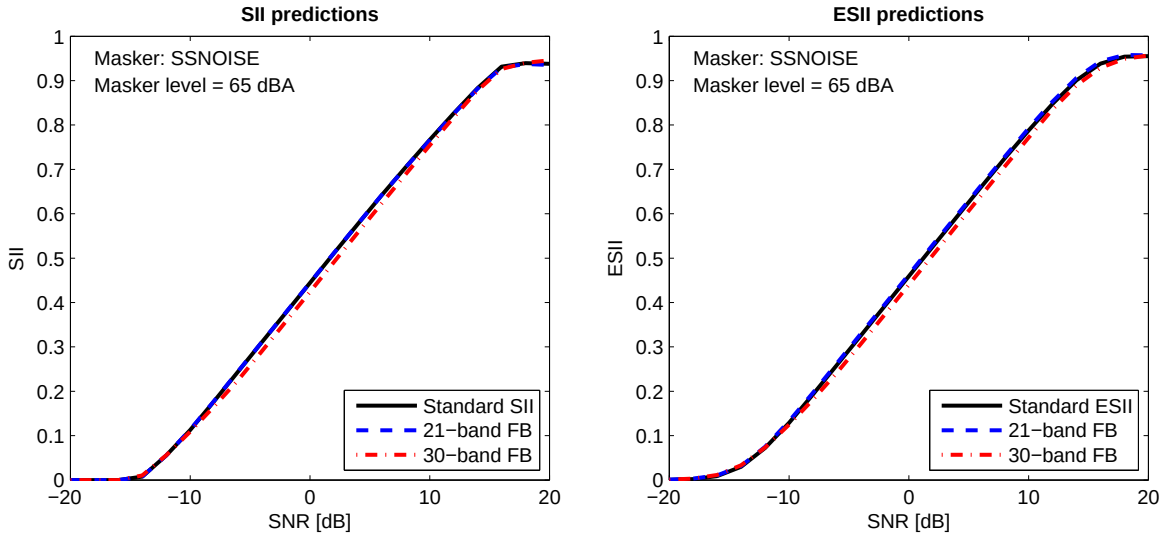


Figure 4.7: *SII (left panel) and ESII (right panel) prediction using the standard and filterbank (FB) implementations. The predictions are obtained using the SSNOISE signal as speech under continuous masking conditions. The ESII predictions shown use overlapping Hanning windows.*

analysis and synthesis, then the re-synthesized signals were used with the critical-band computational procedure. While the predictions obtained with both filterbank implementations differ only slightly, the proposed 21-band filterbank offers a closer experimental match to the predictions obtained with Müsch's implementation.

4.1.5.3 Alternative model implementations

Two alternative implementations of the EC process have been investigated at earlier stages of this work. The first, considers an implementation of the equalization transformation based on Wiener filters and was published in [54]. The implementation omits the processing errors associated with the EC model, but shows that a Wiener-based model can be a promising solution with low computational cost. However, no further development was carried out in this direction as it raises a difficult research challenge: the coefficients of the Wiener filter are not directly related to the acoustic concepts being modeled in the way that the attenuation and delay parameters are; this makes the artificial processing errors more of a challenge to quantify.

The second alternative implementation considered a representation of artificial processing errors using neural networks and was developed as a course project in the Winter term of 2009 [52]. The paper shows that a neural-network-based approach is a promising solution that provides a single-pass alternative to the iterative optimization routines currently used in the EC model. Extending this implementation further would require a significant amount of training data relating internal representations of the model to speech intelligibility in order to train the neural network. Since this data is not available at this point, no further development was carried out with this approach. Once the proposed model has received sufficient validation, this type of training data can be generated and the approach would then be an interesting extension to the proposed model which would significantly improve computational speed.

In short, both alternative implementations showed a promising outlook. However, no further development of these methods is investigated in this thesis.

4.2 Hearing-device processing

A hearing device under test is required to evaluate the model presented in Section 4.1. In this work, signal processing is performed using an in-house hearing device simulator implemented in Matlab. The simulator provides a set of typical functions commonly performed in real hearing devices thus providing an efficient test bench. A version of the simulator was available within our group [158]. Extensions were needed in order to adapt it to be used in this work to evaluate the binaural speech intelligibility predictor.

4.2.1 The hearing-aid simulator toolbox (HASTB)

The hearing aid simulator is a monaural tool which implements a virtual hearing aid and a virtual user. The user's profile includes the subject's hearing thresholds, UCL, and MCL. Figure 4.8 shows the architecture of the hearing aid simulator. The virtual hearing aid implements typical hearing aid functionality including:

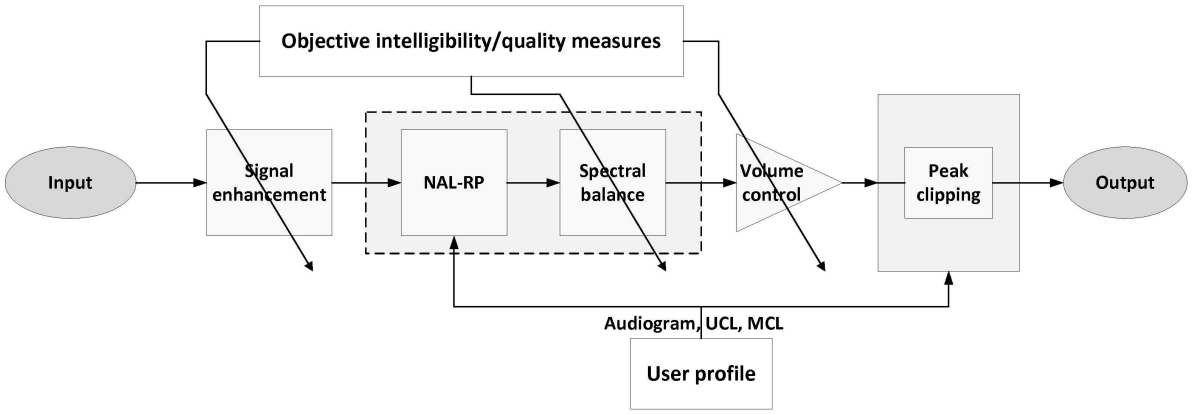


Figure 4.8: *Schematic diagram of the monaural hearing aid simulator. The NAL-RP and spectral balance circuits are combined into a single filter for computational efficiency.*

- A signal enhancement module which acts on the input signal and can be a placeholder for any off-the-shelf single-channel denoising algorithm. Only one algorithm [6] was included in this module before its functionality was extended.
- The NAL-RP linear hearing-loss prescription [26] based on a subject's audiogram.
- A spectral balance function to allow a user to emphasize low or high frequencies using a spectral tilt with different slopes.
- A volume control to increase or decrease the output volume.
- An output limiter which performs peak-clipping operation at the subject's UCL.

The simulator is packaged as a toolbox (HASTB v1.4 [158]) which, in addition to the hearing aid functionality, includes speech intelligibility and quality predictors using Kates and Arehart's three-level approach [124, 125] (*c.f.* Section 3.2.2.1). Finally, the toolbox also provides Matlab routines to find the optimal control settings (*i.e.*, volume control, spectral tilt, noise reduction strength) that maximize speech intelligibility and comfort for a user with a given hearing loss in a particular listening situation [168].

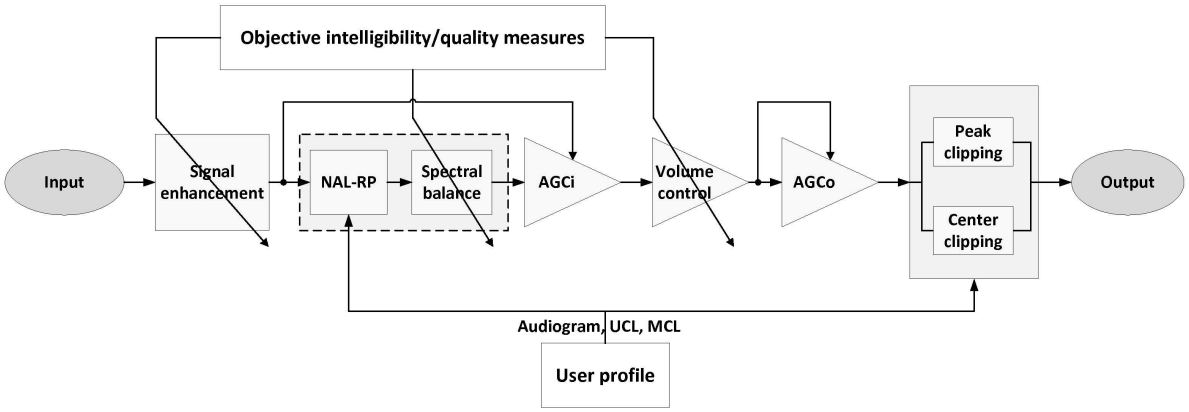


Figure 4.9: *Schematic diagram of the extended hearing device simulator.*

4.2.2 The hearing-device simulator toolbox (HDSTB)

Calling into perspective that a hearing aid is a particular type of hearing device, a major revision of the HASTB was completed to extend its functionality to a more general device. The revised simulator has been renamed the hearing device simulator and packaged in a new toolbox (HDSTB v2.0). Thus the hearing device simulator could represent a hearing aid, a communication headset, a wearable bluetooth receiver, or a pair of active hearing protectors under test, to give a few examples. Figure 4.9 shows the architecture of the hearing device simulator. The extensions included in this current revision include additional NRAs, two compression amplifiers, and a selection of clipping options.

4.2.2.1 Noise reduction

The wavelet denoising algorithm of Bahoura and Rouat [6] included in the signal enhancement module of the original HASTB is a powerful and fairly recent algorithm. Its ability to improve performance in Hidden Markov Model speech recognition experiments has been documented [6]. However, to the best of our knowledge, no data on subjective speech intelligibility using this algorithm have been published to date. Therefore it was fitting to expand the library of algorithms available in the noise reduction module. In addition to the existing wavelet algorithm (WTHR-TSA), Table 4.1 lists the algorithms

Table 4.1: *List of NRAs included in the extended hearing device simulator. The newly added algorithms can be categorized as spectral-subtraction or Bayesian based algorithms.*

Abbr.	Description	Ref
WTHR-TSA	Wavelet thresholding based on time-scale adaptation	[6]
SSboll79	Amplitude spectral subtraction	[17]
SSberouti79	Power spectral subtraction (a posteriori SNR)	[11]
SSscarlat96	Power spectral subtraction (a priori SNR)	[178]
SSPARAM98	Parametric spectral subtraction	[181]
MBSS02	Multiband spectral subtraction	[116]
WFscarlat96	Wiener filter using decision-directed SNR estimation	[178]
MMSE-STSA84	MMSE short-time spectral amplitude estimator	[60]
MMSE-logSTSA85	MMSE log-spectral amplitude estimator	[61]
MMSE-STSA02	MMSE STSA with Gamma-distributed speech priors	[145]
MMSE-logSTSA04	MMSE logSTSA with non-causal a priori SNR estimator	[30]

which were added. These may be classified as algorithms based on spectral subtraction or Bayesian statistical-model methods [203]. Spectral subtraction algorithms operate by dividing a signal into short frames and subtracting an estimate of the noise spectrum from the noisy speech spectrum. The noise estimate is obtained from portions of the signal where only noise is present, and it is assumed that the noise is stationary or slowly varying. Spectral subtraction algorithms tend to produce a particular type of distortion known as "musical noise", which arises primarily from fluctuations in the noise spectrum across frames. Different algorithms have been developed with the goal of reducing or masking the residual musical noise. Bayesian methods use the signal's probability density function to improve on spectral subtraction algorithms by minimizing a cost function. A popular class of these algorithms is one which uses the minimum mean-square error (MMSE) of the short-time spectral amplitude (STSA) as a cost function.

Matlab implementations of the newly added algorithms were obtained from [214]. The implementation requires a short segment of initial silence (noise-only) at the beginning of the signal to estimate the noise spectrum, and a VAD to update and smooth the estimate during silence periods within the signal.

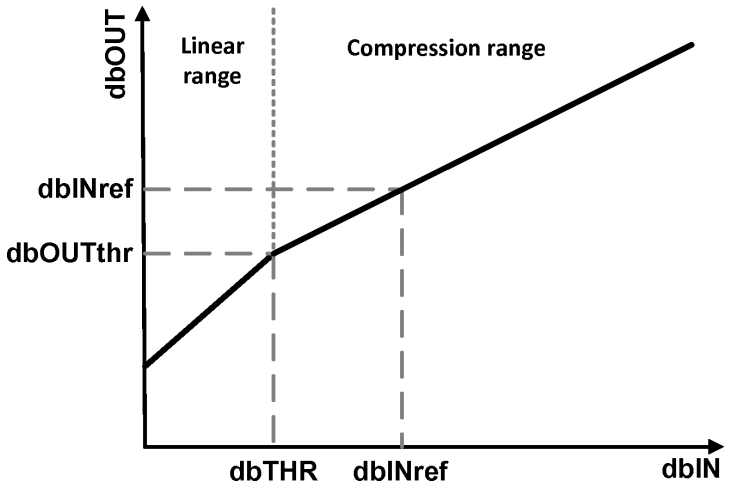


Figure 4.10: *Input/output function of the AGC amplifier in the hearing device simulator.*

4.2.2.2 Compression

An essential extension to the HASTB is the implementation of a module to perform amplitude compression: that is amplification which depends on the level of the input signal (*c.f.* Section 2.4.3). A generic single-channel AGC amplifier, or compressor, was designed with the following parameters: compression ratio (CR), compression threshold ($dbTHR$), input zero-gain reference level ($dbINref$), attack time (T_a) and release time (T_r). The first 3 parameters define the compressor's static behaviour and are used to compute the required target gain as illustrated in the input/output function in Figure 4.10. The computation of the target gain is derived as follows (all levels expressed in decibels): First the gain at threshold ($gainTHR$) is computed as:

$$gainTHR = (dbINref - dbTHR) * (1/CR) \tag{4.17}$$

This gain is applied to any signal that falls in the linear range, *i.e.*, below the compression threshold. Within the compression range, the relationship between the input level ($dbIN$) and output level ($dbOUT$) can be written as:

$$dbOUT = (dbIN - dbTHR)(1/CR) + (gainTHR + dbTHR) \tag{4.18}$$

Therefore, the target gain for a given input level can be computed as:

$$\begin{aligned} targetGain &= dbOUT - dbIN \\ &= gainTHR + [(dbTHR - dbIN)(1 - 1/CR)] \end{aligned} \tag{4.19}$$

The attack and release times define the compressor's dynamic behaviour: *i.e.*, to determine how fast (or slowly) the target gains are reached. This is implemented using a 1 ms exponential-decay accumulator to track the input signal level, compute the target gains, then exponentially increase or decrease the current gain to reach those targets.

The generic compressor is incorporated twice into the hearing device simulator: once as input-controlled (AGCi) and once as output-controlled (AGCo) as shown in Figure 4.9. The AGCi system is placed before the volume controller. It is activated by the signal level before the NAL-RP/spectral balance filter, but acts on the filtered signal. The AGCo system is placed after the volume controller and is activated by the signal level at its input. Finally, all the parameters of the generic compressor are tunable, making it possible to add any number of compression amplifiers to the system.

4.2.2.3 Clipping

Noise reduction and compression algorithms are signal-adaptive processes that rely on smoothly tracking the level of the input signal. Clipping algorithms act instantaneously on each sample in the signal and, thus, tend to produce audible distortions. Two forms of symmetric clipping have been implemented in the hearing device simulator: peak and center clipping. Peak clipping often results from saturation of amplifiers or transducers in a device. Center clipping can be thought of as a crude approximation of the operation of NRAs that tend to reduce the amplitude of low-level signals typically associated with background noise, but also with unvoiced segments of speech (consonants), while keeping the higher-level signals associated with voiced sounds (vowels). The harsh distortions clipping can introduce are scarcely encountered in modern hearing aids which incorporate adaptive circuitry to process signals. However, they may occur in communication systems

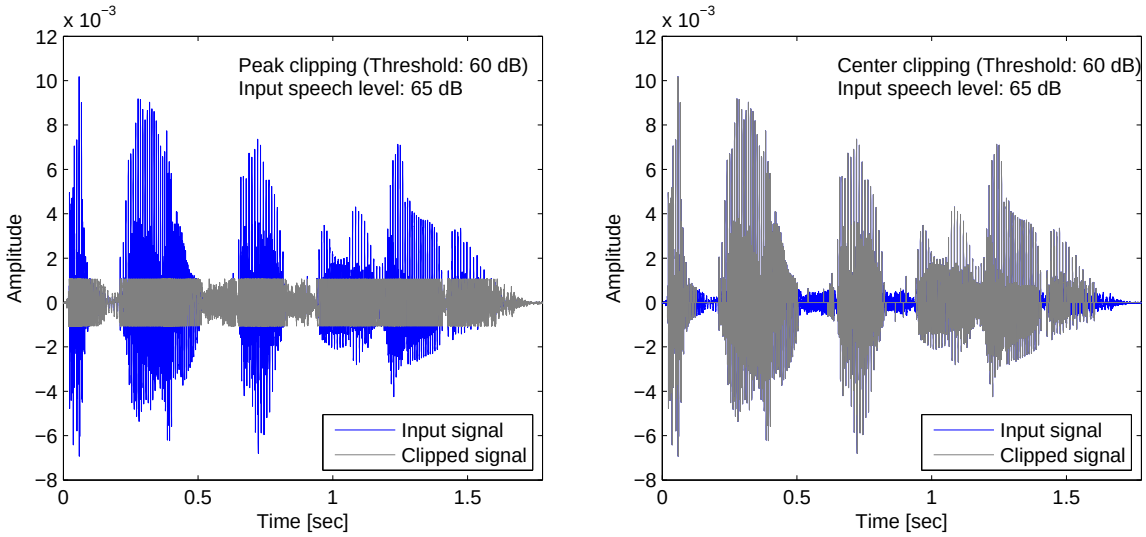


Figure 4.11: *Illustration of peak and center clipping operations.*

or hearing devices that do not use such algorithms.

For an input $x[n]$ and output signal $y[n]$, symmetric peak clipping with a threshold thr is defined mathematically by the following operation:

$$y[n] = \begin{cases} thr & x[n] > thr \\ x[n] & -thr \leq x[n] \leq thr \\ -thr & x[n] < -thr \end{cases} \tag{4.20}$$

Similarly, symmetric center clipping is defined as:

$$y[n] = \begin{cases} x[n] & |x[n]| > thr \\ 0 & |x[n]| \leq thr \end{cases} \tag{4.21}$$

Figure 4.11 illustrates these two forms of clipping applied to a pure speech signal taken from the HINT sentence material. The signals have been scaled to 65 dB, and the clipping threshold set to 60 dB. Two observations can be made from this example. On one hand, peak clipping tends to significantly reduce the energy in the output signal as voiced sounds get clipped. On the hand, center clipping tends to preserve the signal energy, but completely eliminates the unvoiced segments of the speech signal occurring

towards the beginning and ends of utterances. Perceptually, in the absence of noise (such as in this example), a listener can still recognize the words in the processed sentences. The peak-clipped signal sounds more distorted than the center-clipped signal. However, in the case of center clipping, intelligibility tends to degrade at the end of each syllable. These types of effects make clipping conditions interesting to investigate in the presence of noise in the context of the speech intelligibility model.

Finally, three different options are available in the hearing device simulator to specify the clipping threshold: this can be (1) an absolute level specified in decibels as in the above example, (2) a percentage of the signal's cumulative magnitude histogram, or (3) an offset from the subject's UCL specified in decibels. The same options have also been made available for the specification of the compression threshold in the AGCo system (*c.f.* Section 4.2.2.2).

4.2.3 Binaural extension

The hearing device simulator also includes a binaural extension which consists of combining two instances of the monaural virtual hearing device, one for each ear, in order to simulate a binaural listening condition. Key to this extension is the use of head-related transfer functions to represent the transformation of signals from the source to the left and right ears.

The set of HRTFs used in this work were recorded on a KEMAR [25] at the MIT Media Lab [73, 74] and made available for download [72]. The data were measured for different azimuths (covering 360°) and elevations (covering the range from −40° to 90°). At an elevation of 0°, the data is given in 5° increments in azimuth. All HRTFs are recorded as 16-bit data (dynamic range of 96 dB) in a conservatively-measured 65 dB SNR and sampled at 44.1 kHz. Furthermore, the KEMAR on which the data was recorded is equipped with a "normal" pinna on the left and the "large red" pinna model on the right. When reading a pair of left-right HRTFs care must be taken to use a pair of HRTFs at complementary angles corresponding to the same pinna. For example, when

simulating a source located 45° to the right (0° elevation), the HRTFs corresponding to the left pinna at angles 45° and 315° are used to obtain the signals at the left and right ears respectively (the opposite angles need to be used if reading the data corresponding to the right pinna). The downloadable package provides two sets of HRTFs, both of which have been included in the HDSTB:

- A "full" set which contains the 512-point HRTFs along with the responses of a set of 4 different headphones and the Optimus Pro 7 speaker used when recording the HRTFs. The speaker has a non-uniform response. An inverse filter and a minimum-phase inverse filter for the speaker are provided to compensate for this.

- A "compact" set which contains 128-point HRTFs equalized to compensate for the Optimus Pro 7 speaker. The data-reduced HRTFs were derived from the the left pinna of the full set by convolution with the minimum-phase inverse filter. They are provided as a stereo pair for left and right ear responses for a given source position. This particular set of HRTFs has been employed throughout this work.

A linear adjustment (10 dB for the "full" set, and 7 dB for the "compact" set) was applied to the HRTFs prior to convolution with the input signals to obtain the binaural signals. The adjustment was derived such that the low frequencies of the frontal HRTF (0° elevation, 0° azimuth) rolled off towards 0 dB, consistent with the data from [180]. The same adjustment is applied to HRTFs at all azimuth and elevation angles so as not to affect the interaural relationships. Moreover, since both speech and masker signals are convolved with the same HRTFs, the adjustment has no effect on the SNR at each ear. Finally, it is worth noting that a bilateral approach is simulated here, whereby signals are processed independently by the hearing device at each ear. An advanced binaural hearing device simulator test bench would include processing algorithms such as binaural speech enhancement or compression systems. Implementing such advanced strategies into the hearing device simulator is deemed beyond the scope of this thesis, which focuses on methods of modeling speech intelligibility.

4.3 The Matlab speech testing environment

The commercial HINT platform does not offer the flexibility to incorporate signal processing using external Matlab routines. In order to subjectively validate the intelligibility predictions generated with the proposed model, it was necessary to build a system that runs a testing procedure based on the HINT while processing sounds with a variety of algorithms prior to presentation to test participants. The Matlab speech testing environment was designed to have an easy-to-use graphical interface shown in Figure 4.12. The interface allows access to all parameters needed to specify a virtual listener and hearing device (using the HDSTB), along with controls to run the speech test. The HINT practice and test sentence lists were made available strictly for research use.

The Matlab program is designed to permit a range of experimental conditions. Its flexible implementation allows future extensions to include additional languages and processing conditions. To date, the application includes the following features:

- Testing with speech and masker sources located at any azimuth/elevation specification, including two diffuse-masker specifications.
- Testing using the American English and French Canadian HINT sentences.
- Testing under both monaural or binaural listening conditions.
- Two types of maskers: continuous (SSNOISE) and intermittent (SSINT16) noise.
- Different hearing devices as per Section 4.2, with the ability to add new ones.
- Three testing paradigms: (1) Fixed speech level with a adaptive masker level, (2) Fixed masker level with a adaptive speech level, and (3) Fixed speech and masker levels with a adaptive distortion threshold.
- Ability to specify test experiments with independent conditions and populations.
- Ability to calibrate the system using a dedicated Calibration screen.

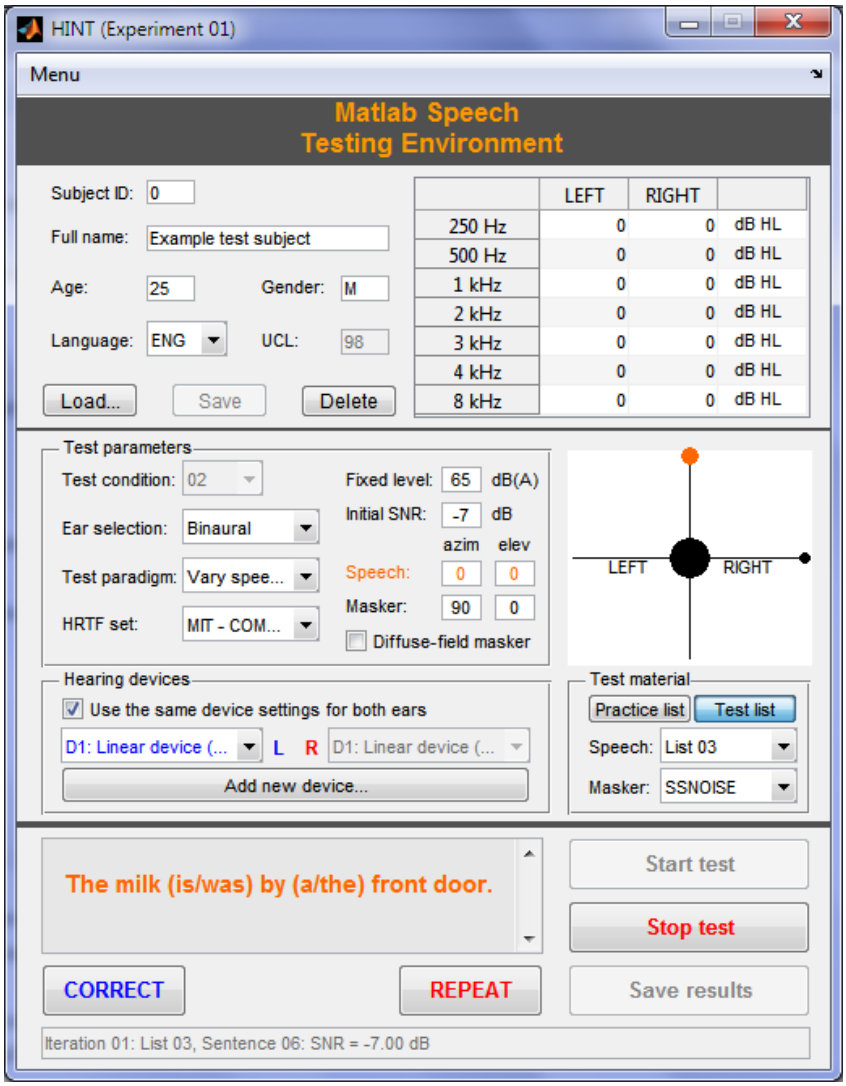


Figure 4.12: *Graphical user interface for the Matlab speech testing environment.*

4.4 Validation approach

The model described in this thesis presents the potential for a wide range of applications involving different listening situations and processing conditions. The variety of conditions could include: multiple spatial sound-source configurations, different signal processing algorithms, a wide range of background maskers, and the consideration of NH and HI subjects. On the one hand, the number of processing steps and novelties incor-

porated into the model require testing with as wide a range of conditions as possible. On the other hand, speech tests impose limitations on the number of test conditions due to the number of available lists of sentences within the speech corpus, and the length of each test session. Therefore, to be feasible within a reasonable time-frame, our strategy to validate the model will focus on a subset of representative experimental conditions. In the remainder of this section, we break down the validation task in two steps: monaural and binaural validation.

4.4.1 Monaural validation

4.4.1.1 Rationale

The goal of the first step is to validate the phase-inversion signal-separation method and the associated Hilbert-based distortion estimation approach used in the nonlinear component of the model (Stage 1). This step will not involve any subjective testing, but will rather consist of a theoretical validation in a monaural context based on the internal representation of signals in the phase-inversion signal-separation model. This was completed at an earlier stage of the work as a simulation-based study. It serves as a proof of concept demonstrating (1) the suitability of the Hilbert-based procedure in capturing the distortion components produced by typical hearing aid processing algorithms, (2) the benefit of using an approach that distinguishes between speech and noise distortion, and (3) the potential incorporation of the distortion estimators in intelligibility prediction.

4.4.1.2 Simulated conditions

The study was conducted by simulating one NH subject and three HI subjects with various degrees of hearing loss. Simple listening conditions were considered in this step using a continuous speech-shaped masker without simulating spatial distribution. In order to test the distortion estimators, we focused on the following four types of hearing aid processing algorithms, which were included in the hearing device simulator at the

time:

- Linear processing (including the NAL-RP prescription for HI subjects)
- Inter-syllabic compression (AGCi).
- Compressive output limiting (AGCo).
- Hard output limiting (peak clipping)

We report on this work in Chapter 5, which has been published in [58] as a revision of [57]. This contribution adds substantial support to Hagerman and Olofsson's signal-separation model [87], which has been used extensively in the research literature with only limited validation.

4.4.2 Binaural validation

4.4.2.1 Rationale

This step consists of conducting a series of experiments to collect subjective intelligibility data and comparing this data with model predictions generated under the same experimental conditions. This is an iterative process in which the measured SRT is used as a starting SNR for the speech and noise mixture at the input of the model. With each iteration, the model outputs a prediction value, the BSII, and the SNR is adaptively adjusted, according to the procedure described in Section 4.1.3.1, until a target BSII value is reached (*e.g.*, 0.375). The final SNR is the predicted SRT which corresponds to 50% intelligibility. The validation process includes a fitting stage in which modeling parameters are manipulated under a limited range of baseline conditions in order to produce predictions that optimally fit the measured data.

4.4.2.2 Experimental conditions

As mentioned in Section 3.1.1.1, the HINT corpus contains 12 lists of 20 sentences of speech material that may be used for testing (practice lists are not calibrated for use in

the testing procedure). Since sentences cannot be repeated during the test, any subjective experiment can have a maximum of 12 conditions. Therefore, we have chosen a number of representative testing conditions which were arranged in four different experiments, each performed on a separate population of subjects, to investigate the following effects:

1. Effect of spatial sound-source distribution
2. Effect of two clipping algorithms
3. Effect of two compression algorithms
4. Effect of two noise-reduction algorithms

For experiments involving noise reduction or compression, the adaptive nature of the algorithms suggests performing measurements with a wide range of continuous and fluctuating masking signals. However, due to limitations both in time and the number of conditions possible within a test session, two types of maskers are considered in these experiments: the speech-shaped noise (SSNOISE) and an intermittent noise (SSINT16). These two signals provide the best and worst-case masking scenarios allowing us to test the limits of the model's applications (*c.f.* Section 4.1.4.1).

Ideally, NH and HI subjects would be tested in these experiments. HI subjects are particularly interesting as they are candidates for hearing aid fittings. However, the difficulty of recruiting HI subjects to collect a sufficient amount of subjective data to validate the model makes this a lengthy endeavor. For time restrictions, only NH participants are considered for this thesis. It is worth mentioning that the distortions tested in this work are commonly used in electronic hearing protectors and open-ear headsets which are often worn by NH subjects in the workplace [78, 79]. Testing HI subjects and modeling supra-threshold deficits of cochlear hearing-loss is the subject of a future extension of this work. Several studies have investigated ways to apply correction for different aspects of hearing impairment to SII [76, 77] and ESII [5, 170, 173, 184] predictions. This will be further addressed in Section 8.2.2.

We will report on the subjective portion of these experiments as a separate study in Chapter 6. Comparison with model predictions under the experimental conditions considered is deferred to Chapter 7.

4.5 Conclusion

In this chapter we have presented a three-stage model to predict speech intelligibility under binaural listening situations in the presence of nonlinear signal processing conditions. The conceptual and implementation details of the model have been thoroughly discussed along with the details of two software systems: the hearing device simulator, and the Matlab speech testing environment. The development of these tools has played a crucial role in permitting this doctoral work to come to fruition. The model has been validated in multiple steps carried out at different stages during the development. The validation strategy has been presented in detail in the previous section.

The topics discussed in this chapter highlight important research contributions, which will be expanded on throughout the remainder of this thesis: (1) The novel speech and noise distortion estimation method which make the first stage of the model (*c.f.* Section 4.1.1) will be evaluated in Chapter 5; (2) the new three-stage binaural speech intelligibility model presented throughout Section 4.1 will be validated in its entirety in Chapter 7; (3) the Matlab speech testing environment developed during this work has proven to be a powerful tool with tremendous benefits to our research group; and (4) the experiments carried out for this work to collect the subjective SRT data reported in Chapter 6 are, to the best of our knowledge, the first of their kind using the HINT test material. These measurements would not have been possible without the new testing environment, whose future outlook shows promising applications.

Chapter 5

Monaural Evaluation of the Phase-Inversion Method

Introduction

In the previous chapter we proposed a three-stage model to predict speech intelligibility in a binaural listening condition when using nonlinear hearing devices. In this chapter we present a validation of the phase-inversion method used in the first stage of the model. This is performed in a monaural context independent of subsequent binaural processing in the model. The chapter is largely based on a recent simulation-based study published in [58] as a revision of [57]. The main contribution of this chapter is to serve as a proof of concept validating the phase-inversion signal-separation method by:

- modeling distortion components present in the speech and noise estimates recovered with the phase-inversion method,
- demonstrating the benefit of using a modeling approach that distinguishes between speech and noise distortion, and
- incorporating the distortion measure in intelligibility prediction.

The study presented in this chapter was completed at an earlier stage in this doctoral work and is limited to a few processing systems, which include two compressive nonlinearities (AGCi and AGCo systems), as well as peak clipping, which may be interpreted as instantaneous compression with an infinite ratio. Center clipping and noise-reduction algorithms were not yet included in the hearing aid simulator. Moreover, the work was carried out before completing the short-time extension of the model (*c.f.* Section 4.1.4). Thus, the speech-modulated noise signal (SSMOD01) was not available at the time of the study. The spectral estimates shown in this chapter are based on the first sentence of the HINT test material and the speech-shaped noise signal (SSNOISE). As for SII predictions, these are averaged over the 10 sentences that constitute the first list of the original grouping of the HINT sentences [153].

The remainder of this chapter will be organized as follows: In Section 5.1, an overview of the proposed evaluation and the simulation setup is presented. This is followed by the results obtained when simulating normal-hearing and hearing-impaired users using common hearing aid processing steps in Section 5.2. These results include spectral estimates and SII predictions obtained using two approaches: the phase-inversion method and the coherence function. In Section 5.3, we discuss the validity of the phase-inversion signal-separation model by comparing the SII predictions obtained with the two methods and correlating them with relevant findings established in the literature. Concluding remarks are presented in Section 5.4.

5.1 Evaluation overview

5.1.1 Rationale

Stage 1 of the binaural model presented in the previous chapter consists of (1) signal separation based on the phase-inversion method and (2) distortion estimation based on a Hilbert transform measure (*c.f.* Section 4.1.1). In this chapter the speech and noise

distortion estimators in equations (4.13) and (4.14) are used to evaluate the signal-separation power of the phase-inversion model under different processing conditions. For a linear system, the recovered output signals $u'(t)$ and $\tilde{u}'(t)$ (corresponding to the input signals $u(t)$ and $\tilde{u}(t)$) will form a Hilbert pair and the term $S_{u'u'}(f)$ will be the same in equations (4.9) and (4.11). If the system is not linear and $S_{u'u'}(f)$ still cancels out through equation (4.13), then the model assumptions hold for this system: *i.e.* the nonlinear distortions are fully captured by the error signals $e_1(t)$ to $e_4(t)$. If $S_{u'u'}(f)$ does not cancel out completely, then comparing $S_{spd}(f)$ and $S_{c_1c_1}(f)$ from (4.9) can provide an indication of how distorted the speech estimate recovered with the phase-inversion method is. A similar reasoning also applies to $S_{nzd}(f)$ and $S_{d_1d_1}(f)$ from (4.10). As such, it is useful to define a speech-to-distortion ratio (SpDR) and a noise-to-distortion ratio (NzDR) based on these estimates:

$$SpDR(f) = \frac{S_{c_1c_1}(f)}{S_{spd}(f)} \tag{5.1}$$

$$NzDR(f) = \frac{S_{d_1d_1}(f)}{S_{nzd}(f)} \tag{5.2}$$

These ratios can be used in the context of speech intelligibility predictions based on the SII [4], for example, in assessing the impact of distortions in the speech and noise signal estimates on the computation of the audibility function in the different frequency bands. The ratios can also be frequency-weighted (using the SII band importance function) to produce a single value.

In addition, the coherence function introduced in Section 3.2.1.1 is used to complement the evaluation. The MSC estimated using equation (3.10) is used to obtain spectral estimates of the output speech and noise signals as in equations (3.8) and (3.9). Thus, the estimates $\widehat{S}_{u'u'}(f)$ and $\widehat{S}_{v'v'}(f)$ are compared to those obtained with the phase-inversion method. We recall, however, that the estimation of the MSC is subject to sources of variance and bias, the effect of which will also be considered in the evaluation.

Table 5.1: *Hearing thresholds at standard audiometric frequencies and uncomfortable levels for the simulated subjects (in dB HL).*

Subject	**250**	**500**	**1k**	**2k**	**3k**	**4k**	**6k**	**8k**	**UCL**
NH	0	0	0	0	0	0	0	0	98
HI01	20	25	30	40	35	45	40	55	102
HI02	30	40	50	50	55	65	60	65	107
HI03	80	75	60	55	65	80	75	70	115

5.1.2 Simulation setup

The evaluation framework used in this study consists of computer simulations using an early version of the MATLAB hearing device simulator presented in Section 4.2.2. One NH and three HI subjects are simulated (*c.f.* Table 5.1). The audiograms for subjects with mild-moderate (HI01), moderately-severe (HI02) and severe (HI03) hearing loss were taken from a local database. For each subject, a UCL representative of an average subject with a similar degree of hearing loss was used [159]. Six hearing aid systems were simulated as outlined in Table 5.2. The AGC system is a single-channel inter-syllabic compression amplifier with a threshold of 50 dB SPL, a 3:1 ratio, and attack and release times of 5 ms and 50 ms respectively. Peak clipping is performed at the subject's UCL. Compression limiting is implemented with a threshold set at UCL – 5 dB, a 10:1 ratio, a 2 ms attack and 20 ms release time. The test sentences used in these simulations are taken from the original HINT test [153], composed of 25 lists of 10 sentences, digitized and re-sampled at 20 kHz (*c.f.* Section 3.1.1.1).

The power spectral densities presented in this work are computed according to Welch's averaged periodogram method [209], using a 512-point Hamming window with 50% overlap. The effect of window length on the coherence-based spectral estimates is discussed in Sections 5.2.1.1 and 5.2.3.4.

Finally, intelligibility prediction is performed based on the speech and noise spectra

Table 5.2: *Hearing aid systems used during the evaluation.*

Hearing Aid	**Prescription**	**Output Limiting**
HA01	NAL-RP	none
HA02	NAL-RP + AGC	none
HA03	NAL-RP	compressive
HA04	NAL-RP + AGC	compressive
HA05	NAL-RP	Peak clipping
HA06	NAL-RP + AGC	Peak clipping

obtained with the coherence and phase-inversion methods using the standard SII (ANSI S3.5–1997 (R2012) [4]). The third-octave band procedure is used with the band audibility function for short passages (Table B.2 in [4]). For the phase-inversion method, an adjustment is made to the SII computation by including the speech distortion spectrum computed using equation (4.13) as an additional noise source in the equivalent noise spectrum level. The effect of this adjustment is investigated in Section 5.2.3.5.

5.2 Results

Simulations were performed for all subjects and hearing aids, at soft (50 dB SPL), normal (65 dB SPL) and loud (80 dB SPL) speech presentation levels. A wide range of input SNRs is considered using a fixed speech level while varying the noise level. We examine the effect of each hearing aid system from Table 5.2 on the recovered speech and noise spectra, with insight from the new distortion measures derived in equations 4.13 to 5.2, as well as on SII predictions. Wherever applicable, asterisk (*) markers denote the coherence method, while square markers denote the phase-inversion method.

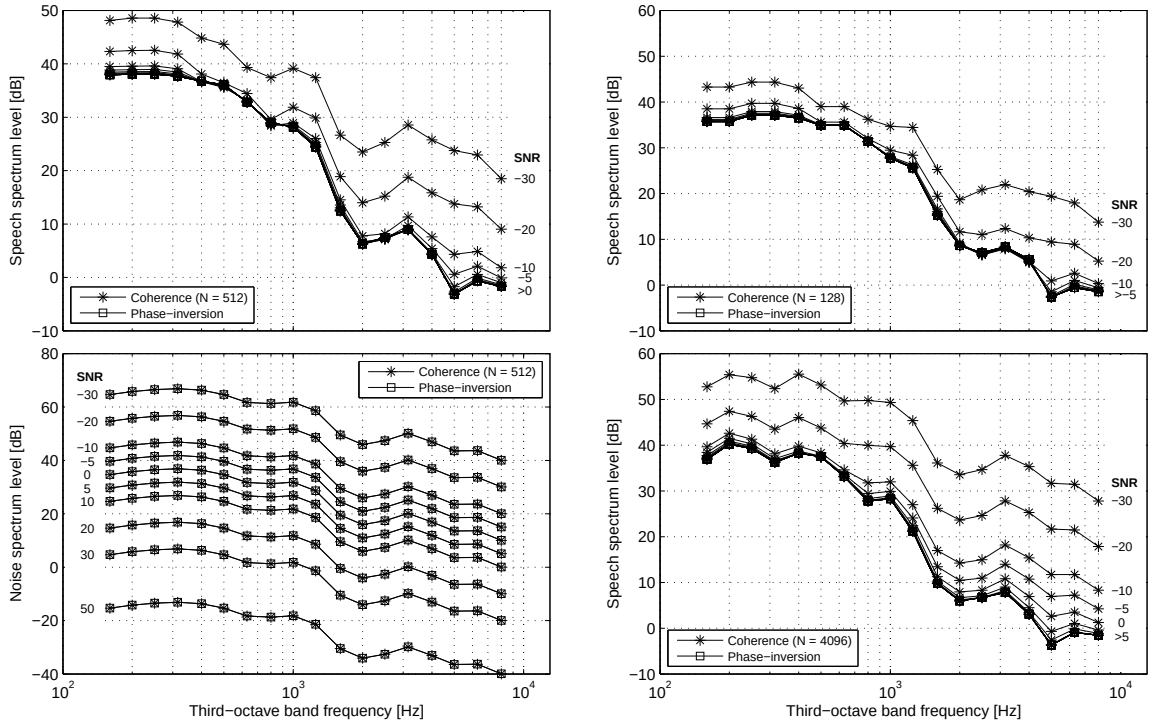


Figure 5.1: *Recovered spectra obtained under additive noise conditions for a normal hearing subject with no hearing aid processing. In the left panels, both speech and noise spectra are shown using the phase-inversion and coherence method using a window size of 512 points. In the right panels, only the speech spectra are shown, with the coherence function computed at windows sizes of 128 and 4096 points. Speech level: 65 dB SPL.*

5.2.1 Recovered speech and noise spectra

5.2.1.1 No hearing aid processing

Normal-hearing subjects do not require a hearing aid. In the absence of processing, simulations amount to testing the effect of additive noise. In this case, the phase-inversion method recovers the exact signals at the output (equations (4.5) and (4.6) will have no error terms). The left panels of Figure 5.1 show the speech and noise spectra recovered with the coherence and phase-inversion methods at different SNR levels for the NH subject. The noise spectra from both methods overlap for all SNR levels (bottom panel). For the speech spectra (top panel), the coherence method (computed using 512-point windows) deviates from the phase-inversion method at low SNR levels, where the MSC

is near zero. This can be attributed to averaging bias affecting the coherence estimation. With these simulations, we can determine the window size that limits the bias effect, and the SNR at which to perform subsequent simulations, so that the bias can be ignored when analyzing the recovered estimates.

Similar simulations were performed with window sizes ranging from 128 to 4096 points to study the effect of averaging bias. The speech estimates at the limits of this range are shown in the right panels of Figure 5.1. Using a shorter window permits averaging more segments, which limits the bias effect at the expense of frequency resolution. Conversely, using a longer window gives a better resolution, but increases the bias. In our simulations, a 512-point window provided an ideal trade-off between frequency resolution and bias. The noise estimates were not affected by this bias effect.

5.2.1.2 Linear and compressive processing

Spectral estimates recovered with the coherence and phase-inversion methods at the output of the linear and compressive hearing aids (HA01, HA02 and HA03) are shown in Figure 5.2. Also shown on these plots are the spectra of the speech and noise distortion along with the SII-weighted SpDR and NzDR. In the left panels, subject HI02 is simulated using a loud speech presentation level to ensure HA02 is not operating in linear mode below the compressor's knee point. The spectra obtained with HA01 serve as a baseline to isolate the effect of the AGC system. As expected, the output of HA02 is approximately 7.5 dB below the output of HA01 due to the compressor's attenuation gain. For each device, the two methods produce overlapping speech and noise spectra. The low level of distortion, reflected in the high SpDR and NzDR, confirms the suitability of the phase-inversion signal-separation model under these processing conditions. In the right panels, the same simulation is repeated with HA03. The effect of compressive output limiting is negligible as evidenced by the overlapping signal and distortion spectra for both HA01 and HA03.

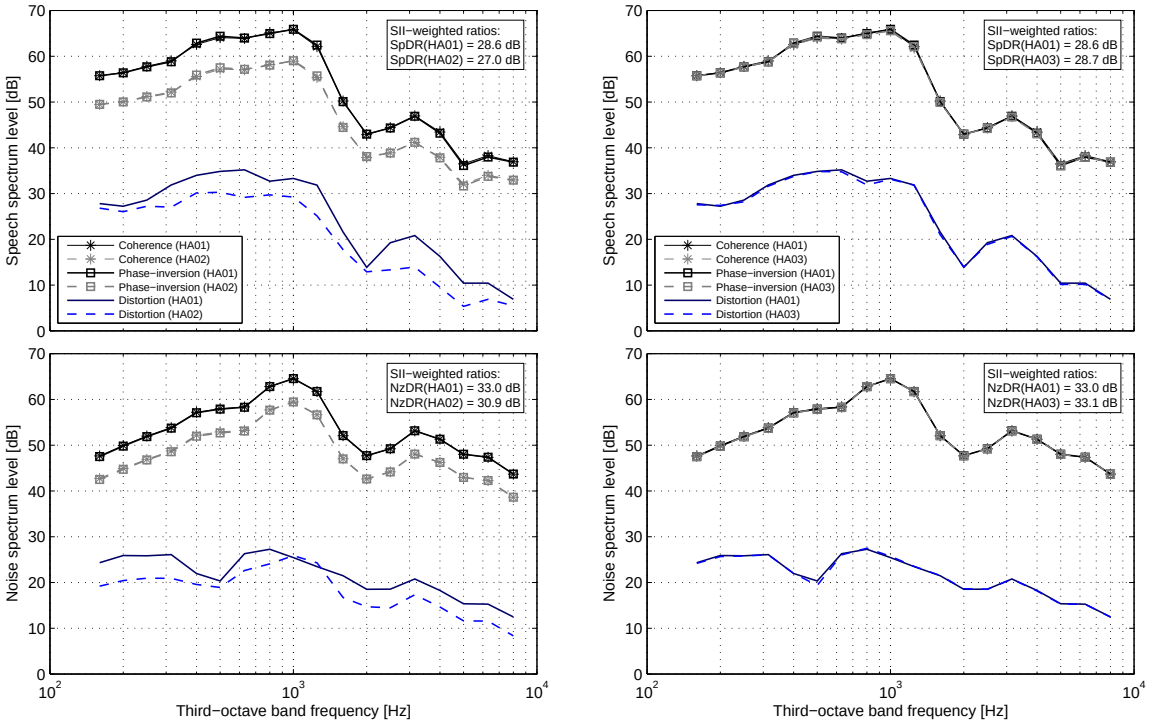


Figure 5.2: *Recovered spectra obtained under linear and compressive processing for subject HI02. Speech and noise spectra estimated with the coherence and phase-inversion methods are shown for pairs HA01-HA02 (left panels) and HA01-HA03 (right panels). Also shown are the speech and noise distortion spectra for the phase-inversion method. For a given device, the spectral estimates recovered with both methods overlap. For HA03, the plots also overlap with those of HA01. Speech level: 80 dB SPL, SNR: 5 dB.*

5.2.1.3 Peak clipping

The effects of peak clipping (HA05) is shown in the left panel of Figure 5.3. The plots show the speech, noise and distortion spectra at the output of the pair of hearing device HA01-HA05 for the same subject and presentation level as in Figure 5.2. With HA01 serving as baseline, the plots illustrate the importance of the proposed speech and noise distortion measures: although the speech and noise spectra from HA05 seem to overlap those from HA01 for both methods, the distortion spectra indicate that the recovered estimates are distorted. In fact, the distortion spectra are consistent with the fact that peak clipping tends to produce high-frequency harmonics of the original signal. For example, in this simulation, the (unweighted) SpDR reaches levels as low as 7 dB over

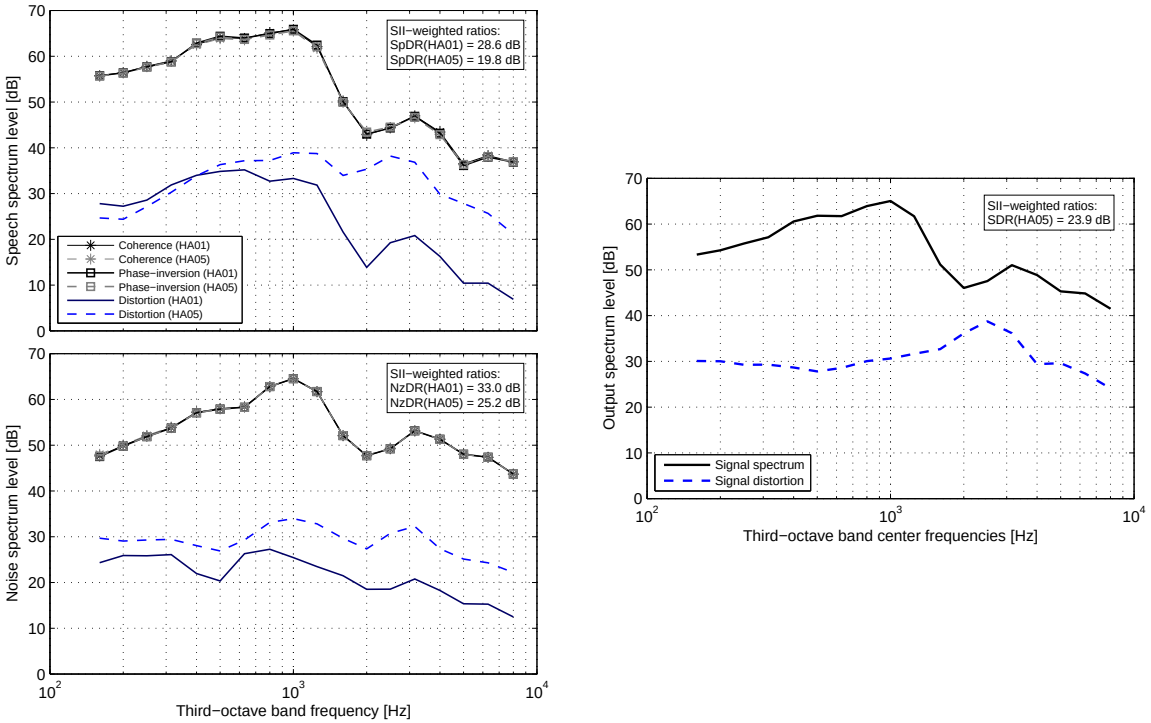


Figure 5.3: *Recovered spectra obtained under linear and peak clipping processing for subject HI02. Speech and noise spectra estimated with the coherence and phase-inversion methods are shown for the pair HA01-HA05 (left panels). The spectral estimates recovered with both methods overlap for both devices, but the distortion spectra provide additional information. For comparison, the total signal and signal distortion computed according to [154] are also plotted (right panel). Speech level: 80 dB SPL, SNR: 5 dB.*

the frequency range 2-5 kHz which contributes up to 33% of speech intelligibility for the chosen band importance function.

The right panel of Figure 5.3 offers additional support for our methodology. The plot shows the spectrum of the total signal at the output of HA05 along with a distortion spectrum computed according to Olofsson and Hansen's original approach [154]. The SII-weighted SDR falls within the range of our SpDR and NzDR measures shown in the left panel. While the SDR provides an indication that the output signal is distorted, our approach using separate estimators for speech and noise distortion confirms that the distortion affects the recovered speech signal more. This information will be useful when the recovered estimates are used for intelligibility prediction (*c.f.* Section 5.2.3).

5.2.2 Speech-to-distortion and noise-to-distortion ratios

The last section illustrated the utility of the new speech and noise distortion measures in a particular listening condition. In this section, we use the SII-weighted SpDR and NzDR to analyze the effect of distortions across a range of conditions. Figure 5.4 shows the SII-weighted SpDR and NzDR as functions of input SNR for user HI02 with all six hearing aids from Table 5.2. Simulations are performed at soft, normal, and loud speech presentation levels and shown in the left, middle and right panels respectively. The most significant effects are shown at loud presentation levels where the nonlinear elements in the devices would be most active.

In the absence of nonlinear processing (HA01), the SpDR and NzDR remain constant, and provide a baseline to study the effects of distortion in other hearing aids. AGC processing introduces a fairly low amount of distortion at all presentation levels: the SpDR remains high, within 5 dB below the linear baseline; the NzDR is elevated as well and decreases only slightly at high SNR levels where the noise is very low. Third, the compressive output limiter does not seem to introduce any additional distortion at all presentation levels, as evidenced by the overlap of the plots for HA03 and HA01 (and also the overlap of the plots for HA04 and HA02). Finally, peak clipping is known to introduce the highest amount of distortion. The peak clipper is most active at loud presentation levels (right panels), and particularly at low SNR, where the noise level is high. The distortion decreases (ratios increase) as the SNR is increased. As the overall signal level becomes dominated by the fixed speech signal (input SNR above 0 dB), the SpDR and NzDR reach a stable level. When peak clipping is combined with AGC processing (HA06) the resulting SpDR and NzDR show significant improvements (less distortion). This is consistent with the expectation that the reduced gain of the AGC system would lead to less aggressive clipping. The effect of peak clipping can also be observed at normal presentation level (middle panels) at very low SNR.

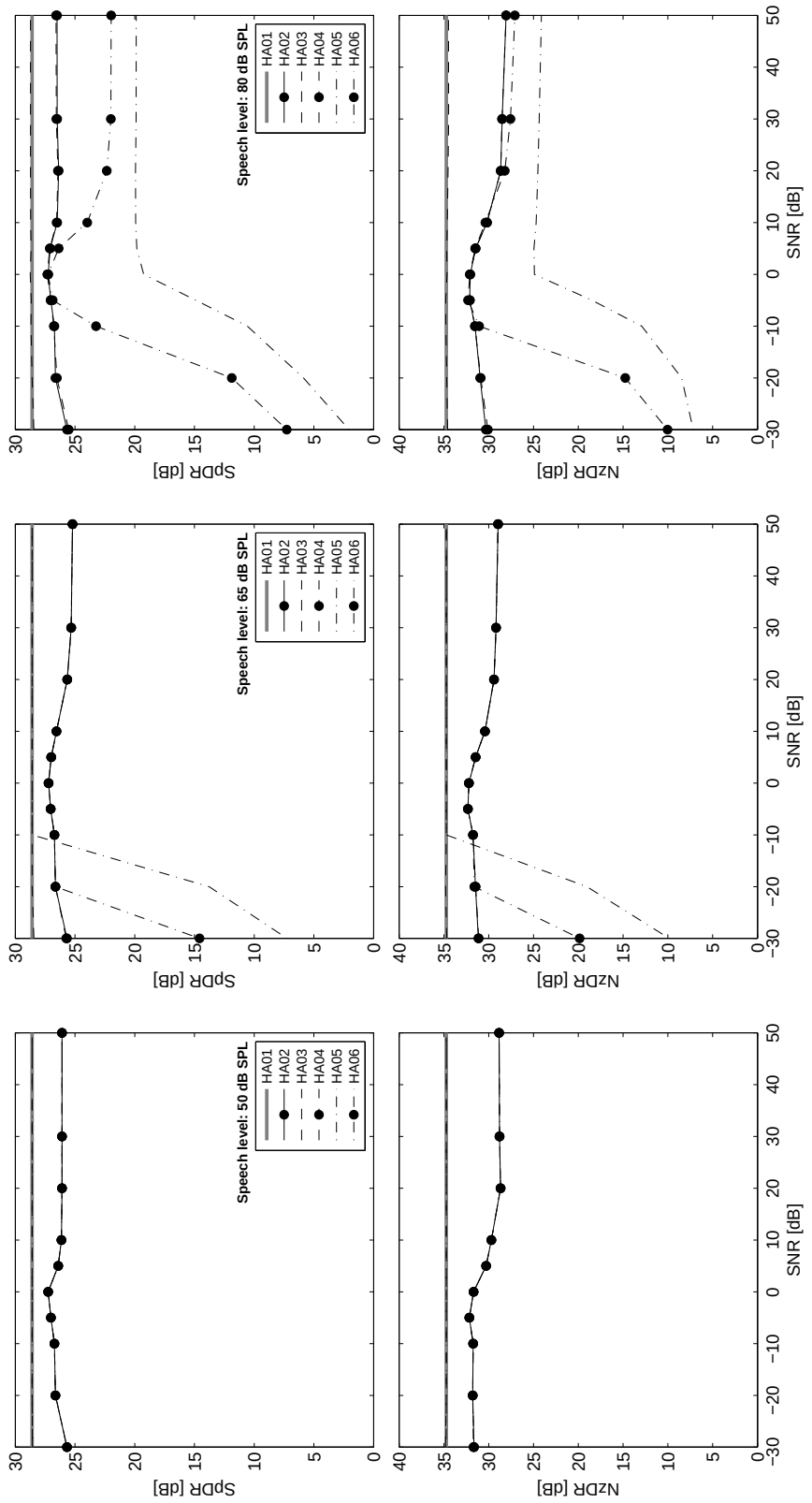


Figure 5.4: *SII-weighted speech-to-distortion and noise-to-distortion ratios computed from the phase-inversion method for subject HI02 when using the hearing aids listed in Table 5.2. Under soft speech level (left panels) plots for HA01, HA03 and HA05 overlap, and so do the plots for HA02, HA04 and HA06. Under normal (middle panel) and loud (right panel) speech levels, the plots for HA01 and HA03 overlap and so do the plots for HA02 and HA04. Solid markers denote AGC hearing aids, while systems based on a linear prescription have no markers.*

5.2.3 SII predictions

SII predictions for all users and hearing aid conditions at normal speech level are summarized in Figure 5.5. In Figure 5.6 we isolate the effects of the AGC system (HA02), compression limiter (HA03) and peak clipper (HA05) for subject HI02 at different speech levels. For each condition, we show the SII for the phase-inversion and coherence method as a function of input SNR. By analyzing these plots in light of findings from the literature, we can make qualitative observations about speech intelligibility predictions computed with the phase-inversion method, while accounting for the proposed distortion measure. For example, linear processing does not effect the MSC or the Hilbert-transform distortion measure. For such systems, we expect the two signal-separation methods to give similar predictions, except perhaps where bias effects in the MSC are significant. For nonlinear systems, the more distortion is introduced, the more we can expect the predictions to deviate since the underlying distortion in each method is affected differently by various nonlinear mechanisms.

5.2.3.1 No hearing aid processing

SII predictions for NH subjects are superimposed on each panel of Figure 5.5 (thick lines). This plot shows that the differences between the two methods observed in Section 5.2.1.1 have no effect on speech intelligibility. The deviations correspond to conditions where the speech level is too low compared to the noise level, and does not contribute to intelligibility. In fact, the SII standard explicitly ensures that any frequency band with a SNR below -15 dB has an audibility factor of zero [4].

5.2.3.2 Effects of hearing loss

Figure 5.5 shows how, at normal speech level, the predicted SII decreases consistently for all hearing aids as the severity of the hearing loss increases. Moreover, the AGC seems to be the only system to produce distortion at this presentation level: The compressive

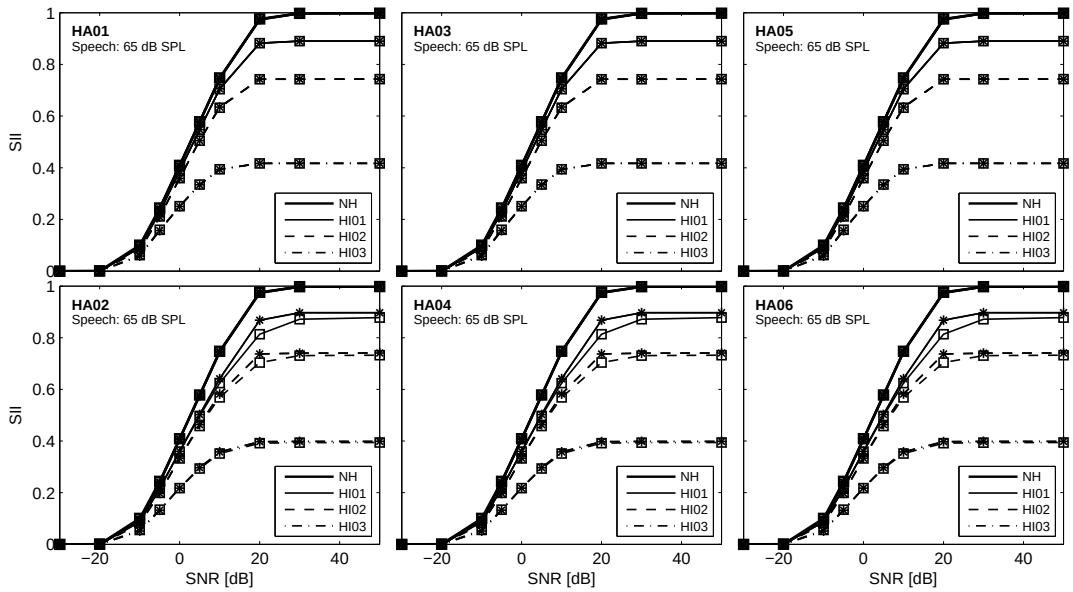


Figure 5.5: *SII predictions as a function of SNR for all combinations of HI subjects and hearing aids from tables 5.2 and 5.1 at normal speech level (65 dB SPL) computed using the coherence and phase-inversion methods. The SII for the NH subject without hearing aid processing is also plotted as a baseline.*

limiter (HA03) and peak clipper (HA05) are not quite active at this point (except perhaps at very low SNR, where the SII is near zero and the distortion has no observable effect). The effect of each of these systems is considered in detail in the next section. The plots for HA04 and HA06 confirm these observations when the nonlinearities are combined.

5.2.3.3 Isolating effects of nonlinearities

The top row of Figure 5.6 shows the SII predictions at low speech level for each nonlinear system. For HA02, the AGC system is operating around the compressor's kneepoint, applying an overall positive gain, which results in improved predictions compared to HA01. The overlapping predictions obtained with the two methods show that, although the AGC system is operating in a nonlinear mode during portions of the speech above 50 dB SPL, the distortion is too small, and the overall behaviour is similar to a linear hearing device. For HA03 and HA05, the output limiters appear to be inactive. These observations were confirmed by producing SII-weighted SpDR and NzDR plots similar

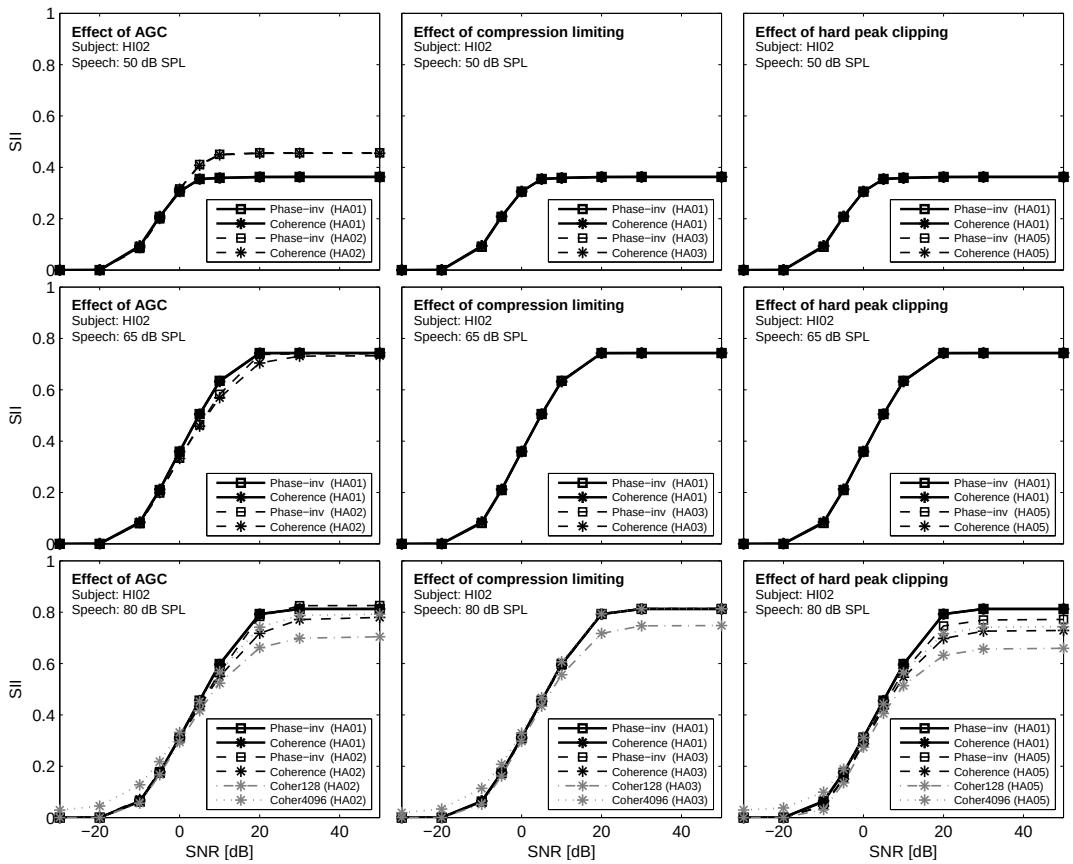


Figure 5.6: *SII predictions showing the effects of nonlinear processing. Subject HI02 is simulated using soft (50 dB SPL), normal (65 dB SPL) and loud (80 dB SPL) speech presentation levels. The panels in the bottom row also show the predictions obtained with the coherence method using different window sizes.*

to those presented in Figure 5.4 at low speech level.

The middle row of Figure 5.6 shows the SII predictions at normal speech level for each nonlinear system. Once more, the AGC system seems to be the only component that affects intelligibility. Here, the compressor is operating above the kneepoint around the zero-gain level (set to 65 dB SPL). Therefore, the target gain of the hearing aid should be, on average, the same as the linear prescription (HA01). The plot for HA02 shows a slight decrease in intelligibility compared to HA01 for an input SNR between -5 and 20 dB. The plot also shows that the AGC has a more pronounced effect on the predictions obtained with the coherence method.

The bottom row of Figure 5.6 shows the SII predictions at loud speech level, at which the nonlinear processing components (AGC, compressive output limiter, and peak clipper) are most active. Compared to linear processing (HA01), predictions for the phase-inversion method show decreased intelligibility for HA02 over the SNR range between 5 and 20 dB, and slight improvements above a 30 dB SNR. No distortion effects are observed for HA03. For HA05, a reduction in SII is observed, especially at high SNR levels. The coherence method is more sensitive to these distortion effects.

5.2.3.4 Coherence and window length considerations

Measurements using window lengths ranging from 128 to 4096 were performed for hearing aids HA02, HA03 and HA05. The SII predictions at the extremes of the range tested are shown in gray lines at the bottom row of Figure 5.6. The results confirm that the 512-point window size presents an ideal tradeoff to avoid bias effects. More specifically, the plots show that: (1) for low SNR levels, the averaging bias could affect speech intelligibility predictions when a very long window is used; (2) for high SNR levels, the group-delay bias leads to deviations in SII predictions when a very short window is used. Finally, even when using a very long window to avoid group-delay bias effects at high SNR levels, predictions with the coherence method are still more sensitive to the distortions introduced by HA02 and HA05 than the phase-inversion method.

5.2.3.5 SII adjustment for phase-inversion method

The SII predictions presented so far included an adjustment which accounts for the effect of speech distortion. The effect of this adjustment is investigated in Figure 5.7 for the conditions presented in Figure 5.6. The square markers are the same values reported in the previous figure. The open circles denote the SII values computed without accounting for speech distortion. These plots reveal that predictions with and without the adjustment overlap for linear processing (HA01) and compression (HA02 and HA03). The plots of SII-weighted speech-to-distortion ratios in Figure 5.4 already predicted this

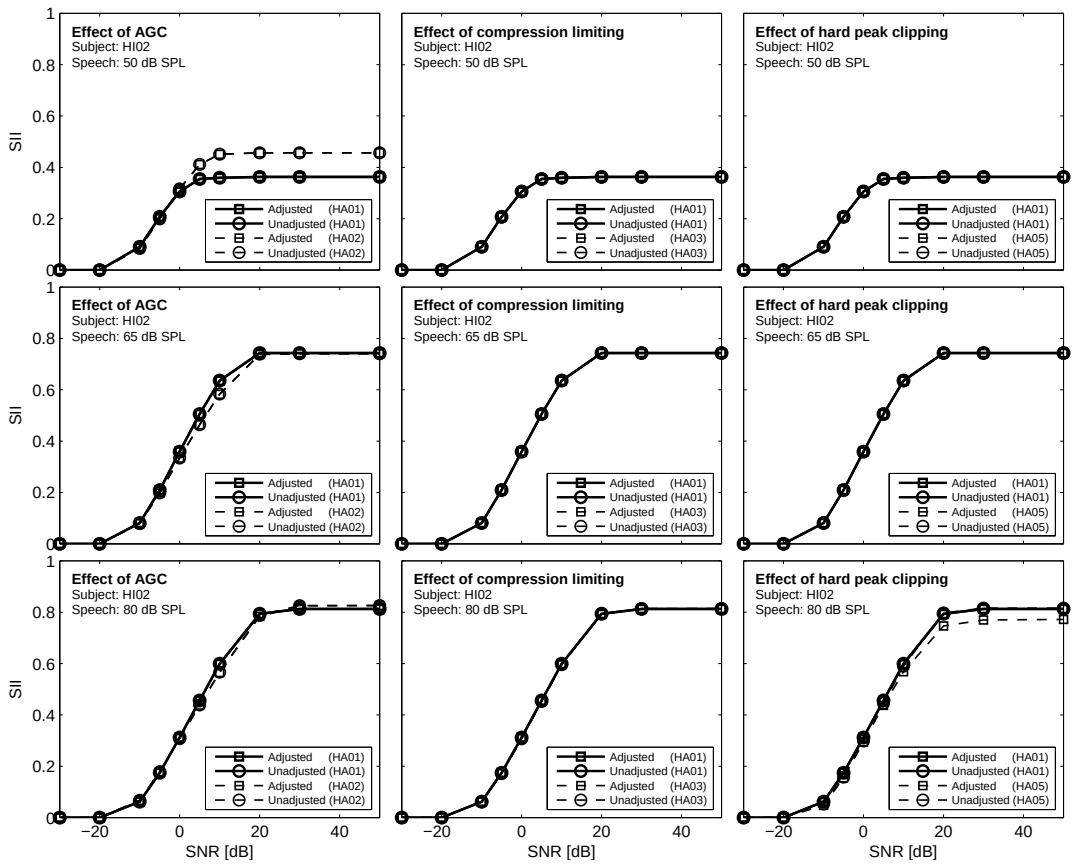


Figure 5.7: *Effect of SII adjustment for the same conditions shown in Figure 5.6. Square markers denote the adjusted SII and circles denote the unadjusted SII.*

result: the distortion produced is very low, except perhaps at very low SNR values where the speech level is too low to contribute to speech intelligibility (SII is near 0). The effect of the adjustment is most clear for peak-clipping (HA05), where the adjusted SII values are as much as 0.05 lower than the unadjusted values at high SNRs for loud speech. In fact, the unadjusted predictions for HA05 were as high as for HA01, a result which could already be predicted by the overlap of the recovered spectra from both hearing aids (*cf.* Figure 5.3).

5.3 Discussion

This work offers an evaluation of the phase-inversion signal-separation method which, to date, has only been verified under very few conditions. The proposed speech and noise distortion measures provide an evaluation framework to assess the suitability of using the phase-inversion method in the context of speech intelligibility predictions when using hearing aids. The results in this work can be qualitatively compared to some relevant studies in the literature.

In a study aimed at "understanding compression" in which 18 studies over the past 30 years are reviewed [121], Kates remarks that results on the benefits of dynamic range compression are mixed, with studies finding advantages, no effect, or even disadvantages with respect to speech intelligibility and quality. A problem for comparing these conclusions is the range of compression parameters (compression ratio, threshold, attack and release times, number of channels) and experimental conditions (choice of stimuli, hearing loss severity) considered. Two studies which use comparable settings to the AGC in HA02 stand out in the literature. In [41], Dreschler et al. measured SRTs using a 60 dBA noise presentation level to evaluate the benefit of single-channel compression compared to linear amplification for a group of subjects. They found a slight difference in favor of the linear hearing aid: the mean SNR to obtain 50% intelligibility was +2 dB (± 2.8 dB) for the linear hearing aid, and +2.9 dB (± 4.7 dB) for the compression hearing aid. For the HINT material in this study, the SII corresponding to 50% sentence intelligibility can be estimated from normative data [204] as 0.375. In Figure 5.6, this SII value corresponds to a 0 dB SNR for HA01 and +1 dB SNR for HA02 (left-most panel of the center row). This is in reasonable agreement with the observations from [41]. Furthermore, Jenstad and Souza in [112] studied the benefits of single-channel compression compared to linear amplification on phoneme recognition in quiet. They used the same presentation levels as those chosen for the current study. Their results can be compared to the SII predictions at 50 dB SNR as shown in the leftmost panels in Figure 5.6, especially since the

plots reach a plateau as the SNR level increases. Our results fall well within the range of benefits reported for individual phonemes in Figure 7 of [112].

The effect of output limiting on speech intelligibility is well documented [35]: compression output limiting has been found to produce no significant shift in SRT levels, while peak clipping is known to negatively affect intelligibility. However, when quantified, it is suggested that the effect of peak clipping may not be large [35, 67, 102]. Furthermore, in the preamble to the three-level coherence-based SII [124], Kates and Arehart report that a single-level approach (like the one reported in this work) underestimates speech intelligibility for peak clipping conditions when compared with subjective scores. The (adjusted) SII predictions shown in this work for hearing aids HA03 and HA05 are in agreement with these observations.

In this work, we have used the `hilbert` function in MATLAB to perform Hilbert transformation when deriving the speech and noise distortion measures used to evaluate the phase-inversion method. The function is a frequency-domain operator which acts on the entire signal length. Thus simulations using different segment sizes for spectral estimation showed no bias or variance effects on our results. However, the Hilbert transform could be implemented using short time-domain approximating filters, or overlap-add techniques. When such implementations are needed, variance and bias effects due to the number and size of segments should be investigated similar to what was done in Sections 5.2.1.1 and 5.2.3.4 for the coherence-based estimation.

The present study focused on fast compressive nonlinearities and peak-clipping distortion commonly found in hearing aids. Additional simulations with slow-acting compression were found to be similar to a linear device. An extension of this work to study the effects of different types of NRAs is being considered for the future. NRAs typically act like a center-clipping mechanism, reducing low-level portions of the input signals, which may be associated with consonants [124]. They tend to improve speech quality and degrade intelligibility. Vocoder effects and other nonlinear mechanisms would also merit further investigation.

It is important to include additional remarks on intelligibility prediction to help clarify future directions in relation to other work. The standard SII is based on long-term spectra and ignores the temporal fluctuations of the signal envelope. An extension to the SII (the ESII), which uses a segmental approach, has been proposed to predict speech intelligibility in fluctuating noise [170, 171]. Other measures, such as those based on the STI (IEC60268-16 [103]), also rely on the modulation of the signal envelope. Taal et al. evaluated several objective measures when using noisy speech processed with ideal time-frequency segregation (ITFS) [197]. They found that methods based on temporal correlation performed rather poorly, while models based on correlation in the joint spectro-temporal domain gave good predictions. They also proposed a simple measure based on spectral correlation which gave the best predictions for ITFS-processed speech. In other work, Jørgensen and Dau proposed a measure based solely on envelope modulation spectra [114]. The measure, which ignores the signal's fine structure, gave good intelligibility predictions under conditions of reverberation and spectral subtraction. These findings suggest that an extension of this work could benefit from the integration of spectro-temporal processing, possibly in a manner similar to the ESII, to capture the time-varying effects of fast-attack AGC systems or NRAs.

Finally, there is an interest in extending this work to develop an objective measure to predict binaural speech intelligibility while wearing hearing aids. The challenge in this task is to isolate the effect of distortions introduced by the hearing devices from the effect of binaural auditory processing. One approach [55] is to use a time-domain signal-separation technique like the phase-inversion method as a preliminary stage to a binaural SII metric [12]. The speech and noise distortion measures proposed in this work provide a foundation for the evaluation framework used to validate this approach.

5.4 Conclusion

This work presents an evaluation of the phase-inversion signal separation method which has been extensively used in hearing aid research since its introduction in [87]. Hearing-aid simulations are performed using common settings for AGC processing, compression limiting and peak clipping. The spectra of the recovered signals are visualized and compared to those obtained using the coherence function. Spectral estimates are used along with new distortion measures to analyze separately the effects of nonlinear processing on speech and noise estimations, especially in the context of speech intelligibility prediction. The proposed distortion measures extend the work of Olofsson and Hansen [154] and are found to be helpful evaluation tools and the results validate the use of the phase-inversion method for signal separation under the processing conditions investigated. The present study shows the phase-inversion method as an attractive solution for signal separation which, combined with the Hilbert-based distortion estimators, can be integrated into a spectro-temporal or binaural approach for objective speech intelligibility prediction.

Chapter 6

Subjective Measurement of Speech Reception Thresholds

Introduction

In order to validate the proposed model, subjective SRTs measured under different binaural listening conditions and nonlinear processing commonly encountered in hearing devices are required. To the best of our knowledge, such perceptual data is not readily available for modeling purposes. On the one hand, most studies using specific signal processing algorithms are performed under monaural listening or using a spatial configuration that does not provide binaural benefit. On the other hand, in studies investigating the effects of hearing devices in different binaural spatial configurations, the details of the underlying algorithms are not accessible. Therefore, it was imperative to perform in-house measurements on a population of subjects using specific algorithms under a variety of binaural listening situations.

This chapter reports on the experiments performed to collect this data, which will be used in the following chapter to compare with model predictions under the same conditions. The experiments presented in this chapter provide the first formal test of the Matlab speech testing environment (*c.f.* Section 4.3). They offer new original data on

the interaction of the following effects on SRT measurements using the American English version of the HINT:

- spatial sound distribution (not limited to 0° and 90°),
- continuous[1] vs intermittent masking, and
- various nonlinear processing algorithms.

Multiple experiments were required in order to test the different listening and processing conditions. These will be described, along with the rest of the experimental procedure, in Section 6.1. The results of each experiment are presented in Section 6.2, followed by a discussion in Section 6.3 and concluding remarks in Section 6.4. The measurements presented in this chapter provide an original data set using specific signal processing algorithms accessible to the research community.

6.1 Methods

6.1.1 Experimental setup

6.1.1.1 Hardware and software equipment

All experiments were performed with subjects seated comfortably inside a double-walled audiometric room (Industrial Acoustics Company, Inc., NY, USA). Tests were carried out using the Matlab speech testing environment. This platform makes use of the HRTFs discussed in Section 4.2.3 to simulate spatial sound-source configurations for binaural hearing, and the Matlab hearing device simulator toolbox (HDSTB v2.0) to simulate hearing devices. A list of all devices used in these experiments is shown in Table 6.1. A detailed description of the hearing device simulator and the operations implemented in each device is provided in Section 4.2.2.

[1] Please refer to the footnote on page 72.

Table 6.1: *List of simulated hearing devices. Also listed are settings for the following device parameters, wherever applicable: clipping or compression threshold (THR), zero-gain reference level (REF), compression ratio (CR), attack time (Ta) and release time (Tr).*

		Parameters				
Abbr.	Device description (dB SPL)	THR (dB SPL)	REF	CR (ms)	Ta (ms)	Tr
BASE	No device					
PC50	Peak clipping at 50 dB	50				
PC60	Peak clipping at 60 dB	60				
CC50	Center clipping at 50 dB	50				
CC60	Center clipping at 60 dB	60				
AGCi	Intersyllabic compression	40	68	2.5	4	40
AGCo	Output limiting	60	68	10	2	20
NRA01	MMSE-STSA84 [60]					
NRA02	MBSS02 [116]					

Outside the audiometric room, a Windows PC (Dell Inc., Intel Core i7-3770 CPU) running Matlab R2013a (64-bit) was used to process signals and play sounds through a USB audio interface (TASCAM US-366). Inside the room, sounds were delivered through a set of headphones (TDH-39P). The system was calibrated on a KEMAR manikin (G.R.A.S., Head and Torso Type 45BA), with a Type 2 IEC711 Ear Simulator (G.R.A.S., RA0045-S4) and a soft pinna (G.R.A.S., KB1065) using a Class 1 Handheld Analyzer (Brüel & Kjær, Type 2250).

6.1.1.2 SRT measurement procedure

A fixed-masker paradigm was used with the masker level fixed at 65 dBA, and the speech level adaptively varied to achieve a desired presentation SNR. For each condition, a HINT list consisting of 20 sentences is chosen. The adaptive protocol is divided in two phases: In the first phase, a coarse estimate of the subject's threshold is calculated from the first 4 sentences using a large step size. The second phase begins with sentence #5 presented at the threshold determined in phase 1, and uses a smaller step size to get a more precise estimate of the threshold. The procedure is summarized as follows:

- Sentence #1 is played at the starting SNR and repeated until the subject repeats all words correctly, increasing the level by 4 dB for each incorrect response. Once a correct response is received, the SNR is decreased by 4 dB for the next sentence.
- Sentences #2–4: For these sentences, the SNR is increased by 4 dB after each incorrect response, and decreased by the same step size after each correct response.
- Sentence #5 is played at a SNR computed as the mean of the final level of the first sentence, the levels of sentences #2–4, and the level at which the fifth sentence would have been presented based on the subject's response to sentence #4.
- Sentences #6–20: For these sentences, the SNR is increased by 2 dB after each incorrect response, and decreased by the same step size after each correct response.

At the end of the list, the SRT for the given test condition is computed as the mean of the levels of sentences #5–20 and the level at which the twenty first sentence would have been presented based on the subject's response to sentence #20. The procedure, which is similar to the HINT protocol [153], corresponds to a sentence SRT measurement, since all the words of a test sentence must be recognized for a correct response to be registered.

6.1.2 Experimental conditions

With a limited number of test lists per session (12 lists, *c.f.* Section 4.4.2.2), the experimental task is broken down into four experiments with four groups of subjects focusing on spatial listening, continuous vs. intermittent masking, and different hearing devices.

6.1.2.1 Experiment A: Spatial release from masking

The goal of the first experiment is to study spatial release from masking (SRM) in the absence of hearing device processing. SRT measurements are performed in a simulated anechoic room (no reverberation) with a frontal speech source while virtually displacing the noise source with the use of HRTFs. Table 6.2 lists the testing conditions for this

Table 6.2: *Listening conditions tested in Experiment A. These conditions focus on spatial listening using a fixed frontal speech source and a continuous masker at different locations.*

Condition	Configuration	Device	Masker	Starting SNR (dB)
C01	S_0N_0	No device	SSNOISE	-2
C02	S_0N_{30}	No device	SSNOISE	-5
C03	S_0N_{60}	No device	SSNOISE	-10
C04	S_0N_{75}	No device	SSNOISE	-7
C05	S_0N_{90}	No device	SSNOISE	-7
C06	S_0N_{105}	No device	SSNOISE	-7
C07	S_0N_{120}	No device	SSNOISE	-10
C08	S_0N_{150}	No device	SSNOISE	-5
C09	S_0N_{180}	No device	SSNOISE	-2
C10	S_0N_{d1}	No device	SSNOISE	-10
C11	S_0N_{d2}	No device	SSNOISE	-10

experiment. The order of test conditions and sentence lists was randomized for each subject. Masker azimuths from 0° to 180° are considered in 30° increments. In order to investigate head diffraction effects (*c.f.* Section 6.3.1), two additional measurements are performed in the vicinity of 90° (conditions C04 and C06). Two conditions with diffuse maskers are also considered. Diffuse maskers are generated by mixing uncorrelated noise, filtered through a number of HRTFs. Condition S_0N_{d1} corresponds to a horizontally diffuse masker mixed from six different azimuths (30°, 90°, 150°, 210°, 270°, 330°). The masker in condition S_0N_{d2} is a 3-dimensional diffuse masker with additional sources at 50° elevation (at azimuths 0°, 64°, 120°, 176°, 240°, 296°) and above the listener's head (90° elevation).

6.1.2.2 Experiment B: Clipping algorithms

Two clipping algorithms were tested in the next experiment: a peak clipper and a center clipper. Since these are not signal-adaptive operations, they were tested in continuous masking only. Ten conditions were tested in this experiment as listed in Table 6.3. These are arranged into three groups defined by the type of clipping operation. The order of conditions and HINT lists was counterbalanced across subjects (*c.f.* Section 6.1.2.5).

Table 6.3: *Listening conditions tested in Experiment B: Two clipping algorithms with different thresholds, are tested in two spatial configurations under continuous masking.*

Condition	Configuration	Device	Masker	Starting SNR (dB)
C01	S_0N_0	No device	SSNOISE	-2
C02	S_0N_{90}	No device	SSNOISE	-7
C03	S_0N_0	PC50	SSNOISE	2
C04	S_0N_{90}	PC50	SSNOISE	-3
C05	S_0N_0	PC60	SSNOISE	2
C06	S_0N_{90}	PC60	SSNOISE	-3
C07	S_0N_0	CC50	SSNOISE	-2
C08	S_0N_{90}	CC50	SSNOISE	-7
C09	S_0N_0	CC60	SSNOISE	-2
C10	S_0N_{90}	CC60	SSNOISE	-7

Two clipping thresholds (50 and 60 dB SPL) were considered for each type of distortion following informal listening tests. Peak clipping is typically encountered as a hard output limiter and would require testing with signals that exceed the clipping threshold (typically set around 90 dB SPL or more). However, subjecting listeners to such elevated presentation levels for an extended duration of time is straining. Therefore, an output limiting scenario is simulated by testing with a 65 dBA fixed masker level and reducing the clipping threshold. The threshold of 60 dB SPL was selected in order to cause audible distortion effects. The threshold of 50 dB SPL was selected as a limit below which speech was unintelligible and the convergence of the adaptive procedure towards became impossible. Center clipping removes low-level signal components, which is often thought of as a crude approximation of NRAs (*c.f.* Section 4.2.2.3). Thresholds below 50 dB SPL were found not to produce enough audible distortion, whereas thresholds above 60 dB SPL made the speech barely intelligible and prevented the test from converging.

Finally, peak clipping is a harsh distortion which clips the high-level components of the signal. As a result, the level of the peak-clipped signal is significantly lower than the signal at the input of the device. In order to quantify the effect of peak-clipping distortion on speech intelligibility compared to baseline conditions (C01 and C02), it is

Table 6.4: *Listening conditions tested in Experiment C: Two compression algorithms are tested in two spatial configurations under continuous and intermittent masking.*

Condition	Configuration	Device	Masker	Starting SNR (dB)
C01	S_0N_0	No device	SSNOISE	-2
C02	S_0N_{90}	No device	SSNOISE	-7
C03	S_0N_0	No device	SSINT16	-12
C04	S_0N_{90}	No device	SSINT16	-17
C05	S_0N_0	AGCi	SSNOISE	-2
C06	S_0N_{90}	AGCi	SSNOISE	-7
C07	S_0N_0	AGCi	SSINT16	-12
C08	S_0N_{90}	AGCi	SSINT16	-17
C09	S_0N_0	AGCo	SSNOISE	-2
C10	S_0N_{90}	AGCo	SSNOISE	-7
C11	S_0N_0	AGCo	SSINT16	-12
C12	S_0N_{90}	AGCo	SSINT16	-17

imperative to isolate this effect from the impact of the reduced signal level. Therefore, in conducting this experiment, the processed signal was amplified by 12 dB for conditions C03 and C04 and by 5 dB for conditions C05 and C06 before presenting it to the listener. These gains were derived empirically in order to normalize the clipped masker to 65 dBA, when a masker is presented at the same level to the input of the simulated hearing device (in the absence of speech). This is unique to peak-clipping conditions, as the problem is not encountered for center clipping or any of the distortions in the other experiments.

6.1.2.3 Experiment C: Compression algorithms

Two AGC systems were tested in the next experiment: an input-controlled (AGCi) and an output-controlled (AGCo) system. The test conditions have been divided into three groups as listed in Table 6.4. For each group, four conditions were considered using the continuous and intermittent maskers at 0° and 90° azimuths. The order of conditions and HINT lists was also counterbalanced across subjects.

The static and dynamic parameters for the two compression amplifiers are listed in Table 6.1. For the AGCi system, parameters typically used for intersyllabic compression

Table 6.5: *Listening conditions tested in Experiment D: Two noise reduction algorithms are tested in two spatial configurations under continuous and intermittent masking.*

Condition	Configuration	Device	Masker	Starting SNR (dB)
C01	S_0N_0	No device	SSNOISE	-2
C02	S_0N_{90}	No device	SSNOISE	-7
C03	S_0N_0	No device	SSINT16	-12
C04	S_0N_{90}	No device	SSINT16	-17
C05	S_0N_0	NRA01	SSNOISE	0
C06	S_0N_{90}	NRA01	SSNOISE	-5
C07	S_0N_0	NRA01	SSINT16	0
C08	S_0N_{90}	NRA01	SSINT16	-5
C09	S_0N_0	NRA02	SSNOISE	0
C10	S_0N_{90}	NRA02	SSNOISE	-5
C11	S_0N_0	NRA02	SSINT16	0
C12	S_0N_{90}	NRA02	SSINT16	-5

were employed. For the AGCo system, dynamic parameters (compression ratio, and attack/release times) typically used for a soft output limiter were used. However, similar to the case with peak clipping (*c.f.* Section 6.1.2.2), the threshold was set to 60 dB SPL, and testing was carried out at 65 dBA, instead of a louder masker. In fact, convergence of the adaptive protocol could not be guaranteed at high presentation levels, especially in intermittent masking conditions which required very low SNR levels[2].

6.1.2.4 Experiment D: Noise reduction algorithms

Two NRAs popular in the speech enhancement literature were tested in the last experiment. A Bayesian statistical-model (NRA01: MMSE-STSA84 [60]) and a multiband spectral-subtraction (NRA02: MBSS02 [116]) algorithm were selected. Table 6.5 lists the testing conditions for this experiment, which are also divided into three groups. The presentation order was also counterbalanced across groups of subjects.

[2]Informal trials have shown that, with the masker level fixed at 85 dBA, as the speech level was increased (to achieve a lower SNR), the AGCo would limit the increased speech signal. As a result, a threshold was never reached as subjects failed to repeat sentences correctly, even at unreasonably high speech levels (*e.g.*, 130 dBA).

6.1.2.5 Counterbalancing

The order of experimental conditions and HINT list presentation was randomized in Experiment A. For the other experiments, where test conditions are grouped according to the simulated hearing device, the order was counterbalanced among subjects as follows: all subjects started with the base conditions (no device). Next, half the subjects for each experiment were tested with the first distortion group (peak clipping, AGCi, or NRA01) then with the second (center clipping, AGCo, or NRA02). The second half of the subjects were tested in the opposite order. Similarly, within each group, conditions were also counterbalanced in subgroups, first according to distortion threshold (Experiment B) or masker type (Experiments C and D), then according to masker azimuth. The base group for Experiment B consists of only 2 test conditions, which were counterbalanced according to masker azimuth. Finally, the order of list presentation was selected for each subject such that no list was presented for the same condition across subjects.

6.1.3 Subjects

A total of 32 normal-hearing subjects, aged 20 to 28 yr, were recruited to participate in the experiments described above. Thus, each experiment comprised 8 subjects. Prior to the test, each subject was screened through an otoscopic examination of the ear canal, tympanometry, as well as pure-tone audiometry. Subjects with abnormal visual evaluation (*e.g.*, evidence of ear infection, cerumen accumulation, etc.) or middle-ear pressure reading, or with hearing thresholds over 25 dB HL for any of the audiometric frequencies in any ear were excluded from the study. All subjects who presented themselves for testing were compensated for their participation. Testing was conducted under a certificate for research with human subjects from the Office of Research Ethics and Integrity at the University of Ottawa. The data collection was carried out by two masters' students in audiology (M.H.Sc.) at the University of Ottawa.

6.2 Results

A total of 360 SRT measurements were performed during this experimental work. A listing of all the measured SRTs, along with the mean, standard deviation (SD), and 95% confidence interval (CI) for each experimental condition is provided in Appendix B. According to the adaptive experimental procedure, each SRT measurement is an average of 17 levels (*c.f.* Section 6.1.1.2). Each individual SRT for which the SD was above 3.5 dB was considered to be inaccurate (the convergence of the procedure for such measurements was questionable), and was therefore discarded. A total of 9 such measurements are indicated in the tables in Appendix B.

The results of experiment A are shown in Figure 6.1. The left panel shows the mean SRT measured across 8 subjects under different spatial configurations (*c.f.* Table 6.2). The error bars represent the SD. The right panel shows the SRM, *i.e.*, the improvement in SRT with respect to the condition S_0N_0 (masker at 0° in Figure 6.1). The results of experiments B, C and D are reported in Figure 6.2. In these experiments, two spatial configurations (S_0N_0 and S_0N_{90}) and different hearing devices are simulated as outlined in Tables 6.3 to 6.5. For experiments C and D, continuous (SSNOISE) and intermittent (SSINT16) maskers are considered. In the figure, the mean SRT is shown with the SD in the left panel for each experiment. SRM is presented in the middle column and masking release due to the device under test in the right column. The latter is computed as the difference between the SRT for each of the base conditions (no device) and the corresponding measurement with the device. Positive values of masking release indicate improved speech intelligibility due to the spatial configuration or device under test, whereas negative values indicate loss of intelligibility.

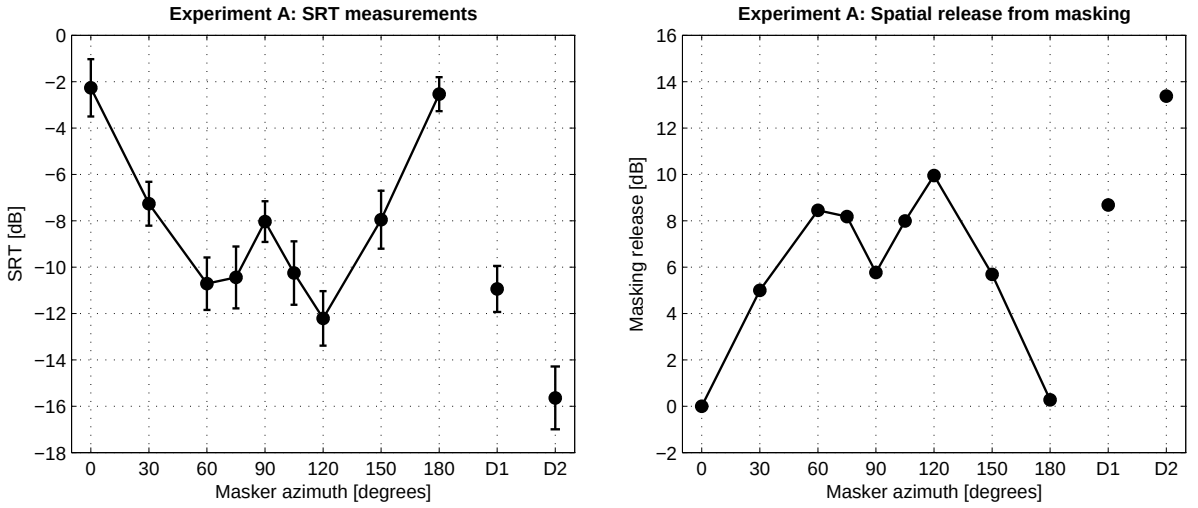


Figure 6.1: *Subjective SRTs from Experiments A showing the effect of noise azimuth and two diffuse maskers (speech azimuth is* 0°*). Spatial release from masking with respect to the* 0° *masker condition (S_0N_0) is shown in the right panel.*

6.3 Discussion

6.3.1 Spatial listening

Spatial listening ability can be quantified by studying the SRM (*c.f.* Section 3.3.1). Figure 6.1 reports the results for continuous masking at several masker azimuth and two diffuse masking conditions. The worst spatial configuration for speech intelligibility is when the speech and masker sources are co-located in the front: S_0N_0. The SRT is approximately the same in the S_0N_{180} condition, when the noise is placed directly behind the listener (SRM of 0.3 dB). Lateral movement of the masker source leads to improved intelligibility with a maximum masking release of approximately 10 dB around 120°. The maximum does not occur at 90° as one would expect. This is due to a common feature of the reproduction of sounds through headphones using KEMAR-based HRTFs, which will be further discussed below. When using a horizontally diffuse masker (D1), masking release is about 1.3 dB below the maximum SRM at 120°. For the 3-dimensional diffuse masker (D2), the additional vertical spatial cues lead to an increased SRM of 13.3 dB.

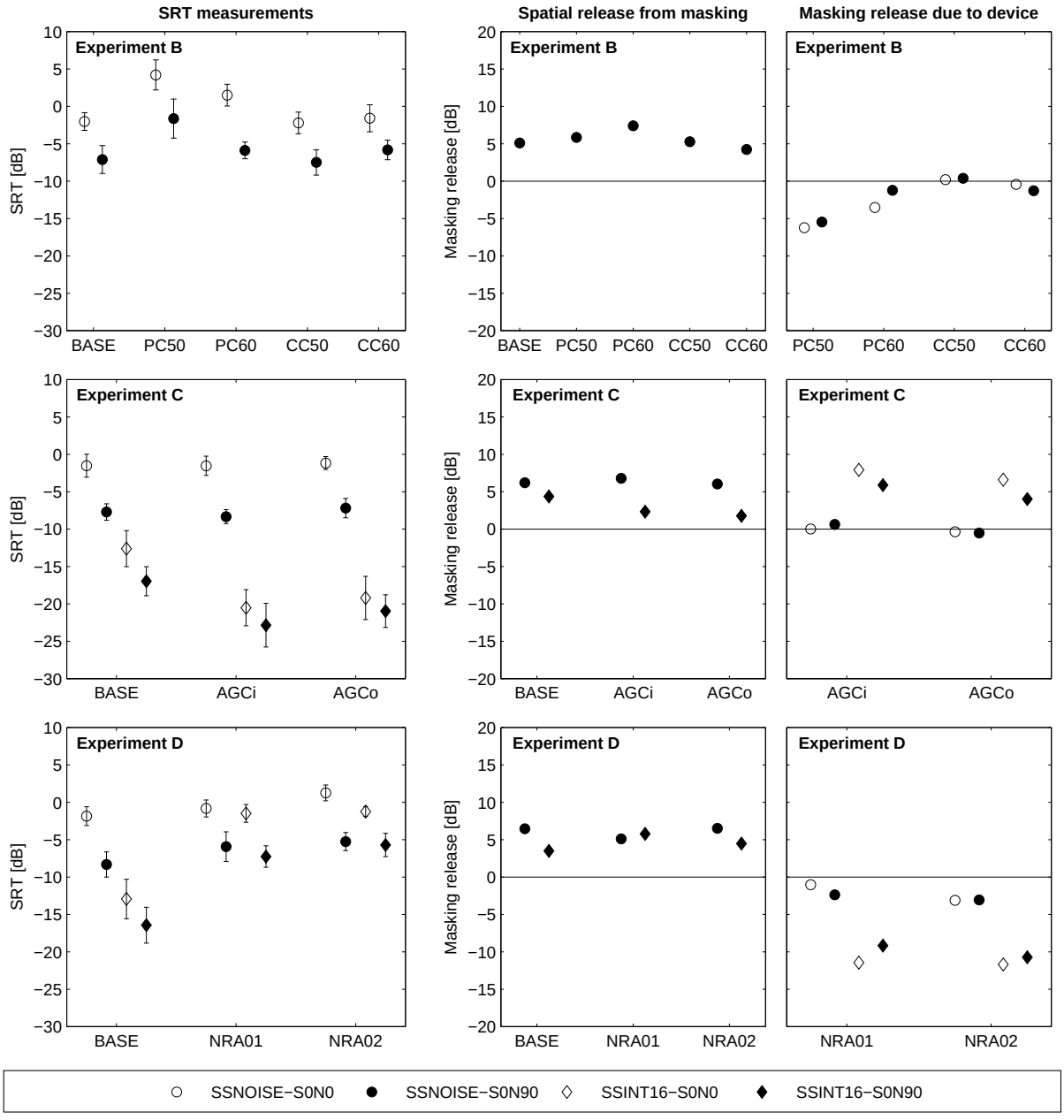


Figure 6.2: *Subjective SRTs from Experiments B, C and D showing the effects of clipping (top row), compression (center row), and noise reduction (bottom row) algorithms. For each experiment, masking release due to spatial configuration and device under test are shown in the center and right panels respectively. "BASE" conditions indicate no device processing.*

Table 6.6: *Comparing SRT and SRM for baseline conditions across experiments with normative data for the HINT [204]. All values are in dB.*

	Exp A		Exp B		Exp C		Exp D		All		Norms	
	SRT	SD	SRT	SD	SRT	SD	SRT	SD	SRT	SD	SRT	SD
S_0N_0	-2.3	1.2	-2.0	1.2	-1.5	1.5	-1.9	1.3	-1.9	1.3	-2.6	1.0
S_0N_{90}	-8.0	0.9	-7.1	1.9	-7.7	1.1	-8.3	1.7	-7.8	1.5	-10.1	1.3
SRM	5.5	1.2	5.1	1.2	6.2	1.7	6.5	2.1	5.9	1.6	7.5	

6.3.1.1 Comparison with HINT norms

Conditions S_0N_0 and S_0N_{90} were tested in all 4 experiments under continuous masking. The mean measured SRT and SD for these conditions from each experiment as well as an average computed across all 32 subjects are compared to normative data [204] for HINT in Table 6.6. Also included in the comparison is the computed SRM between the two conditions. The mean measured SRTs in our experiments are on average 0.7 dB above the norms in the S_0N_0 condition, and 2.3 dB below the norms in the S_0N_{90} condition, which leads to a lower SRM. The differences observed are likely due to the processing of sounds through different sets of HRTFs. Our implementation uses the HRTFs measured on KEMAR in an anechoic room from the MIT Media Lab [73] (*c.f.* Section 4.2.3). The HINT system uses HRTFs measured on KEMAR at 0° and ±90°, however, not in an anechoic room. Therefore, to eliminate acoustic effects in the environment, the measured HRTFs were normalized by a reference recording in the same room with KEMAR removed, using a flat-response microphone positioned at the center of the KEMAR head [185]. The authors observed that this approach tends to produce "similar or larger amounts of spatial release from masking" when compared with sound-field measurements.

6.3.1.2 Comparison with other SRM data

The left panel of Figure 6.3 presents the SRM data from Experiment A as a function of masker azimuth (excluding the diffuse masking conditions) compared with similar data published in the literature [12, 22, 163, 165, 166]. With the exception of 90°, the results

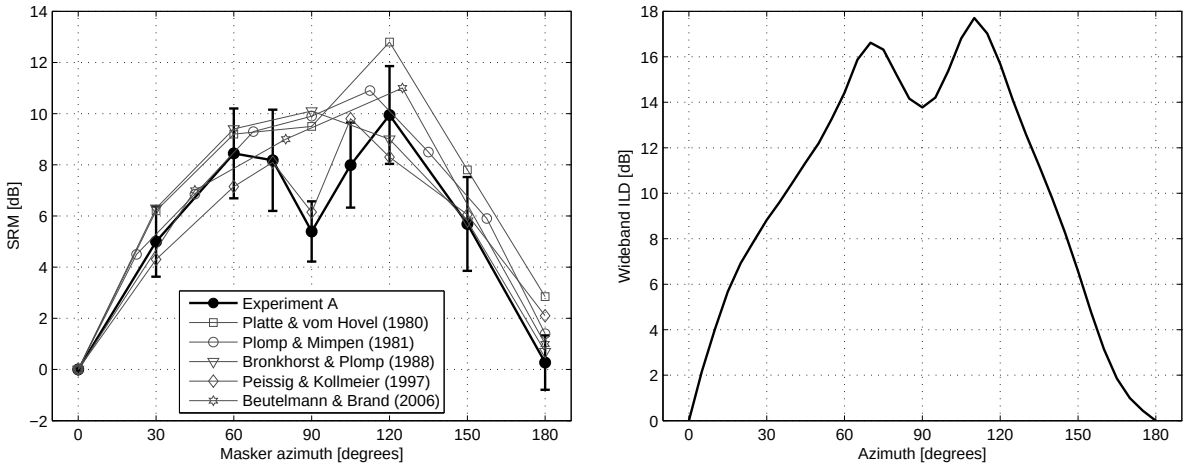


Figure 6.3: *Comparison of SRM results with data from different studies (left panel). Wideband ILD for KEMAR-based HRTF as a function of noise azimuth (right panel).*

seem to fit well within the range of values reported from all 5 studies compared to in the figure. All studies, except for the work of Bronkhorst and Plomp [22], report maximal binaural benefit when the masker is located behind the interaural axis (azimuth between 105° and 120°), as we have observed in our own measurements.

The drop in SRT at 90° can be attributed to a decreased ILD due to diffraction effects when the sound source is aligned with the interaural axis: the head acts as a diffraction object to the incident sound wave which leads to increased interaural cross talk at the contralateral ear [163, 213]. Using the set of HRTFs from MIT (the same as the ones used in this work), Xie shows that this effect is frequency dependent and most pronounced for frequencies between 1.4 and 2.8 kHz [213]. A plot of wideband ILD as a function of azimuth for this set of HRTFs is shown in the right panel of Figure 6.3. It is computed as the ratio of the wideband power of the HRTF measured at the right and left ears for each azimuth. The ILD plot features a drop at 90° azimuth which accounts for a drop of about 3 dB in the measured SRTs. The asymmetry with respect to the interaural axis illustrated in the ILD plot is also consistent with the reported measurements.

The dip in the SRM plot at 90° is consistent with the earphone measurements reported in [163], one of the studies shown in Figure 6.3. Both this and our studies used "raw"

HRTFs to simulate binaural hearing. Two other studies were conducted in the sound field [165, 166]. The SRT trend in those studies exhibits an inflection point instead of a dip, which still reflects a reduced binaural benefit around 90° azimuth. This may be due to the averaging effects of small movements of the subject's head when listening in sound field. The measurements in [22] were conducted under earphone listening using KEMAR-based HRTFs which had been normalized with an equalization filter as suggested in [129]. This normalization appears to compensate for the diffraction effects and also remove the asymmetry of the measured SRTs with respect to the interaural axis. Finally, in [12], the azimuth axis was sampled in a way to avoid these effects.

6.3.2 Continuous vs. intermittent masking

Next, measurements with continuous and intermittent masking in experiments C and D (middle and bottom rows in Figure 6.2) are compared. We focus on the baseline conditions C01 to C04 in both experiments (no hearing device). First, we note that the standard deviation of the measurements with the intermittent masker (2–2.5 dB) is larger than for the continuous masker (1.1–1.7 dB). Similar observations have been frequently reported in the literature (*e.g.*, [68, 135, 147, 164, 170–172]).

Two additional observations arise from the SRTs for conditions C01 to C04. First, we compute the masking release due to the intermittent masker for both spatial configurations (difference between the SRTs for C01 and C03 at S_0N_0, and for C02 and C04 at S_0N_{90}). A larger release is observed in the S_0N_0 condition (around 11 dB) than in the S_0N_{90} condition (between 8–9 dB). Moreover, it seems that listeners experienced larger SRM benefit under continuous masking (around 6.5 dB) than under intermittent masking (3.5–4.4 dB). These findings are consistent with recently published measurements using the Canadian French version of the HINT [135]. Experiments using an intermittent masker are designed to test the ability to listen in the "dips". These observations suggest that listeners tend to benefit more from dip listening in a difficult spatial configuration (*e.g.*, S_0N_0) compared to an easier one (*e.g.*, S_0N_{90}). An alternative perspective

Table 6.7: *Comparing the projected percentage increase/decrease in speech intelligibility obtained with peak and center clipping under continuous masking with respect to the corresponding baseline condition (no device).*

		% intelligibility change			# Subjects with
Condition	Device	Mean	SD	Range	improvement
S_0N_0	PC50	−62	19	−104 to −42	0/8
S_0N_0	PC60	−38	11	−52 to −20	0/6[3]
S_0N_0	CC50	2	10	−21 to +9	7/8
S_0N_0	CC60	−4	16	−20 to +25	3/8
S_0N_{90}	PC50	−55	21	−69 to −9	0/8
S_0N_{90}	PC60	−12	10	−24 to −1	0/8
S_0N_{90}	CC50	4	15	−10 to +33	3/8
S_0N_{90}	CC60	−13	17	−32 to +16	3/8

is that listeners tend to take advantage of binaural cues more effectively under a more difficult masking condition (*e.g.*, continuous masking) compared to an easier one (*e.g.*, intermittent masking). In summary, it appears that the benefits due to binaural cues and dip listening are complementary but not additive.

6.3.3 Effects of distortion

6.3.3.1 Clipping algorithms

The results reported in the top row of Figure 6.2 show the effects of the peak and center clipping algorithms with the different clipping thresholds considered. The change in SRT can be expressed as a percentage increase or decrease in speech intelligibility using the 10% per dB slope of the performance-intensity function for the American English HINT [204]. The mean, SD and range of projected percentage change in speech intelligibility for each condition are reported in Table 6.7. The table also lists the number of subjects (out of the total for each given condition) who showed improvement in speech intelligibility compared to the corresponding baseline condition.

[3]The data points for two subjects were discarded for this condition due to an elevated measurement standard deviation (*c.f.* Appendix B).

Table 6.8: *Comparing the projected percentage increase/decrease in speech intelligibility obtained with AGCi and AGCo under continuous masking with respect to the corresponding baseline condition (no device).*

Condition	Device	% intelligibility change Mean	SD	Range	# Subjects with improvement
S_0N_0	AGCi	0	25	−33 to +34	4/8
S_0N_0	AGCo	−4	20	−21 to +28	4/8
S_0N_{90}	AGCi	6	10	−8 to +22	5/8
S_0N_{90}	AGCo	−5	16	−26 to +18	3/8

Overall, all listeners showed reduced intelligibility with peak clipping conditions as reflected in the elevated SRTs. Perceptually, the peak-clipped signals sound highly distorted, with the distortion increasing as the clipping threshold is decreased: the greater loss in intelligibility reported for PC50 is consistent with the fact that the clipper is more active when using a lower clipping threshold. For the center-clipping scenarios considered, the range of results obtained shows no significant effect on speech intelligibility. Perceptually, while the processed signals sound cleaner, this benefit appears to be offset by the suppression of low-level portions of speech at the end of each syllable. However, the results suggest that using a lower clipping threshold (50 dB) offers a better trade off between these two effects with respect to speech intelligibility. Finally, we note that in all clipping conditions considered the spatial cues which lead to positive SRM values appear to be preserved in the processed signals (*c.f.* Figure 6.2, top row, center panel).

6.3.3.2 Compression algorithms

The results reported in the middle row of Figure 6.2 show the effects of the two compression algorithms considered. Under continuous masking, speech intelligibility seems to be unaffected by the two algorithms. This is also reflected in Table 6.8, where the projected percentage change in speech intelligibility with respect to the baseline is listed for both algorithms under the two spatial configurations considered. While the AGCi system seems to modestly outperform the AGCo system in the S_0N_{90} condition on aver-

age, the results in the table show that there is no evidence that either systems provides a significant benefit to speech intelligibility in continuous noise. Similar findings have been previously reported in the literature under comparable conditions [41, 131]. In addition, it is worth noting that the SRTs reported for AGCo are lower than those reported for the PC60 device from Experiment B. The reduced intelligibility under peak clipping explains why compression amplifiers have become the method of choice to limit the output of modern hearing devices (*c.f.* Section 2.4.3).

While no clear intelligibility benefit of AGC processing is observed under continuous masking, an important finding from the results reported in this work is that both compression algorithms have led to large improvements in speech intelligibility under intermittent masking conditions. This is most clearly illustrated in the right-most middle panel in Figure 6.2. In the S_0N_0 condition, for example, masking release due to the AGCi and AGCo algorithms is around 7.9 and 6.6 dB respectively. To explain these improvements, Figure 6.4 illustrates the effect of each algorithm on the unprocessed signals for two SNR levels, corresponding approximately to the SRT in the no processing condition (-12 dB, left panel) and with the compression algorithms (-20 dB, right panel). In addition to the unprocessed and processed signals, the figure displays the average power in the ON and OFF segments expressed in dB SPL, as well as the power ratio of OFF to ON segments. Whereas only a short section of speech from the first HINT sentence is displayed, the average powers are computed over all the (ON or OFF) segments from a signal containing a concatenated sequence of all 20 sentences of the first HINT list masked by an intermittent signal of equal length. The OFF-to-ON ratio provides an estimate of the SNR in each scenario, since the OFF segments consist of the speech-only signal and the ON segments are dominated by noise. Changes in the OFF-to-ON ratios[4] reported

[4]Changes in the ON-to-OFF ratio depend on the behaviour of the algorithms in each segment and the SNR of the speech and noise mixture. During ON segments, the signal power is on average equal to the compressors' zero-gain reference level (68 dB SPL). Any power difference in these segments compared to the unprocessed signals can be attributed to the compressors' release times affecting the beginning of each ON segment. During OFF segments, the signal level under both SNR scenarios falls within the compression range of the AGCi system. For the AGCo, the average signal level is in the compressor's linear range, however, some portions of the speech signal will fall in the compression range.

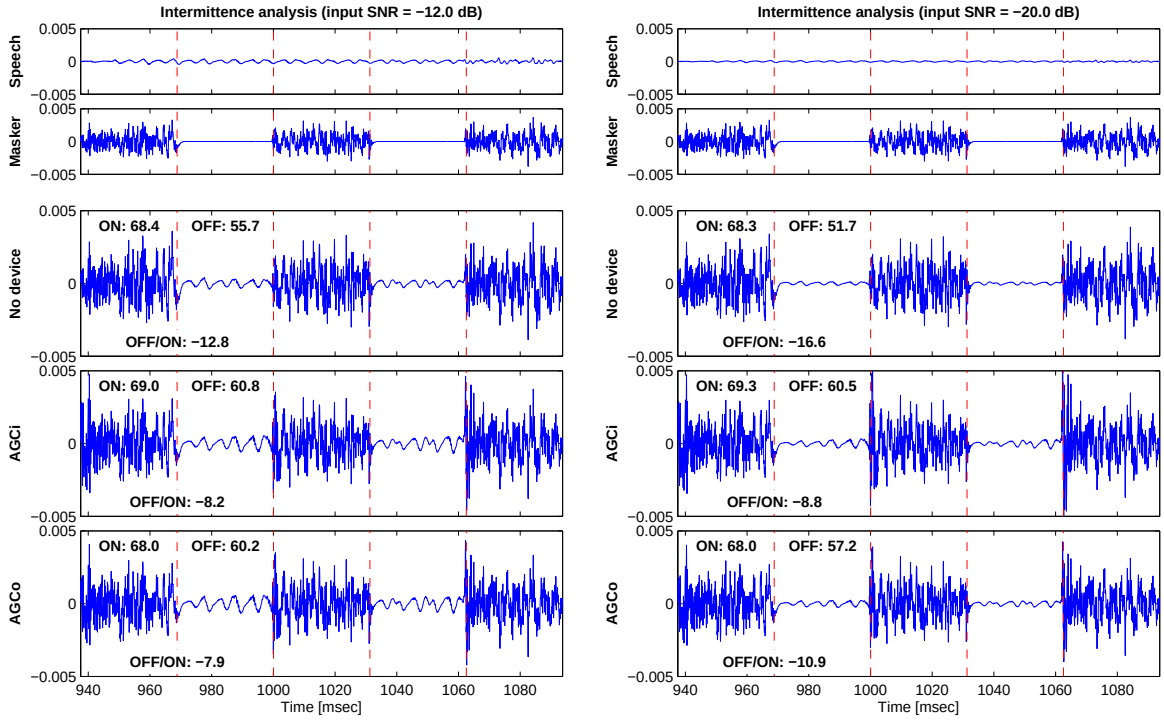


Figure 6.4: *Unprocessed speech, noise, and mixed signals, as well as signals processed through AGC algorithms in the S_0N_0 condition for the intermittent masker. Input SNRs of –12 (left panel) and –20 dB (right panel) are chosen, corresponding approximately to the SRT for the no processing and AGC conditions respectively. The average power in the ON and OFF segments expressed in dB SPL, and the power ratio of OFF to ON segments are also shown for the unprocessed and processed signal mixtures.*

in the figure account, in large part, for the improvements in the measured SRT: for the -20 dB SNR simulation, the improvement in the OFF-to-ON ratio due to the AGCi and AGCo algorithms are about 7.8 and 5.7 dB, which is comparable to the masking release observed in Figure 6.2. Other contributing factors include forward and backward masking effects, as well as the dynamic behaviour of the compression algorithms in response to fast fluctuations of the signals within the segments which are not entirely captured in the average powers. A similar situation is likely at play to explain the large benefits also observed at S_0N_{90} for the AGCi and AGCo algorithms in intermittent masking.

Not accounting for the dynamic range of speech (-12 to 18 dB around the mean), the static gain applied by the AGCi and AGCo systems for the -12 dB SNR simulation can be calculated to be 7.4 and 7.2 dB respectively. In the -20 dB SNR simulation the respective gains are around 9.8 and 7.2 dB.

Finally, we note the effects of the compression algorithms considered on the SRM in Figure 6.2 (middle row, center panel). The SRM seems to be unaffected by the algorithms under continuous masking. However, when speech intelligibility is improved under intermittent masking, it seems that listeners tend to decrease their reliance on binaural cues for speech perception (SRM values are smaller for the processed signals than the corresponding base condition without processing).

6.3.3.3 Noise reduction algorithms

The results reported in the bottom row of Figure 6.2 show the effects of the two NRAs considered. All masking release values in the right panel are negative suggesting that, on average, neither of the algorithms produced an improvement in speech intelligibility. This is in line with similar findings in the literature where no single-microphone algorithm has been found to significantly improve speech intelligibility [100, 197] and often led to reduced intelligibility. Further, under continuous masking and in the most basic spatial configuration where binaural benefits are minimal (S_0N_0), the MMSE-STSA84 algorithm (NRA01) seems to outperform the MBSS02 algorithm (NRA02). This is, at least qualitatively, in line with the results reported by Hu and Loizou in [100], where the effects of several algorithms on speech intelligibility were investigated[5].

Table 6.9 lists the mean, SD, and range of projected percentage change obtained across all 8 subjects for each of the two algorithms under continuous masking conditions with respect to the corresponding baseline condition. A straightforward quantitative comparison of these numbers with results reported in the literature using NRA01 is not possible: the selection of speech material, masker and SNR used in other studies have significant effects both on the behaviour of the algorithms and speech intelligibility[6]. For NRA02, Kamath used sentences from the HINT corpus mixed with speech-shaped noise

[5]The list of algorithms in [100] does not include the MMSE-STSA84 algorithm [60], but did include the revised logMMSE-STSA85 algorithm [61], which only differs from its predecessor in that it minimizes the MMSE of the log-spectra when enhancing noisy speech. In an earlier study [99], these two algorithms received similar ratings with respect to signal distortion, background intrusiveness, and overall quality. The two are often cited in the literature as part of a family of algorithms [6, 89, 176].

Table 6.9: *Comparing the projected percentage increase/decrease in speech intelligibility obtained with NRA01 and NRA02 under continuous masking with respect to the corresponding baseline condition (no device).*

Condition	Device	% intelligibility change Mean	SD	Range	# Subjects with improvement
S_0N_0	NRA01	−10	15	−33 to +9	3/8
S_0N_0	NRA02	−31	20	−50 to +13	1/8
S_0N_{90}	NRA01	−24	21	−64 to +5	1/8
S_0N_{90}	NRA02	−31	25	−55 to +18	1/8

at a 0 dB SNR and found an average decrease of 22% in percentage word correct scores [117]. Only a single subject showed an improvement in speech intelligibility, and the effect of the algorithm ranged from a 13% improvement to a degradation of 55%. Our results for the S_0N_0 condition in Table 6.9 fall within a very similar range.

When using the intermittent masker (SSINT16), both algorithms cause a large degradation in speech intelligibility: for example, SRTs in the S_0N_0 condition are increased by 11.4 and 11.7 dB for NRA01 and NRA02 respectively. Musical noise is audibly present when listening to the processed signals, even for NRA01, which has been shown to eliminate musical noise under stationary and poorly stationary maskers [27]. For both algorithms, the estimation of the noise spectrum is based on the assumption of a slowly varying noise environment: it is initially estimated based on a segment of pure noise, assumed to be virtually stationary during speech activity, and updated during speech-absent segments detected by a VAD. In a highly non-stationary masker such as SSINT16, it appears that updating the noise estimate fails and the benefit of dip lis-

[6]For example, Hu and Loizou used the IEEE sentences [192] in their intelligibility study. They did not consider a speech-shaped masker, but instead used "babble", "car", "street" and "train" noises mixed at 0 and 5 dB SNR [100]. Taal et al. performed speech intelligibility measurements using NRA01 with speech material from the Danish Dantale II corpus [208] corrupted by speech-shaped and cafeteria noise. Subjects' percentage correct scores are reported for the set of SNR values {-8.9, -7.7, -6.5, -5.2, -3.1} [195, 197]. Approximately 53% intelligibility is achieved with the unprocessed speech-shaped noise corrupted signal at -8.9 dB SNR, and around 95% at -3.1 dB. The performance of the algorithm compared to the unprocessed noisy speech is drastically different at each of these SNR, an observation consistent with the findings of Hilkhuysen et al. [89].

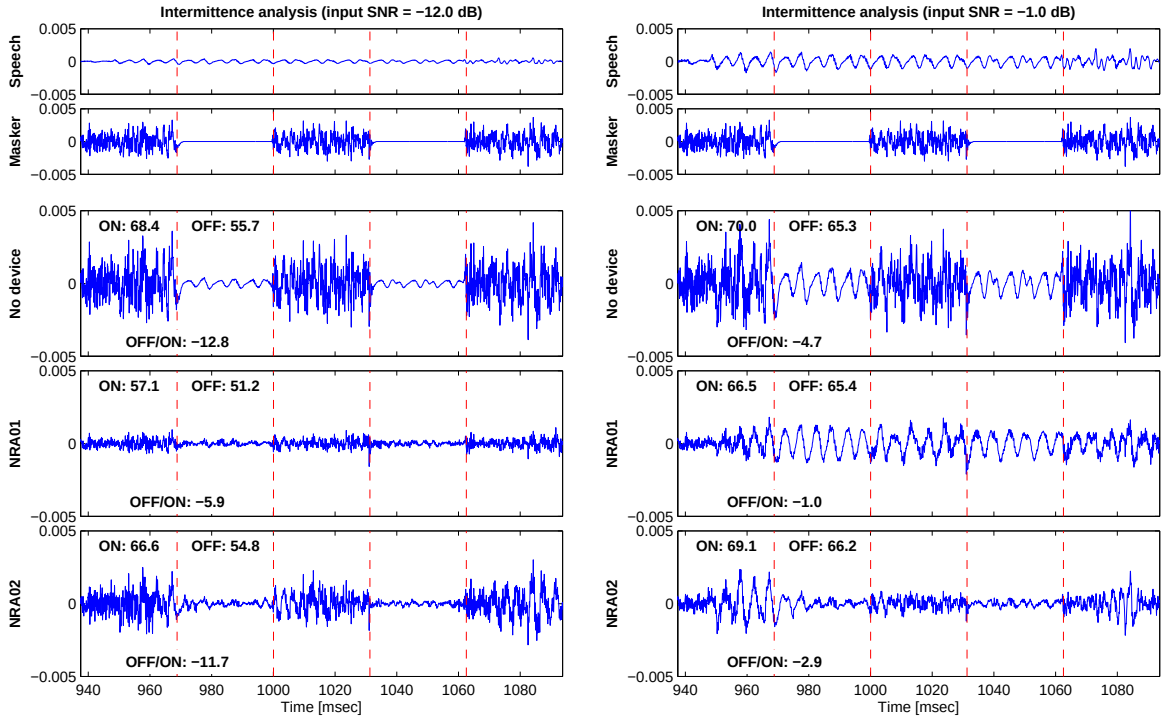


Figure 6.5: *Unprocessed speech, noise, and mixed signals, as well as signals processed through noise reduction algorithms in the S_0N_0 condition for the intermittent masker. Input SNRs of -12 (left panel) and -1 dB (right panel) are chosen, corresponding approximately to the SRT for the no processing and NRA conditions respectively. The average power in the ON and OFF segments expressed in dB SPL, and the power ratio of OFF to ON segments are also shown for the unprocessed and processed signal mixtures.*

tening is impaired. Figure 6.5 provides a visual illustration using a speech and noise mixture processed through each of the two algorithms. Two SNRs are simulated which approximately correspond to the SRT in the no processing condition (-12 dB, left panel), and with the NRAs (-1 dB, right panel). Several observations can be made from this figure. First, the changes in the OFF-to-ON ratio cannot explain the decreased intelligibility due to the NRAs. In fact, the ratio changes reported are in the positive direction (+3.7 and +1.8 dB for NRA01 and NRA02 respectively). This is in contrast to the AGC systems where positive changes in the OFF-to-ON ratio correctly predicted the observed masking release. Second, at -12 dB SNR, there seems to be a spill over of residual noise from the ON segments to the OFF segments in the processed signals for

both algorithms. This appears to happen also at -1 dB SNR for NRA02. This could indicate that the operation of the NRA and/or VAD is occasionally breaking down under these conditions for which they were not designed. Third, it is clearly apparent that, unlike for AGC systems, the nature of the residual speech signal in the gaps is severely affected by the algorithm. Last, the performance of the algorithms depends on the SNR as well as the type of speech segments (voiced vs. unvoiced). The portion of speech shown in the figure corresponds to a voiced-speech segment (underlined) from the first HINT sentence: "(A/The) boy fell from the window." Other portions of the signal (not shown here) demonstrate that the algorithms act differently on lower-energy unvoiced segments (such as the consonant /f/ in 'fell') and higher-energy voiced segments (such as the vowel /o/ in the word "boy"). It is possible that the changes in the OFF-to-ON ratio observed are dominated by the behaviour of the algorithms in high-level segments, and thus fail to account for the scoring method used in this experiment: the adaptive procedure adopted from the HINT considers full sentence intelligibility, whereby a sentence is scored as incorrect if the listener repeats one word incorrectly.

Finally, both algorithms seem to preserve the binaural cues which lead to positive SRM values in Figure 6.2 (bottom row, center panel). It is notable that SRM values under intermittent masking seem to be slightly larger with the processed signals (especially for NRA01) than the corresponding base condition without processing. It appears that, with speech intelligibility reduced due to the presence of musical noise and the noise-filled gaps, listeners tend to increase their reliance on binaural cues for speech perception.

6.4 Conclusion

The measurements presented in this chapter offer a comprehensive dataset on the interaction of the effects of spatial sound distribution, masker intermittence and nonlinear processing on speech perception. Four independent experiments, testing a total of 32 subjects, were conducted under different binaural scenarios simulated through headphones

and the use of HRTFs. In the absence of hearing device processing, the SRM in response to lateral displacement of the continuous masker ranged between 0 and 10 dB, and increased up to 13 dB for diffuse masking conditions. The choice of signal processing algorithms (clipping, compression, and noise reduction) and masker type (continuous and intermittent) considered has led to a wide range of masking release in our results: (1) the SRM between the S_0N_0 and S_0N_{90} conditions ranged between 2.1 and 7.8 dB; (2) masking release due to the device under test ranged from -11.7 to 7.7 dB; whereas (3) masking release due to intermittence (compared to continuous masking) ranged between 0.5 and 18.7 dB.

The results discussed in this chapter confirm previous findings in the literature and extend knowledge, especially with regards to the interaction of spatial distribution with masker intermittence and the effect of adaptive signal processing algorithms. In particular, the results show that SRM is highly dependent on test conditions related to the type of masker and signal processing. For instance, consistent with the findings in [135], our results showed a decrease in SRT under intermittent masking (compared to continuous masking) in both S_0N_0 and S_0N_{90} conditions. The SRM between the two spatial configurations also decreased. A further decrease in SRT occurred with AGC processing which improved speech intelligibility under intermittent masking, in large part due to a change in the ratio of signal powers in the ON and OFF segments of the masker. The improvement at S_0N_0, however, was greater than at S_0N_{90}, which led to a further decrease in the SRM. Conversely, an increase in the SRT occurred for NRAs which degraded speech intelligibility.The SRM between the two spatial configurations was also increased and was comparable to the continuous masking condition.

Finally, the listening conditions and algorithmic details of the nonlinear processing systems tested are fully described in this chapter. Therefore, in addition to providing a valuable data set for validating the model proposed in this doctoral thesis (*c.f.* Chapter 7), the experiments reported in this chapter contribute new original data accessible to the research community.

Chapter 7

Binaural Validation of the Proposed Model

Introduction

Chapter 4 presented a new model designed to predict speech intelligibility under conditions of binaural listening and nonlinear signal processing. Various modeling components and design choices have been validated in Sections 4.1.4.1 and 4.1.5.2, while other parameters still require some fine tuning using subjective data. Furthermore, a simulation-based study was conducted in Chapter 5 to validate the use of the phase-inversion signal-separation method and distortion estimators introduced in Stage 1 of the model (*c.f.* Section 4.1.1). However, this theoretical validation dealt with a limited set of monaural processing conditions and did not include a comparison with subjective listening tests.

This chapter is dedicated to completing the model validation by comparing SRT predictions to sentence SRT data presented in Chapter 6. The validation framework, briefly introduced in Section 4.4.2, is detailed in Section 7.1. Results are shown in Section 7.2 starting with the model validation and fitting in the absence of hearing device processing then by introducing different signal processing conditions. A discussion of the

results, along with suggestions for future refinements to the model, is presented in Section 7.3, followed by concluding remarks in Section 7.4.

7.1 Overview

The validation process is illustrated in Figure 7.1, which consists of two parallel paths. Similar signal processing is applied in each path. First, monaural signals (scaled to obtain a desired SNR) are processed through a set of HRTFs in order to simulate a given binaural listening scenario. In the subjective path, the binaural speech and noise signals are mixed first, then processed through the hearing device simulator, before being presented to the test subject using an adaptive procedure that is known to converge to the SRT that corresponds to 50% sentence intelligibility. The experimental procedure and the results obtained under a variety of listening and processing conditions have been reported in Chapter 6. In the prediction path, the binaural signals are mixed using the 2x2 Hilbert mixer in four different ways before they are passed through the hearing device simulator. The simulator is integrated in the binaural intelligibility model (*c.f.* Figure 4.2), which outputs a prediction index (the BSII) for a given listening scenario. This index is mapped to a predicted SRT by iteratively adjusting the SNR until the BSII reaches a reference value which corresponds to 50% intelligibility[1].

The model validation is carried out by comparing the predicted and mean measured SRT computed across all subjects for a range of experimental conditions. In the figures presented in this chapter, the error bars associated with the mean measured SRT correspond to the 95% confidence interval across subjects for a given experiment. These values are tabulated along with the rest of the experimental data in Appendix B. The model's predictive performance is assessed by computing the Pearson's correlation coefficient (ρ) and the root-mean-square error (RMSE) (ϵ) between the predicted and mean measured SRT for a given set of conditions.

[1]In this work, the mean measured SRT was used as a starting point to obtain the first prediction index in this iterative search procedure.

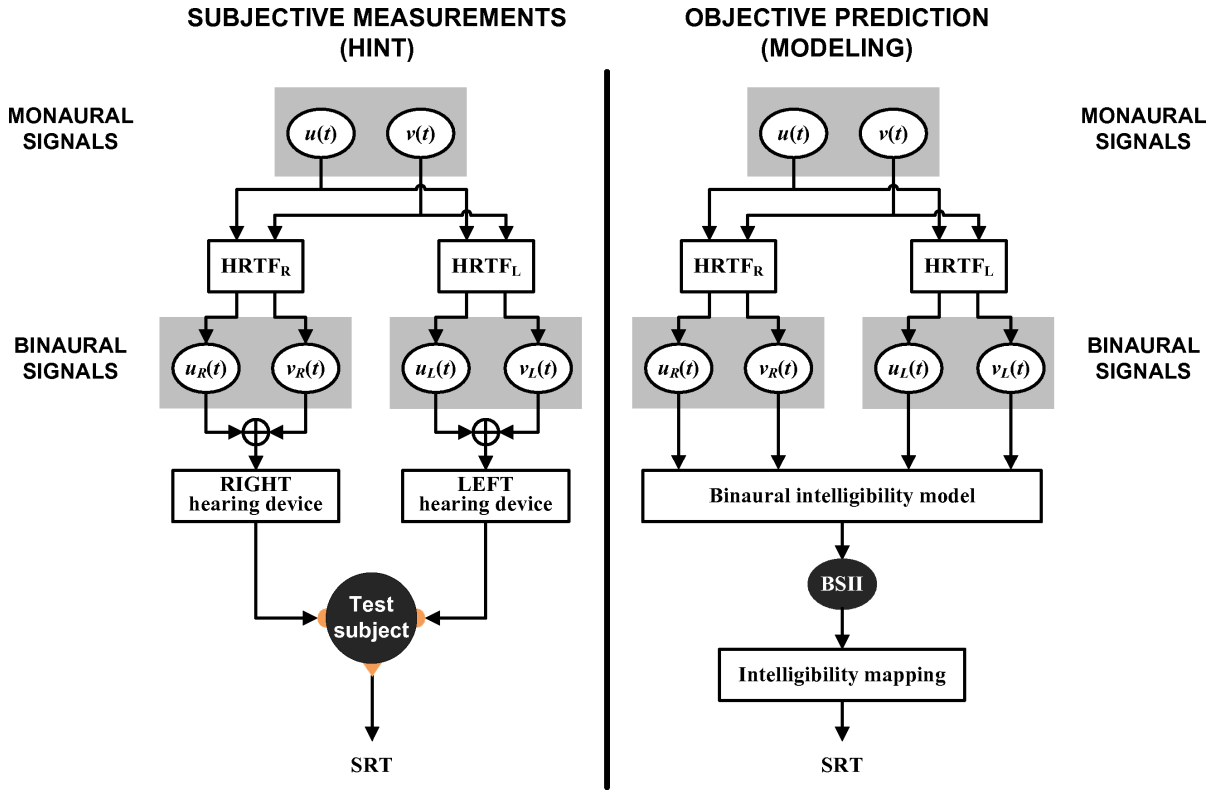


Figure 7.1: *Binaural validation strategy: comparing subjective measurements and objective predictions of speech intelligibility. The hearing device is incorporated within the binaural intelligibility model in the objective prediction pathway.*

This validation process requires some parameter fitting in order to match the model predictions with subjective data. In Section 7.2.1.1, the target BSII reference value which corresponds to 50% intelligibility is explored. The reference value for the standard SII using the HINT speech material (0.375, *c.f.* Section 3.1.2.3) may not correspond to our experimental data, which differ slightly from the HINT norms (*c.f.* Section 6.3.1.1). A new reference value is determined empirically using the long-term implementation of the model (1FRAME) under continuous masking in the absence of signal processing. Another modeling parameter considered is the windowing method used in the short-time implementation of the model. In Section 7.2.1.2, we investigate the two windowing procedures proposed in Section 4.1.4.2: the 24-ms Hanning windows with 50% overlap (HANN24), and the non-overlapping 12-ms rectangular windows (RECT12).

Finally, distortion modeling under nonlinear signal processing conditions is addressed in Section 7.2.2 where three distortion modeling approaches are considered. The choice is set through a selection parameter in the "distortion control panel" block in Stage 1 of the binaural intelligibility model depicted in Figure 4.2:

- DIST0: No distortion modeling is included. Model predictions are based on the speech and noise signals recovered at each ear through the phase-inversion signal-separation method (*c.f.* equations (4.5) and (4.6)).
- DIST1: The speech and noise distortion estimates in equations (4.13) and (4.14) are removed from the recovered speech and noise signals respectively.
- DIST2: Same as DIST1 with the speech distortion estimate also added to the recovered noise signal, thus acting as an additional noise source.

The approach adopted here is an initial attempt to incorporate distortion modeling in the speech intelligibility model. More complicated schemes to automate the modeling decision will be considered in future revisions of the model. These, along with additional model fitting parameters, are discussed in Section 7.3.2.

7.2 Results

7.2.1 Validation and fitting without distortion

7.2.1.1 Target SII reference

The left panel of Figure 7.2 presents a comparison of model predictions using a range of target BSII values between 0.370 and 0.410 to the SRT measured at different masker azimuths in Experiment A. The diffuse-masking conditions are not included in the fitting process as they are not commonly studied in subjective studies (*c.f.* Section 6.3.1). A reference BSII value of 0.400 appears to provide the optimal fit within the 95% CI on the measurements at most azimuths while minimizing the RMSE.

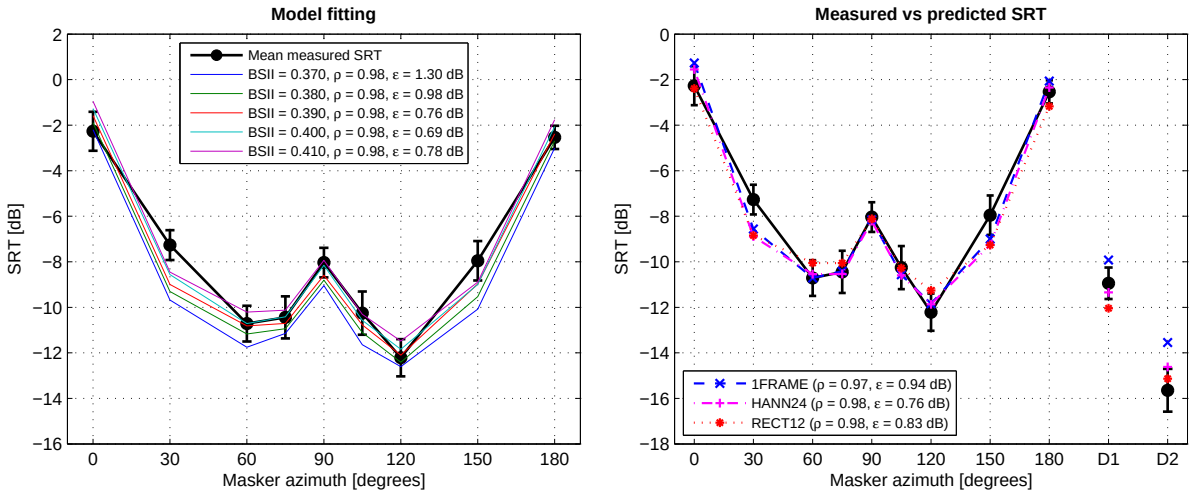


Figure 7.2: *Measured and predicted SRTs for different spatial configurations in the absence of a hearing device (Experiment A). Model fitting with 1FRAME implementation to determine optimal BSII value (left panel). Predictions with all 3 model implementations using the optimal BSII value (right panel).*

The right panel of Figure 7.2 compares predictions obtained using the three model implementations (1FRAME, HANN24, and RECT12) to the subjective SRTs measured in Experiment A. As would be expected under continuous masking, the three implementations yield similar predictions under most spatial configurations. Most predictions fall within the 95% CI of the measured data, except in the S_0N_{30} and S_0N_{150} configurations where the model seems to slightly overestimate intelligibility. Moreover, the results show that the model can account for masking release when 2-D and 3-D diffuse maskers are considered. However, it appears that the long-term implementation (1FRAME) begins to underestimate intelligibility as additional masker sources are mixed. It is possible that mixing uncorrelated sources introduces small short-time spatial and spectral fluctuations in the masker which are not captured in the single-frame approach.

7.2.1.2 Short-time window selection

Figure 7.3 compares model predictions to subjective SRTs measured under continuous and intermittent masking in the absence of signal processing. The mean measured SRT and 95% CI shown in this plot are computed over 16 subjects tested under these con-

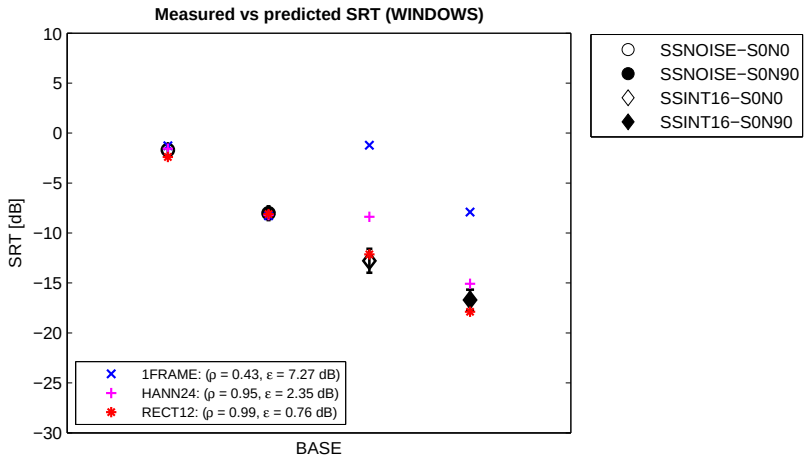


Figure 7.3: *Measured and predicted SRTs under continuous and intermittent masking in the absence of a hearing device. Predictions obtained with all 3 model implementations are shown as indicated in the inner legend. The outer legend indicates the measured data points.*

ditions (BASE conditions of Experiments C and D). As expected, the long-term model implementation largely underestimates speech intelligibility under intermittent masking; since it does not account for short-time amplitude changes in the masker, it fails to model the intelligibility benefits of dip listening (*c.f.* Section 6.3.2). Of the two windowing procedures, the RECT12 model better captures the effects of intermittent masking.

7.2.2 Distortion modeling

In this section, we test the model against the data collected with human subjects under the nonlinear processing conditions in Experiments B, C and D in Chapter 6. For the static algorithms (peak and center clipping), only continuous masking conditions were considered; in this case, predictions obtained with the 1FRAME and RECT12 model implementations are compared to subjective SRT data. For algorithms with dynamic characteristics that depend on the input signal envelope (compression and noise reduction) both continuous and intermittent masking conditions have been tested. For these algorithms, only the RECT12 model is implemented since the 1FRAME model cannot account for intermittent masking release even under baseline conditions (Figure 7.3).

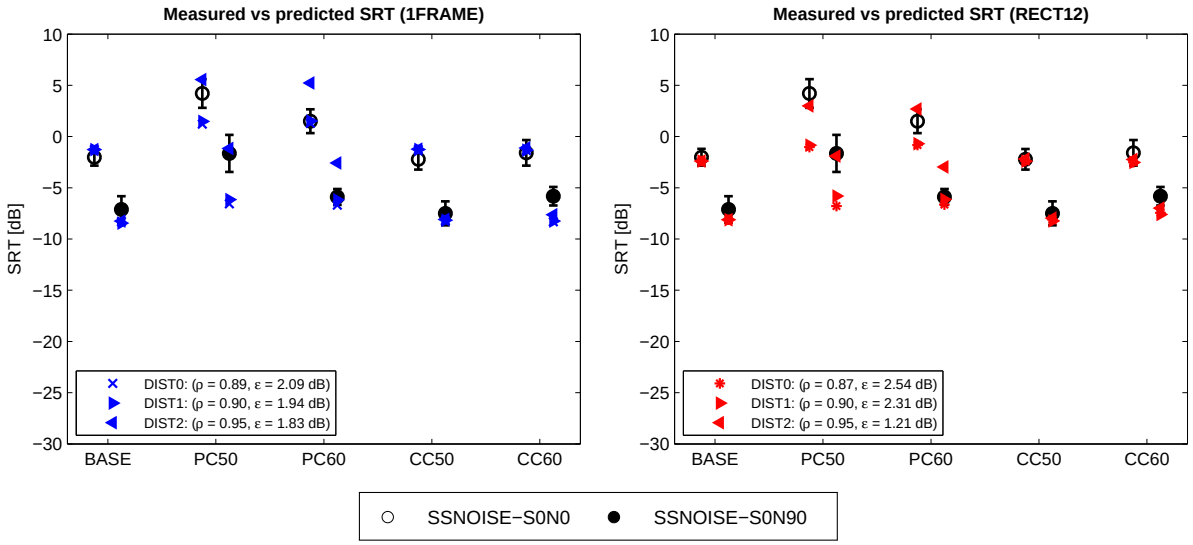


Figure 7.4: *Measured and predicted SRTs under peak and center clipping conditions using the 1FRAME (left panel) and RECT12 (right panel) model implementations. On the horizontal axis, test conditions are grouped by hearing device as defined in Table 6.1. The inner legends show the model performance with different distortion modeling approaches. The outer legend indicates the measured data points.*

7.2.2.1 Clipping algorithms

Figure 7.4 provides a comparison of model predictions with subjective SRTs under peak and center clipping conditions (Experiment B). Overall, predictions obtained with DIST2 (left-pointing triangle markers) best fit the subjective data, with the RECT12 implementation providing the best performance. However, of the two implementations, the 1FRAME predictions seem to offer a better fit with DIST1 (right-pointing triangle markers). Close inspection of the plots reveals additional details. On the one hand, center clipping seems to cause minimal distortion, as evidenced by the clustering of all 3 model predictions around a single point for each condition. This is supported by the experimental data discussed in Section 6.3.3.1. Peak clipping, on the other hand, causes significant distortion which increases as the clipping threshold is decreased. The figure shows that of the modeling approaches considered, DIST2 better captures the distortion when it severely degrades speech intelligibility (PC50), whereas DIST1 provides a better fit when speech intelligibility is preserved (PC60).

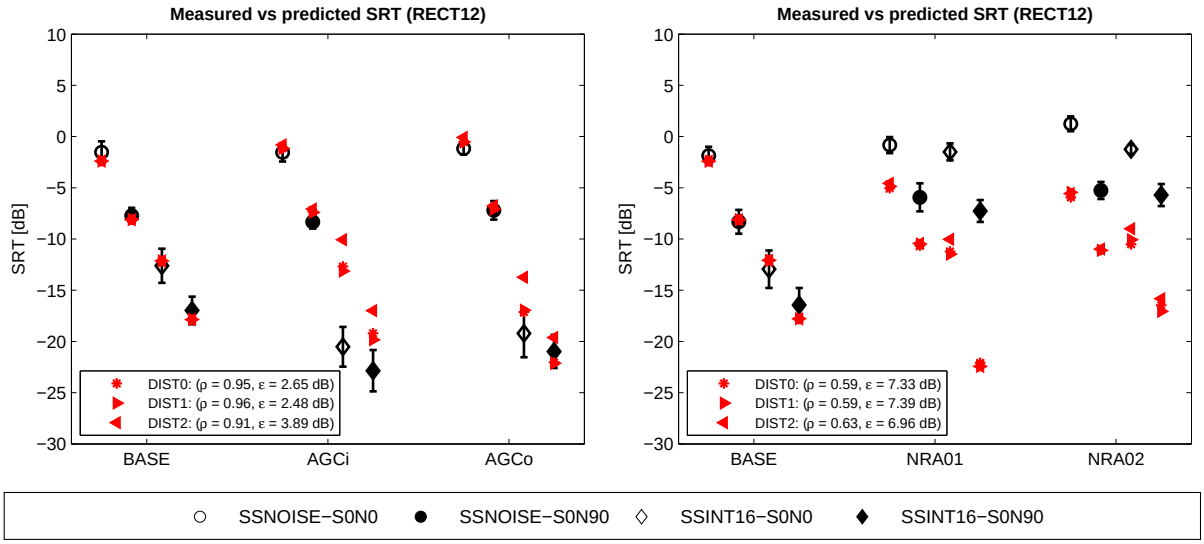


Figure 7.5: *Measured and predicted SRTs under compression (left panel) and NRA processing (right panel) conditions. Predictions using the RECT12 model implementation are shown. Test conditions are grouped by hearing device as defined in Table 6.1.*

7.2.2.2 Compression algorithms

The left panel of Figure 7.5 provides a comparison of model predictions with subjective SRTs when using compression algorithms (Experiment C). Overall, predictions using DIST1 best fit the subjective data. Specifically, under continuous masking conditions in which the effects of the algorithms on speech intelligibility are minimal (*c.f.* Section 6.3.3.2), all three modeling approaches produce similar predictions which fit the measured data points. However, under intermittent masking, where the algorithms improve speech intelligibility, two observations can be made: (1) adding the speech distortion component as an additional noise source (DIST2) degrades the predictive performance of the model; (2) DIST1 and DIST0 perform better with the AGCo algorithm than with AGCi, for which the model predictions appear to underestimate speech intelligibility (higher SRT).

7.2.2.3 Noise reduction algorithms

The right panel of Figure 7.5 provides a comparison of model predictions with subjective SRTs when using NRAs (Experiment D). Overall, there is little effect of distortion

modeling with this data set, however, DIST2 slightly improves on the predictions compared to DIST1 and DIST0. Nevertheless, the model predictions still largely overestimate speech intelligibility (lower SRTs) when the NRAs are used, especially under intermittent masking conditions. In fact, the model predicts improvements in speech intelligibility compared to the baseline conditions, whereas the subjective data shows loss of intelligibility. This observation is consistent with the discussion in Section 6.3.3.3.

7.3 Discussion

7.3.1 Summary

7.3.1.1 Modeling spatial effects

This chapter provides new insight on the modeling of speech intelligibility, binaural hearing, and nonlinear distortions. Overall, in the absence of signal processing algorithms, model predictions exhibit a high correlation with the measured SRT with low prediction errors. The performance indicators in Figure 7.2 for 1FRAME ($\rho = 0.98$, $\epsilon = 0.69$ dB) and RECT12 ($\rho = 0.98$, $\epsilon = 0.83$ dB) fall within the range reported for Beutelmann and Brand's original EC/SII ($\rho = 0.94$, $\epsilon = 1.2$ dB), as well as the revised BSIM ($\rho = 0.99$, $\epsilon = 0.5$ dB) and stBSIM ($\rho = 0.98$, $\epsilon = 0.6$ dB) models[2][13].

7.3.1.2 Modeling masker-type effects

With respect to the short-time extension of the model, the RECT12 windowing method was found to outperform the HANN24 method, even though both windows were designed

[2]These figures cited from [13] are reported here for reference only. A comparison of the models based strictly on these figures would not be fair, since they correspond to spatial conditions similar to those reported in Figure 7.2 as well as other room reverberation conditions not included in our experiments. A more direct comparison is possible with the figures reported in [12], where the authors computed a correlation coefficient $\rho = 0.97$ under similar spatial configurations (not including diffuse masking) in anechoic room conditions, and a mean absolute error (MAE) of 1.6 dB. Since the RMSE gives a relatively high weight to large errors compared to the MAE, the performance indicators reported in Figure7.2 represent an improvement over the numbers reported in [12].

to achieve the same effective length. The model predictions using the HANN24 approach appear to underestimate the benefits of dip listening. This may be attributed to the tapered edges of the Hanning window blurring out the sharp edges of the gaps in the intermittent masker. This led to exaggerated forward and backward masking effects which result in a large underestimation of speech intelligibility. Therefore, while the Hanning window produced good predictions using the continuous speech-shaped masker in this work, and a multi-talker babble or single-talker modulated masker in [13], it may require tuning for listening environments that include sharply fluctuating maskers, such as the sound of a helicopter, pneumatic riveter or machine gun for example.

Furthermore, no additional fitting was needed to account for the effects of masker type (*c.f.* Section 3.1.2.3). The results show that the reference BSII value used to fit of the 1FRAME implementation with the continuous masker data also corresponds to the SRT using the RECT12 implementation under intermittent masking. This further confirms the model's good predictive power [172].

7.3.1.3 Modeling distortion effects

Under distortion conditions, it is reasonably expected that the performance indicators would reflect a decrease in the model's prediction power. It is also worth noting that the sentence SRTs reported in Chapter 6 often exhibited higher variability under distortion conditions than in the absence of signal processing. Overall, predictions when the speech and noise distortion components are subtracted from the respective speech and noise recovered estimates (DIST1) showed a slight improvement compared to when no distortion modeling was included (DIST0). The difference, while minimal, is consistent across all signal processing algorithms considered. Moreover, under conditions where the algorithm severely degraded speech intelligibility (*e.g.*, PC50 and NRAs), incorporating the speech distortion as an additional noise source (DIST2) helped improve the model predictions. Conversely, when speech intelligibility is slightly affected (*e.g.*, PC60, CC50, CC60, and most AGC conditions), the modeling approach DIST1 produced accurate predictions.

Conditions involving AGCi processing under intermittent masking caused the model to underestimate speech intelligibility, with the DIST1 modeling approach providing improved predictions over DIST0 and DIST2. This did not occur for the AGCo algorithm, for which the model successfully predicted the SRT. The difference between the model predictions for these two systems may be attributed to the nonlinear behaviour of the two compressors during the ON and OFF segments of the masker[3]. The underestimated SRTs indicate that additional distortion modeling is required to account for the AGCi amplifier's highly nonlinear behaviour under intermittent masking.

The model's performance is poor under conditions that involve NRAs, where the predicted results largely overestimated the subjectively measured SRTs. Despite this poor performance, the results presented under these conditions give valuable insight into the phase-inversion signal-separation method, which is used in Stage 1 of the model. The method has been employed by several groups in assessing hearing aid performance due to its inherent mathematical simplicity [34, 75, 152, 155, 182, 186, 187, 212]. To the best of our knowledge, the improvements in SNR reported in these studies were not subjectively validated. The present work shows that, even in common conditions which involve continuous masking, these SNR improvements are largely overestimated. The distortion modeling approach DIST2 described in this work represents the first attempt to correct for this overestimation. It helps improve the model predictions, but falls short of matching the subjectively measured SRTs.

Finally, the discrepancy between the subjectively reduced intelligibility due to the NRAs and the improvements predicted in Sections 6.3.3.3 and 7.2.2.3 could, in part, be attributed to the experimental design and choice of sentence intelligibility assessment. In addition to reducing background noise, the NRA may also suppress low-energy

[3]Recall the compressors' parameters listed in Table 6.1 and the intermittence analysis presented in Section 6.3.3.2. At around -20 dB SNR, the signal power in ON segments is roughly the same as the zero-gain reference level of both compressors (68 dB SPL). Thus, on average no gain is applied to the input signal during these segments. During OFF segments, the average speech level is around 45 dBA, which falls above the AGCi's compression threshold (40 dB SPL), and below the AGCo's threshold (60 dB SPL). Thus, the AGCi amplifier will be operating in the compression region, while the AGCo will be operating mostly in linear mode during these segments.

consonantal phonemes which exhibit spectro-temporal properties similar to the speech-spectrum noise used in these experiments (*e.g.*, /f/, /s/, and /sh/). Therefore, although the algorithm may preserve vowels and high-energy voiced consonants (*e.g.*, /n/, /m/, /l/ and /r/), the differential suppression of such consonantal phonemes would drastically degrade performance in a subjective intelligibility task like the one carried out in our measurements, where the SRT was based on whole-sentence keyword repetition. Other scoring methods by phoneme transcription might have been more appropriate. From a diagnostic perspective, it may also be useful to assess the effects of NRAs with phoneme-recognition tasks using, for example, CVC and VCV utterances, as well as using noise signals not based on the speech spectrum.

7.3.2 Model refinements

The model presented in this work consists of three stages as depicted in Figure 4.1 and explained throughout Section 4.1. Although various model parameters can be adjusted, only the reference BSII value was fitted to the subjective data in order to (1) avoid overfitting, (2) adhere, as much as possible, to previous designs in the literature, and (3) provide a model that can be easily adapted to new speech material. In this section we list possible refinements to each of the three stages in the model that may help produce better predictions. Given that model predictions are highly accurate in all spatial configurations and noise types in the absence of hearing device processing, the primary focus is on distortion modeling improvements in Stage 1 of the model.

7.3.2.1 Stage 1: Distortion modeling

The methodology derived in Section 4.1.1 and integrated into the proposed three-stage model is a novel approach to distortion modeling. The speech intelligibility predictions reported in this chapter highlight the importance of using such an approach which distinguishes between speech and noise distortion. Whereas the predictions for all processing algorithms did benefit from removing both the speech and noise distortion components

from the recovered speech and noise estimates (DIST1), further improvements are necessary to account for the effects of all signal processing algorithms considered. Both underestimation (*e.g.*, AGCi conditions) and overestimation (*e.g.*, NRA) of sentence intelligibility occurred in the predicted SRTs reported in this chapter. DIST1 and DIST2 respectively increased and decreased intelligibility predictions. Other approaches, not reported in this work, were also considered in an attempt to further improve predictions: for example, adding the noise distortion component as an additional speech source led to better performance under AGCi processing conditions.

The next step in refining the model would be to redesign the "distortion control panel" to include an automated decision modeling that would select which of the speech and noise distortion estimates is perceptually important and how it is integrated into the model. The following avenues for research in this area will be considered for future revisions of the model:

- Investigate the possibility of incorporating a multi-level approach to distortion modeling inspired by the work of Kates and Arehart [124]. They argue that, when a signal is degraded by noise or distortion, low- mid- and high-level segments of the signal contribute unequally to speech intelligibility. Such an approach, which implicitly incorporates the differences between consonantal and vowel phonemes, proved highly effective in accounting for various types of distortion in different studies [29, 124, 127]. Theoretically, such an approach could be implemented in the short-time version of the model by applying different weights to the speech and noise distortion components in each segment depending on the signal level.

- Investigate the coherence relationships between the input signals (speech or noise) and the recovered signals using the phase-inversion method (signals $c_1(t)$ and $d_1(t)$ in equations (4.5) and (4.6)). Such relationships may be a good indication of the "amount of residual noise" in the recovered speech estimate or vice versa, and could be exploited, for example, to derive weighting factors for each distortion

components in order to correct the recovered estimates and extend the simple modeling approaches used here (DIST1 and DIST2).

- Investigate the relationship between the speech and noise distortion estimates $S_{spd}(f)$ and $S_{nzd}(f)$. This relationship may reflect the degree of statistical dependence between the error signals $e_1(t)$ and $e_2(t)$ obtained during the measurements using the phase-inversion method (*c.f.* equations (4.1) and (4.2)). Statistical dependence (or independence) may be a valid indicator of the underlying distortion, and may be used to derive a selection or weighting mechanism to extend the distortion modeling approach used here.

- Investigate whether the speech-to-distortion and noise-to-distortion ratios (SpDR and NzDR respectively) formally derived in equations (5.1) and (5.2) in Chapter 5 could be used to identify perceptually relevant distortion components.

7.3.2.2 Stage 2: Binaural hearing

Except for the implementation of a 21-band Gammatone filterbank that matches the frequency bands of the SII critical-band computational procedure, the EC processing performed in the Stage 2 of the model was maintained identical to the model proposed by Beutelmann and Brand in [12]. In particular, the artificial processing errors are derived according to equations (4.15) and (4.16), with parameters taken from [206]. These parameters were derived to fit subjective BMLD data from [15, 51, 134]. Modifying these parameters may provide a better fit to the data collected in our measurements. A preliminary investigation showed that increasing the parameter $\sigma_{\epsilon 0}$ to 2.5 dB, and $\sigma_{\delta 0}$ to 100 μs had the following effects:

- predictions using the 1FRAME model implementation fit the subjective data for Experiment A with a target BSII value of 0.375 (the same as the reference value determined for HINT using the standard SII),

- model predictions fit within the 95% CI of the mean for all spatial configurations,

- higher correlation ($\rho = 0.99$) with the subjective SRTs was achieved with lower prediction error ($\epsilon = 0.70$ dB).

This seems to be a promising avenue for refining the model. Nonetheless, care must be taken to avoid overfitting. In addition, the effects of such modifications on predictions under intermittent masking and/or using the short-time implementation of the model should be taken into consideration.

7.3.2.3 Stage 3: Intelligibility prediction

In the final stage of the model, intelligibility prediction is performed using the standard band-importance function listed in Table 1 of ANSI S3.5–1997 (R2012) [4]. Other functions have been derived for specific test materials (*c.f.* Table B1 in [4]). A revision of the proposed model could include an improved estimation of a band-importance function to fit the HINT material using a simple approach recently proposed by Kates [123].

Finally, it is worth noting that, even with a perfect signal estimation and/or distortion modeling approach, using the SII in the last stage of the model to predict sentence intelligibility may prove to be inadequate to model the effects of certain nonlinearities, such as NRAs (*c.f.* Section 7.3.1.3). Should a differential diagnostic approach be adopted to assess such algorithms using phoneme recognition tasks that distinguish between consonants and vowels, the appropriate band-importance function in Table B1 in [4] would be required to model those effects.

7.4 Conclusion

This chapter provided a comprehensive validation of the binaural speech intelligibility model proposed in this thesis. In the absence of hearing device processing, the model successfully predicted speech intelligibility in all spatial configurations using the 1FRAME implementation under continuous masking and the RECT12 implementation under both

continuous and intermittent masking conditions. Two simple approaches were considered to model the distortion effects of different signal processing algorithms: DIST1 was enough to successfully predict the sentence SRTs under some processing conditions (PC60, CC50, CC60, and most AGC conditions), DIST2 gave accurate predictions with the PC50 algorithm, while under- and over-predictions occurred for certain AGCi and NRA conditions respectively.

Despite the latter shortcomings, the research presented in this chapter provides several important contributions. First, the results provide valuable insight into a popular, but previously unverified, research methodology (the phase-inversion signal-separation method). Furthermore, future revisions of the model will extend the distortion modeling approaches presented here by using an automated decision mechanism which selects and/or weighs the contribution of different distortion components to speech intelligibility. Finally, the consideration of additional material and/or scoring methods focusing on speech perception at the phoneme level would provide valuable insights for diagnostic purposes.

Chapter 8

Conclusion

Introduction

This doctoral thesis described the development of a new model for binaural speech intelligibility prediction for nonlinear hearing devices. This chapter summarizes the contributions to the fields of binaural hearing, speech intelligibility, distortion modeling, and assessment of hearing device processing in Section 8.1, and outlines limitations along with future research efforts to address them in Section 8.2. This is followed by concluding remarks in Section 8.3.

8.1 Summary of contributions

The work presented in this thesis addresses an important gap in the literature: the task of predicting speech intelligibility in a binaural listening scenario while accounting for nonlinear processing conditions commonly encountered when using hearing devices, such as hearing aids, electronic hearing protectors, and communication headsets. The main challenge in this work has been to derive a model to isolate the effects of nonlinear distortions produced by hearing devices from the effects of binaural cues in the signals received at each ear. The reference-based model, presented in detail in Chapter 4, can

be summarized in the following three stages:

- Stage 1 is a monaural speech and noise signal estimation and recovery step. An extension of the phase-inversion signal-separation method [87] is performed using novel speech and noise distortion estimators derived by extending the work in [154]. The resulting speech and noise estimates recovered at the output of this stage incorporate the effects of nonlinear distortions due to hearing device processing at the left and right ears.

- Stage 2 implements the concepts of EC theory in order to model the benefits of binaural listening. It includes a Gammatone analysis filterbank and internal masking signals to account for hearing loss at each ear. The stage is an extension of the work of Beutelmann and Brand [12], with several optimizations to the EC process designed to improve the model's computational performance.

- Stage 3 predicts speech intelligibility using the SII (ANSI S3.5–1997 (R2012) [4]), or the short-time extended version, the ESII [170].

The originality of this work lies in the three-stage architecture adopted as a solution to the modeling challenge stated above. Several reference-based monaural measures have been discussed that can predict speech intelligibility while accounting for various forms of nonlinear distortion (*c.f.* Section 3.2.2). These measures typically rely on auditory models or some type of processing that transform the processed and reference signals into internal representations in the TFS or envelope (modulation) domain. Instead of using a coherence or correlation-based metric between the processed and reference signals, the first stage of the model looks at the relationship between a set of signals related through the Hilbert transform to produce time-domain estimates that can be further analyzed to extract binaural cues. The reference signals are essentially used to produce the input mixtures required for multiple measurements on the hearing device. In this way, the combination of the phase-inversion method with a Hilbert-based distortion metric presents a promising avenue for distortion modeling.

In the absence of detailed acoustic and perceptual data to validate the objective model for common signal processing algorithms, subjective data were collected under a variety of spatial listening situations. This was carried out in four experiments in Chapter 6 investigating 11 spatial configurations, two types of maskers, and different algorithms, which included peak and center clipping, two compression systems (AGCi and AGCo), and two NRAs. Sentence SRTs were collected using the Matlab speech testing environment and the HINT speech material, under listening and processing conditions not possible with the commercial HINT platform. These measurements confirm previous findings in the literature and extend knowledge, especially with regards to the interaction of spatial distribution with masker intermittence and the effects of adaptive signal processing algorithms.

Model validation carried out in Chapter 7 focused on evaluating the long-term and short-term model implementations, as well as comparing different distortion modeling approaches. In the absence of hearing device processing, the model successfully predicted speech intelligibility in all spatial configurations using the 1FRAME long-term implementation under continuous masking, and using the RECT12 short-term implementation under both continuous and intermittent masking conditions. Performance indicators used to evaluate the model under these conditions compared favorably with those in [12] and [13]. Two simple distortion modeling approaches were investigated with varying levels of success with different distortion mechanisms.

8.2 Thesis limitations and future works

8.2.1 Distortion modeling

The model predictions obtained with the distortion modeling approaches considered were mixed: DIST1 gave accurate predictions with PC60, CC50, CC60, and most AGC conditions, DIST2 gave accurate predictions for PC50 and slightly improved the results

for NRAs, while under- and over-predictions occurred for certain AGCi and NRA conditions respectively. The results suggest that the popular use of the phase-inversion signal-separation method, previously unverified with subjective data, must be cautioned, particularly under highly non-stationary masking conditions for AGCi, and even under the most common masking conditions with NRAs.

Several avenues have been identified in Section 7.3.2.1 for future revisions of the model or testing methodology. One option is to redesign the "distortion control panel" using an automated decision mechanism which selects and/or weighs the contribution of different distortion components to speech intelligibility. To this end, helpful extensions may be found in a multi-level distortion modeling approach (*e.g.*, [124, 127]), or in the relationships between different pairs of signals. Incorporating additional modeling elements from correlation-based metrics in the envelope domain (*e.g.*, [29, 114, 115, 196, 197]) is another research avenue worth exploring. Finally, another option consists in revising the testing methodology to capture the differential effects of NRAs on different components of speech. This may require additional test material and/or scoring methods that focus on speech perception at a finer scale than whole sentences, owing to the very different characteristics of the different phonemes of speech.

8.2.2 Hearing-loss modeling

The model has been validated with subjective intelligibility scores from NH subjects. HI subjects are normally considered at a later stage in the development once the intelligibility model has been refined. At this stage, the proposed model accounts for hearing loss by incorporating the subject's hearing thresholds as an internal masking noise (*c.f.* Section 4.1.2.2). This approach has been validated in a variety of binaural listening scenarios under stationary masking in [12]. For the short-time ESII model, the inclusion of an internal masking noise has been shown to successfully predict subjective SRTs for HI subjects under stationary masking but not under intermittent masking [173].

Two approaches for hearing-loss modeling will be considered in future revisions of

the model. The first consists of shifting the psychometric function that relates the intelligibility index to SRT scores. This approach has been used to adjust SII predictions for HI subjects in real-life noises [76] and when using hearing protection devices [77]. The second approach consists of modeling the widening of auditory filters due to cochlear hearing impairment. This has been successfully incorporated in a recent monaural speech intelligibility model [127]. To that end, Hohmann's implementation of the gammatone filterbank [91] would allow this solution to be easily integrated into the second stage of the model (*c.f.* Section 4.1.2.1).

8.2.3 Other modeling considerations

In this work, short-time processing has been performed using fixed-length windows. More advanced windowing mechanisms, such as frequency-dependent window-lengths or a forward masking function as proposed in the monaural ESII [170, 171], have not been investigated, and could be addressed in future revisions. However, the effect of such processing on the binaural processing mechanisms are not yet clear. For example, the ability of the model to account for large ITDs in the S_0N_{90} condition, may be negatively impacted by shorter windows in some frequency bands.

The vast majority of spatial configurations considered in this work lie in the horizontal plane. Testing in the 3-dimensional plane was limited to a single condition: Experiment A, condition C11 (*c.f.* Section 6.1.2.1). The response of the pinna to the elevation of a sound source results in high-frequency cues that aid in sound localization. In theory, these cues are properly captured in the EC model given the use of a 21-channel analysis filterbank to perform subband processing. The results reported in Section 7.2.1.1 with the short-time model implementation show that this is indeed the case for the 3D diffuse masker. In the future, additional testing using single speech-noise point sources in the 3-dimensional space would be beneficial to explore this further.

Finally, the Matlab hearing device simulator was used to process signals in both the modeling and experimental parts of this thesis. The next step would be to perform a

study using recordings from real hearing devices. When performing such recordings, care must be taken to (1) ensure signal alignment between different measurements required for the model and (2) avoid time-varying memory effects that may affect those measurements (*c.f.* Section 4.1.5.1). Furthermore, testing was performed using stationary and intermittent speech-shaped maskers. These represent the best and worst-case energetic masking scenarios that test the limits of the model's application. Future studies could complement this work using noise recordings from real-life environments as in [172].

8.3 Outlook

In conclusion, the research presented in this thesis is an important step towards characterizing speech intelligibility under binaural listening situations involving nonlinear processing conditions. On the one hand, the experimental measurements presented expand our understanding in the field of speech intelligibility. On the other hand, the modeling results show a promising potential and serve as a basis for further refinements to the model proposed in this work. Several publications have arisen from this work, and a number of publications is being considered in the near future. Finally, the present research has important applications both among clinicians and engineers in the fields of hearing aid fitting (*e.g.*, selection of parameters and algorithms to optimize speech intelligibility) and algorithm design (*e.g.*, development of binaural signal processing algorithms based on speech intelligibility).

Appendix A

Analysis-Synthesis Filterbanks

Introduction

An analysis filterbank is used in stage 2 of the model to split the input signals into a number of frequency bands to model the peripheral auditory system (*c.f.* Section 4.1.2.1). In this appendix we discuss the filterbank implementation used in the model of Beutelmann and Brand [12], and we document the modifications employed in our model.

A.1 Gammatone analysis filterbank

Gammatone filters are frequently used to represent auditory filters. Their impulse response is the product of a gamma distribution with a sinusoidal tone expressed as:

$$g(t) = at^{n-1}exp^{-2\pi bt}cos(2\pi ft + \phi), \quad (\text{for } t > 0) \tag{A.1}$$

where t and f represent the time and frequency variables respectively, a is the amplitude, n is the filter order, b its bandwidth, and ϕ is the phase. These filters are based on the shape and bandwidth of auditory filters derived from notched-noise measurements [160]. Figure A.1 shows the impulse and magnitude response of a fourth order gammatone filter ($n = 4$) centered at $f = 1000$ Hz with $a = 10^4$, $b = 160$, and $\phi = 0$.

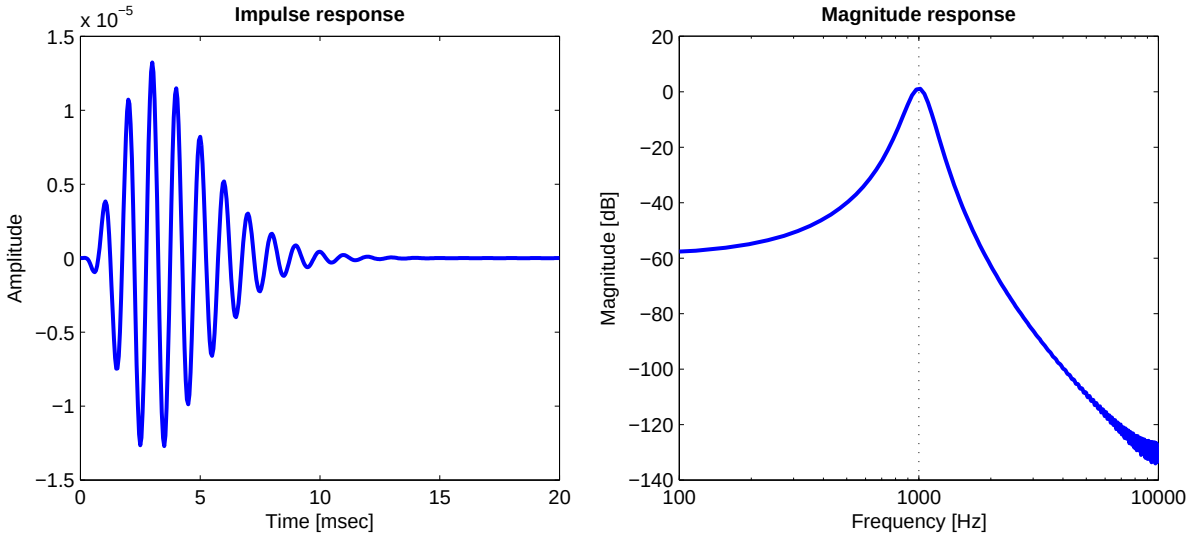


Figure A.1: *Impulse and magnitude response of a fourth gammatone filter centered at 1 kHz. The filter bandwidth is selected according to the critical-band specifications in [4].*

A.1.1 30-band Gammatone analysis filterbank

Thus, to simulate peripheral auditory processing, the input signal is convolved with a bank of filters with impulse response as in equation (A.1). The filterbank employed in Beutelmann and Brand's model [12] uses an all-pole approximation of the gammatone filters proposed by Hohmann [91]. The 30-band filterbank is implemented using fourth-order filters ($n = 4$) with center frequencies between 146 and 8346 Hz equally spaced on the equivalent-rectangular-bandwidth (ERB [81]) scale and bandwidth of 1 ERB each. The magnitude response of the filterbank is plotted in Figure A.2, along with the overall filter response (the sum of all the filters).

A.1.2 21-band Gammatone analysis filterbank

The 30 bands used for the analysis filterbank in [12] do not match any of the computational procedures in the standard SII [4]. Thus, the binaural model requires a synthesis stage (*c.f.* Section A.2) which is immediately followed by a second analysis stage corresponding to the chosen SII procedure. In the revised BSIM model [13], this was addressed

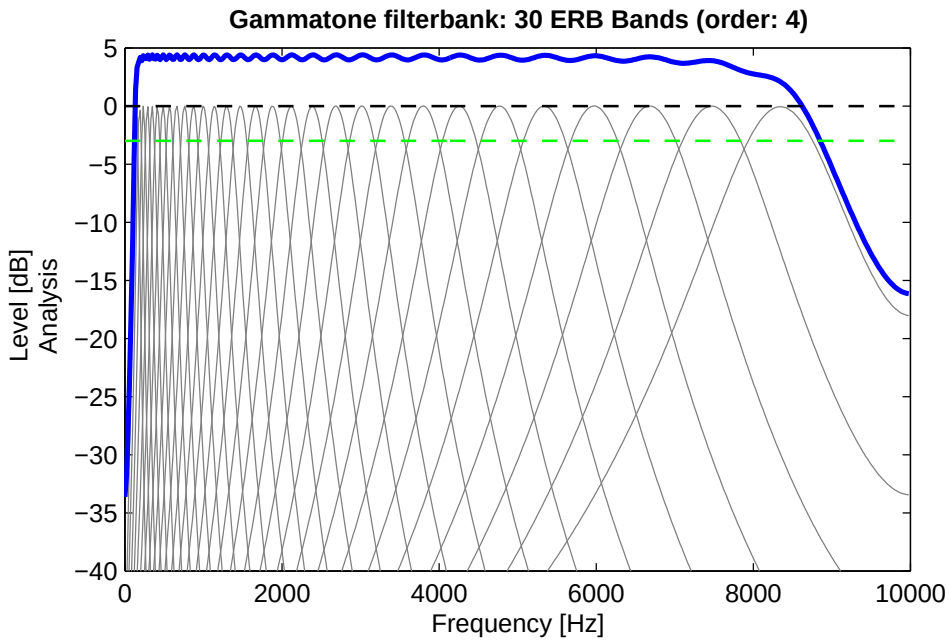


Figure A.2: *Magnitude response of the gammatone filterbank used in [12]. The 0 dB and -3 dB points are indicated with black (dark) and green (light) dashed lines respectively.*

by adapting the standard SII computation to the analysis filterbank. This implies a deviation from the ANSI standard, and the resulting model is fitted to intelligibility scores obtained with the Oldenburg Sentence Test. Our solution was to design a filterbank according to the center frequencies and bandwidths of the 21-band critical-band procedure as outlined in Table 1 of ANSI S3.5–1997 (R2012) [4]. Using this decomposition means the SII computation in Stage 3 of the model can be performed directly on the processed signals without adjusting the band-importance functions. The magnitude response of the 21-band critical-band filterbank is shown in Figure A.3. The overall response of the 30-band filterbank is overlaid on the plot to allow a comparison of the two.

Some additional notes on the proposed analysis filterbank design are necessary. In order to maintain as far as possible the approach used in Beutelmann and Brand's original model, Hohmann's implementation is still being used to design each of the 21 filters in the filterbank. 12^{th}-order filters are needed to achieve the center frequencies and bandwidths that match the 21-band critical-band specifications in ANSI S3.5–1997 (R2012), while

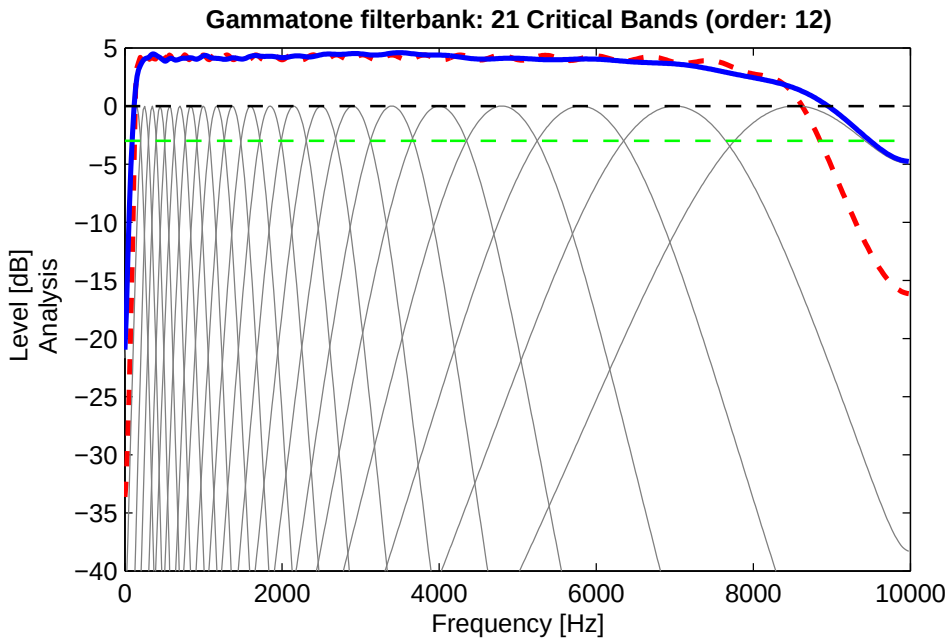


Figure A.3: *Magnitude response of the 21-band gammatone filterbank. The 0 dB and -3 dB points are indicated with black (dark) and green (light) dashed lines respectively. The response of the 30-band filterbank is included for comparison (thick dashed line).*

matching as closely as possible the overall response of the 30-band analysis filterbank. However, the all-pole approximation introduces an error, which is negligible for most filters, but can cause significant distortion in filters with a short impulse response (*c.f.* Section 2.1 in [91]). This affects the last filter of the 21-band filterbank, for which the bandwidth is large. Therefore, this filter was designed using a slightly narrower bandwidth specification (1500 Hz instead of 1800 Hz) in order to avoid these distortion effects. This modification is not expected to affect the model predictions since the band in question has an audibility factor of 0.43%. Still, the overall magnitude response of the 21-band analysis filter shown in Figure A.3 is comparable to the response of the 30-band filterbank over most of the frequency range. In fact, the proposed filter better captures the high-frequency content of the signal.

Table A.1: *Summary of differences between the subband decompositions used for the EC/SII processing combination.*

Specifications	EC30	EC21
Number of bands (EC)	30	21
Center frequencies	ERB-spaced	According to [4]
Bandwidths	1 ERB	According to [4]
Analysis filters	Gammatone, 4^{th} order	Gammatone, 12^{th} order
Synthesis required?	Yes (using[91])	No
Number of bands (SII)	18 (third-octave bands)	21 (critical-bands)

A.1.3 Subband decomposition summary

As we have seen, the filterbank implementation affects, not only the stage of the model that performs EC processing, but also the stage that computes the SII. Table A.1 summarizes the differences discussed so far between the two implementations of the EC/SII processing combination: namely the model of Beutelmann and Brand [12] (EC30), and our own implementation (EC21).

A.2 Gammatone synthesis

A number of subband signals are obtained at the output of the EC process. These signals are combined into a single broadband signal using a synthesis filterbank. Hohmann's filterbank implementation also provides a synthesis operation with a desired analysis-synthesis group delay [91]. The operation consists of delaying each of the filter's impulse responses to have all the envelope maxima aligned at the desired group delay. A weighted sum of these (delayed) responses results in the re-synthesized signal. The weights applied are numerically optimized in order to achieve a flat analysis-synthesis frequency response using a procedure described in [88]. The left panel of Figure A.4 illustrates the analysis-synthesis process using the Matlab implementation provided by the author. The figure shows plots of the power spectral density (PSD) at the input, subband, and output stage for three different wideband signals (white, pink, and speech-shaped noise). These

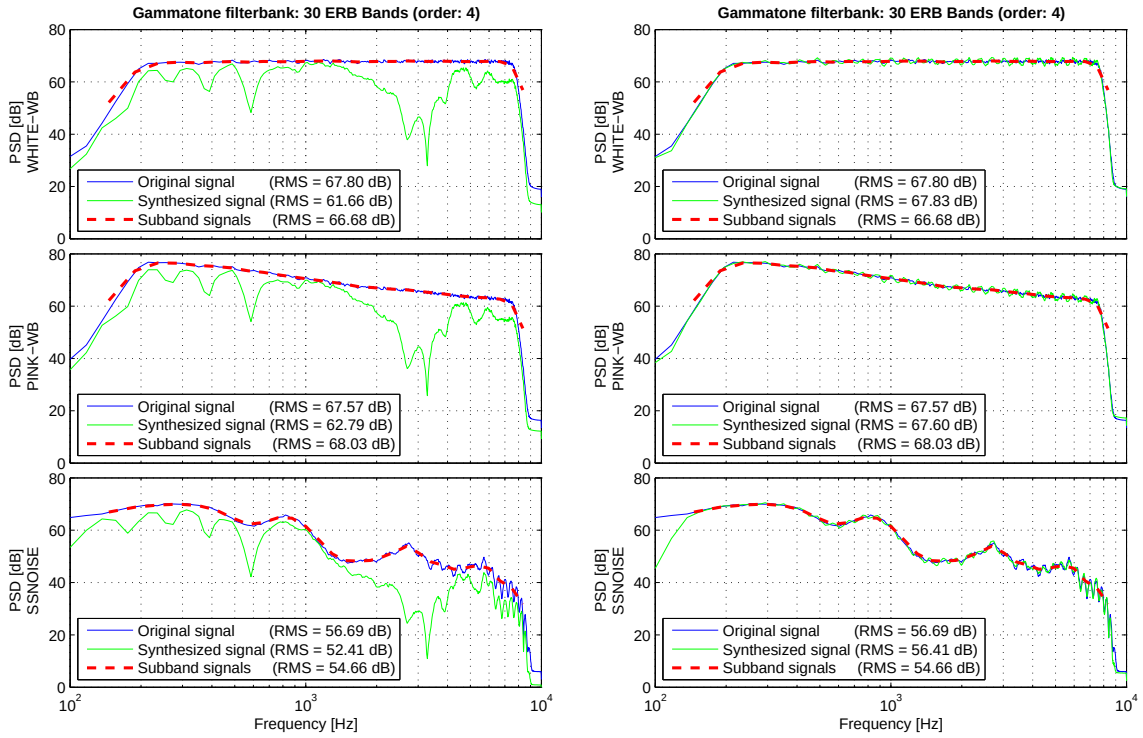


Figure A.4: *PSD of the 30-band ERB-spaced filterbanks' output to white, pink and speech-shaped noise using the synthesis procedure according to [91] (left panel), and using a time-reversal synthesis (right panel).*

plots show that the analysis-synthesis process does not achieve a perfect reconstruction, perhaps suggesting that the optimized weights used in the synthesis process did not succeed in achieving a flat frequency response.

An alternative synthesis method often used in the literature consists of time-reversing the impulse responses of the analysis filterbank to obtain a synthesis filterbank [86, 104, 133]. Thus, synthesis is performed by convolving the subband signals with the time-reversed filters and summing up the outputs. In the z-transform domain, time-reversal corresponds to inverting the system response. Therefore, the z-transform of the impulse response of the m^{th} synthesis filter can be writen as:

$$H_m(z) = \frac{1}{G_m(z)} \tag{A.2}$$

where $G_m(z)$ is the z-transform of the impulse response of the m^{th} analysis filter. This

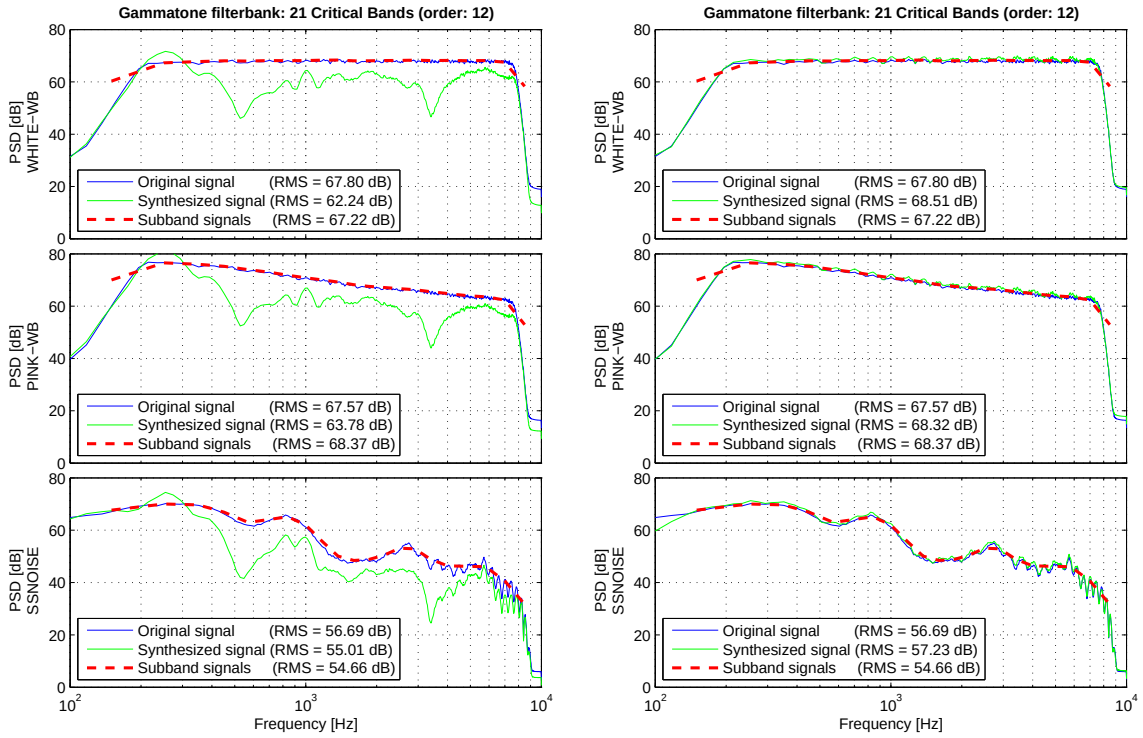


Figure A.5: *PSD of the 21-band critical-band filterbanks' output to white, pink and speech-shaped noise using the synthesis procedure according to [91] (left panel), and using a time-reversal synthesis (right panel).*

approach has the drawback that any zero of the analysis filter which falls outside the unit circle becomes an unstable pole of the synthesis filter. This is not a problem here since Hohmann's implementation is an all-pole approximation of gammatone filters. The right panel in Figure A.4 shows the PSD of the input, subband and output signals of the same 30-band filterbank but using the time-reversed synthesis operation. The analysis-synthesis process is also illustrated using both synthesis approaches for the 21-band filterbank in Figure A.5.

Finally, it is worth noting that, even though the re-synthesized signals are no longer needed in our model, the synthesis operation (*c.f.* Section A.2) was included in the implementation for listening purposes to aid in examining the effects of EC processing.

Appendix B

Experimental Results

In this appendix, we present the experimental results reported in Chapter 6 of the thesis. The tables list the SRT measured for each experimental conditions for each subject, along with the mean measured SRT, standard deviation (SD) and 95% confidence interval (CI) for each experiment. Individual data points with a standard deviation above 3.5 dB are indicated in **bold** font, and have been discarded (*c.f.* Section 6.2).

Table B.1: *Subjective SRTs measured in Experiment A*

	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12
S01	-2.3	-8.4	-10.5	-12.4	-8.6	-10.7	-12.4	-9.2	-3.2	-12.4	-17.3	
S02	-3.2	-6.5	-11.7	-8.1	-6.8	-8.1	-12.1	-5.4	-2.0	-10.4	-15.8	
S03	-1.5	-7.2	-9.9	-11.9	-7.7	-11.7	-12.1	-7.6	-3.2	-10.7	-13.5	
S04	-1.8	-5.8	-10.1	-10.0	-7.6	-8.7	-10.1	-7.4	-1.7	-9.1	-14.0	
S05	-1.6	-6.6	-9.9	-9.6	-7.5	-10.3	-13.7	-7.6	-1.8	-11.1	-15.3	
S06	-2.5	-8.4	-12.8	-11.0	-9.4	-11.6	-13.5	-8.8	-2.7	-11.5	-15.8	
S07	-0.5	-7.4	-11.3	-10.2	**-5.0**	-9.5	-12.6	-8.9	-2.0	-10.7	-16.1	
S08	-4.6	-7.8	-9.5	-10.3	-8.7	-11.4	-11.2	-8.8	-3.6	-11.7	-17.2	
Mean	-2.3	-7.3	-10.7	-10.4	-8.0	-10.3	-12.2	-8.0	-2.5	-10.9	-15.6	
SD	1.2	0.9	1.1	1.3	0.9	1.4	1.2	1.2	0.7	1.0	1.4	
CI	0.9	0.7	0.8	0.9	0.7	0.9	0.8	0.9	0.5	0.7	0.9	

Table B.2: *Subjective SRTs measured in Experiments B, C and D*

	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12
Experiment B												
S01	-0.7	-5.4	4.2	-0.9	3.5	-5.1	-0.8	-8.7	-3.2	-7.0		
S02	-0.9	-8.0	3.3	-1.6	2.7	-5.6	-2.3	-7.4	1.0	-4.8		
S03	-2.5	-8.1	3.0	-1.6	1.3	-6.5	-3.0	-8.0	-3.6	-5.4		
S04	-2.1	-6.7	3.9	-5.9	-0.1	-6.6	0.0	-8.9	-0.9	-6.8		
S05	-1.4	-4.2	9.0	2.4	**5.2**	-4.0	-1.6	-4.2	-0.1	-4.2		
S06	-1.2	-5.9	4.2	1.0	**8.2**	-5.0	-2.1	-5.7	-0.3	-4.5		
S07	-3.7	-8.9	3.4	-3.5	0.0	-6.7	-4.4	-8.5	-3.9	-6.1		
S08	-3.7	-9.6	2.6	-3.0	1.6	-7.4	-3.6	-8.7	-1.6	-7.8		
Mean	-2.0	-7.1	4.2	-1.6	1.5	-5.9	-2.2	-7.5	-1.6	-5.8		
SD	1.2	1.9	2.0	2.6	1.5	1.1	1.4	1.7	1.8	1.3		
CI	0.8	1.3	1.4	1.8	1.2	0.8	1.0	1.2	1.3	0.9		
Experiment C												
S01	1.2	-7.7	-13.9	-14.5	-1.5	-8.2	-20.6	-21.7	-1.6	-6.4	-16.9	-21.2
S02	-3.2	-9.4	-13.8	-16.6	-1.4	-8.7	**-21.7**	-20.9	-1.4	-6.8	-17.8	-19.8
S03	-2.9	-8.7	-11.1	-15.9	-0.2	-10.0	-17.3	-18.9	-1.4	-6.2	-17.1	-18.9
S04	-1.4	-8.2	-15.5	-19.6	-2.8	-7.8	-21.8	-26.8	0.5	-8.8	-22.9	-24.2
S05	-3.3	-6.4	-10.4	-17.3	0.0	-8.6	**-19.1**	-23.1	-0.4	-6.1	**-19.2**	**-24.3**
S06	-1.0	-6.4	-9.7	-14.6	-1.5	-8.1	-18.8	-20.3	-1.2	-5.9	-17.6	-19.7
S07	-0.5	-8.1	-10.8	-17.8	-3.9	-8.6	-20.2	-24.7	-1.7	-8.8	**-17.9**	-19.1
S08	-1.2	-6.8	-15.8	-19.4	-1.1	-6.7	-24.3	-26.5	-2.2	-8.7	-22.9	-23.7
Mean	-1.5	-7.7	-12.6	-17.0	-1.5	-8.3	-20.5	-22.9	-1.2	-7.2	-19.2	-21.0
SD	1.5	1.1	2.4	2.0	1.3	0.9	2.4	2.9	0.9	1.3	2.9	2.2
CI	1.1	0.8	1.7	1.4	0.9	0.6	1.9	2.0	0.6	0.9	2.3	1.6
Experiment D												
S01	-1.4	-7.7	-11.4	-17.2	0.6	-4.5	-0.5	-8.4	1.6	-5.5	-0.5	-5.0
S02	0.7	-8.9	-11.2	-15.2	-0.2	-2.5	-0.3	-7.8	-0.7	-3.6	-0.9	-5.0
S03	-2.2	-9.8	-14.7	-17.5	-2.5	-8.1	0.3	-7.8	1.4	-6.7	-1.0	-4.3
S04	-1.6	-10.3	-16.0	-17.3	-1.7	-6.8	-1.6	-9.1	1.8	-4.8	-2.4	-8.1
S05	-3.4	-8.4	-10.5	-14.9	-0.1	-5.5	-2.8	-6.8	1.6	-5.1	-0.5	-6.2
S06	-3.2	-9.2	-16.0	-17.2	-1.5	-8.7	-2.6	**-8.2**	1.5	-3.8	-1.6	-6.9
S07	-1.8	-5.0	-9.3	-12.0	-1.7	-5.5	-2.1	-6.0	0.1	-6.8	-1.5	-3.4
S08	-2.0	-7.3	-14.5	-20.1	0.4	-5.9	-2.4	-5.0	2.7	-6.0	-1.6	-6.8
Mean	-1.9	-8.3	-12.9	-16.4	-0.8	-5.9	-1.5	-7.3	1.2	-5.3	-1.2	-5.7
SD	1.3	1.7	2.6	2.4	1.1	2.0	1.2	1.4	1.1	1.2	0.6	1.6
CI	0.9	1.2	1.8	1.7	0.8	1.4	0.8	1.1	0.7	0.8	0.4	1.1

Bibliography

[1] ANSI S3.2–2009. American national standard: Methods for measuring the intelligibility of speech over communication systems. *American National Standards Institute*, Melville, NY 2009.

[2] ANSI S3.42–1992 Part 1 (R2012). American national standard: Testing hearing aids with a broad-band noise signal. *American National Standards Institute*, New York, NY 1992.

[3] ANSI S3.5–1969. American national standard: Methods for the calculation of the articulation index. *American National Standards Institute*, New York, NY 1992.

[4] ANSI S3.5–1997 (R2012). American national standard: Methods for the calculation of the speech intelligibility index. *American National Standards Institute*, New York, NY 1997.

[5] Corrections Standards Authority. Hearing standard for selection of entry-level correctional officers. Technical report, State of California, Department of Corrections and Rehabilitation, Sacramento, CA, 2011.

[6] M. Bahoura and J. Rouat. Wavelet speech enhancement based on time-scale adaptation. *Speech Comm.*, 48(2):1620–1637, December 2006.

[7] E. Bàràny. A contribution to the physiology of bone conduction. *Acta Otolaryngol*, Suppl. 26:1–223, 1983.

[8] E. Bedrosian. A product theorem for hilbert transforms. *Proc. of the IEEE*, 51(5): 868–869, May 1963.

[9] J.G. Beerends, A.P. Hekstra, A.W. Rix, and M.P. Hollier. Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment part II-psychoacoustic model. *J. Audio Eng. Soc.*, 50(10):765–778, October 2002.

[10] J. Bench, A. Kowal, and Bamford J. The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *Br. J. Audiol.*, 13(3):108–112, August 1979.

[11] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, volume 4, pages 208–211, Washington, DC, April 1979.

[12] R. Beutelmann and T. Brand. Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.*, 120(1):331–342, July 2006.

[13] R. Beutelmann, T. Brand, and B. Kollmeier. Revision, extension, and evaluation of a binaural speech intelligibility model. *J. Acoust. Soc. Am.*, 127(4):2479–2497, April 2010.

[14] R.C. Bilger, J.M. Nuetzel, W.M. Rabinowitz, and C. Rzeczkowski. Standardization of a test of speech perception in noise. *J. Speech Hear. Res.*, 27(1):32–48, March 1984.

[15] H.C. Blodgett, L.A. Jeffress, and R.H. Whitworth. Effect of noise at one ear on the masked threshold for tone at the other. *J. Acoust. Soc. Am.*, 34(7):979–981, July 1962.

[16] J.B. Boldt and D.P.W. Ellis. A simple correlation-based model of intelligibility for

nonlinear speech enhancement and separation. In *Proceedings of EUSIPCO*, pages 1849–1853, Glasgow, Scotland, August 2009.

[17] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.*, 27(2):113–120, April 1979.

[18] J. Braash. Modelling of binaural hearing. In J. Blauert, editor, *Communication Acoustics*, pages 75–108. Springer, The Netherlands, 2005.

[19] A.J. Brammer, G. Yu, E.R. Bernstein, M.G. Gherniak, J.B. Tufts, and D.R. Peterson. Predicting the intelligibility of speech corrupted by nonlinear distortion. In *Canadian Acoustics*, 38(3), pages 80–81, Victoria, BC, October 2010.

[20] A.J. Brammer, G. Yu, E.R. Bernstein, M.G. Gherniak, J.B. Tufts, and D.R. Peterson. Intelligibility of speech corrupted by nonlinear distortion. In *Proc. Int. Congress on Noise as a Public Health Problems (ICBEN)*, London, UK, July 24–28 2011.

[21] J. Breebaart, S. van de Par, and A. Kohlrausch. Binaural processing model based on contralateral inhibition. I. model structure. *J. Acoust. Soc. Am.*, 110(2):1074–1088, August 2001.

[22] A.W. Bronkhorst and R. Plomp. The effect of head-induced interaural time and level differences on speech intelligibility in noise. *J. Acoust. Soc. Am.*, 83(4):1508–1516, April 1988.

[23] A.W. Bronkhorst and R. Plomp. Binaural speech intelligibility in noise for hearing-impaired listeners. *J. Acoust. Soc. Am.*, 86(4):1374–1383, October 1988.

[24] D.S. Brungart, P.S. Chang, B.D. Simpson, and D.L. Wang. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *J. Acoust. Soc. Am.*, 120(6):4007–4018, December 2006.

[25] M.D. Burkhard. *Manikin measurements. Conference proceedings.* G.R.A.S. Sound and Vibration, Skovlytoften 33, 2840 Holte, Denmark, 1976.

[26] D. Byrne and H. Dillon. The National Acoustic Laboratories' (NAL) new procedure for selecting the gain and frequency response of a hearing aid. *Ear and Hearing*, 4 (7):257–265, August 1986.

[27] O. Cappe. Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. *IEEE Trans. Audio, Speech, and Language Processing*, 2 (2):345–349, April 1994.

[28] G.C. Carter, C.H. Knapp, and A.H. Nuttall. Estimation of the magnitude-squared coherence function via overlapped fast fourier transform processing. *IEEE Trans. Audio Electro-acoustics*, 21(4):337–344, August 1973.

[29] C. Christiansen, M.S. Pedersen, and T. Dau. Prediction of speech intelligibility based on an auditory preprocessing model. *Speech Communications*, 52(7-8):678–692, July-August 2010.

[30] I Cohen. Speech enhancement using a noncausal a priori SNR estimator. *IEEE Signal Process. Lett.*, 11(9):725–728, September 2004.

[31] H.S. Colburn. Theory of binaural interaction based on auditory nerve data. II. detection of tones in noise. *J. Acoust. Soc. Am.*, 61(2):525–533, February 1977.

[32] H.S. Colburn. Theory of binaural interaction based on auditory nerve data. II. detection of tones in noise. supplementary material, 1977. AIP document no. PAPS JASMA-91-525-98.

[33] H.S. Colburn and N.I. Durlach. Models of binaural interaction. In E.C. Carterette and M.P. Friedman, editors, *Handbook of Perception, Vol IV, Hearing*, pages 467–518. Academic Press, New York, NY, USA, 1978.

[34] L.E. Cornelisse. Technique to characterize the differential treatment of speech and noise adaptive signal processing devices: using signal inversion to extract estimates of the modified speech and noise components. In *Proc. Int. Hear. Aid Res. Conf.*, Lake Tahoe, CA, August 2008.

[35] T.R. Crain and D.J. Van Tasell. Effect of peak clipping on speech recognition threshold. *Ear and Hearing*, 15(6):443–453, December 1994.

[36] T. Dau, D. Püshel, and A. Kohlrausch. A quantitative model of the effective signal processing in the auditory system. I: Model structure. *J. Acoust. Soc. Am.*, 99(6): 3615–3622, June 1996.

[37] H. Davis. Advances in the neurophysiology and neuroanatomy of the cochlea. *J. Acoust. Soc. Am.*, 34(9):1377–1385, September 1962.

[38] J. Deller Jr., J. Proakis, and J. Hansen. *Discrete Time Processing Of Speech Signals*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1993.

[39] D.D. Dirks, D.E. Morgan, and J.R. Dubno. A procedure for quantifying the effects of noise on speech recognition. *J. Speech Hear. Disord.*, 47(2):114–123, May 1982.

[40] T.R. Dolan and D.E. Robinson. Explanation of masking-level differences that result from inter-aural intensive disparities of noise. *J. Acoust. Soc. Am.*, 42(5):977–981, November 1967.

[41] W.A. Dreschler, D. Eberhardt, and P.W. Melk. The use of single-channel compression for the improvement of speech intelligibility. *Scand. Audiol.*, 13(4):231–236, 1984.

[42] N.I. Durlach. Note on the equalization and cancellation theory of binaural masking level differences. *J. Acoust. Soc. Am.*, 32(8):1075–1076, August 1960.

[43] N.I. Durlach. Equalization and cancellation theory of binaural masking-level differences. *J. Acoust. Soc. Am.*, 35(8):1206–1218, August 1963.

[44] N.I. Durlach. On the application of the ec model to interaural jnds. *J. Acoust. Soc. Am.*, 40(6):1392–1397, December 1966.

[45] N.I. Durlach and H.S. Colburn. Binaural signal detection: equalization cancellation theory. In J.V. Tobias, editor, *Foundation of modern auditory theory*, pages 365–466. Academic Press, New York, NY, USA, 1972.

[46] O. Dyrlund. Characterization of non-linear distortion in hearing aids using coherence: A pilot study. *Scan Audiol.*, 18:143–148, 1989.

[47] O. Dyrlund. Coherence measurements in hearing instruments, using different broadband signals. *Scan Audiol.*, 21:73–78, 1992.

[48] O. Dyrlund, C. Ludvigsen, Å Olofsson, and T. Poulsen. Hearing aid measurements with speech and noise signals. *Scan Audiol.*, 23:153–157, 1994.

[49] Cherry E.C. Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.*, 25(5):975–979, September 1953.

[50] B.A. Edmonds and H.J.M. Culling. The spatial unmasking of speech: evidence for better-ear listening. *J. Acoust. Soc. Am.*, 120(3):1539–1545, September 2006.

[51] J.P. Egan. Masking level differences as a function of interaural disparities in intensity of signal and of noise. *J. Acoust. Soc. Am.*, 38(6):1043–1049, Dec 1965.

[52] N.N. Ellaham. Neural network model of human inaccuracy in binaural speech intelligibility prediction. Technical report, University of Ottawa, April 2009.

[53] N.N. Ellaham, C. Giguère, and W. Gueaieb. Binaural objective intelligibility measurement and hearing aids. In *Canadian Acoustics*, 37(3), pages 136–137, Niagara-on-the-lake, ON, October 2009.

[54] N.N. Ellaham, C. Giguère, and W. Gueaieb. A Wiener-based implementation of equalization-cancellation pre-processing for binaural speech intelligibility prediction. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October 2009.

[55] N.N. Ellaham, C. Giguère, and W. Gueaieb. Prediction of binaural speech intelligibility when using non-linear hearing aids. In *J. Acoust. Soc. Am.*, volume 123 of *Proc. 159th Meeting of the Acoustical Society of America*, page 1848, Baltimore, MD, April 2010.

[56] N.N. Ellaham, C. Giguère, and W. Gueaieb. Objective characterization of binaural speech intelligibility with hearing aids. In *Proc. Int. Hear. Aid Res. Conf.*, Lake Tahoe, CA, August 2010.

[57] N.N. Ellaham, C. Giguère, and W. Gueaieb. Evaluation of two speech and noise estimation methods for the assessment of nonlinear hearing aids. In *Proc. 36th IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 2011.

[58] N.N. Ellaham, C. Giguère, and W. Gueaieb. Evaluation of the phase-inversion signal separation method when using nonlinear hearing aids. *IEEE Trans. Audio, Speech, and Language Processing*, 21(4):879–888, April 2013.

[59] N.N. Ellaham, C. Giguère, and W. Gueaieb. A new research environment for speech testing using hearing-device processing algorithms. In *Canadian Acoustics*, 42(3), pages 89–90, Winnipeg, MB, October 2014.

[60] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.*, 32(6):1109–1121, December 1984.

[61] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.*, 33(2):443–445, April 1985.

[62] J.S. Erkelens, R.C. Hendriks, R. Heusdens, and J. Jensen. Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors. *IEEE Trans. Acoust. Speech Signal Process.*, 15(6):1741–1752, August 2007.

[63] Etymotic Research. The SIN test, 1993. (Compact Disk).

[64] Etymotic Research. Bamford-Kowal-Bench speech-in-noise test, 2005. (Compact Disk).

[65] G. Fairbanks. Test of phonemic differentiation: The rhyme test. *J. Acoust. Soc. Am.*, 30(7):596–600, July 1958.

[66] T.H. Falk, C. Zheng, and W.Y. Chan. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Trans. Aud. Sp. Lang. Proc.*, 18(7):1766–1774, September 2010.

[67] A.L. Fawe. Interpretation of infinitely clipped speech properties. *IEEE Trans. Audio Electro-acoustics*, 14(4):178–183, December 1966.

[68] J.M. Festen and R. Plomp. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *J. Acoust. Soc. Am.*, 88(4):1725–1736, October 1990.

[69] H. Fletcher. Auditory patterns. *Reviews of Modern Physics*, 12:47–65, 1940.

[70] H. Fletcher and R.H. Galt. The perception of speech and its relation to telephony. *J. Acoust. Soc. Am.*, 22(2):89–151, March 1950.

[71] N.R. French and J.C. Steinberg. Factors governing the intelligibility of speech sounds. *J. Acoust. Soc. Am.*, 19(1):90–119, January 1947.

[72] W.G. Gardner and K.D. Martin. HRTF measurements of a KEMAR dummy-head microphone. `http://sound.media.mit.edu/resources/KEMAR.html`, 1994. Accessed on: July 13, 2010.

[73] W.G. Gardner and K.D. Martin. HRTF measurements of a KEMAR dummy head microphone. Technical report, MIT Media Lab, 1994. Perceptual Computing Technical Report #280.

[74] W.G. Gardner and K.D. Martin. HRTF measurements of a KEMAR. *J. Acoust. Soc. Am.*, 96(6):3907–3908, June 1995.

[75] R. Ghent, M. Nilsson, and V. Bray. Validation of the noise reduction index (NRI). In *153rd Meeting of the Acoustical Society of America*, Salt Lake City, Utah, June 2007.

[76] C. Giguère, C. Laroche, S.D. Soli, and V. Vaillancourt. Functionally-based screening criteria for hearing-critical jobs based on the hearing in noise test. *Int. J. Aud.*, 47(6):319–328, June 2008.

[77] C. Giguère, C. Laroche, V. Vaillancourt, and S.D. Soli. Modelling speech intelligibility in the noisy workplace for normal-hearing and hearing-impaired listeners using hearing protectors. *Int. J. Acoust. Vib.*, 15(4):156–167, Decenber 2010.

[78] C. Giguère, C Laroche, and V. Vaillancourt. Speech intelligibility in military noise for normal-hearing and hearing-impaired listeners using level-dependent tactical hearing protectors. In *J. Acoust. Soc. Am.*, volume 131 of *Acoustics 2012*, page 3310, Hong Kong, May 2012.

[79] C. Giguère, C Laroche, V. Vaillancourt, E. Shmigol, T Vaillancourt, J. Chiasson, and V. Rozon-Gauthier. A multidimensional evaluation of auditory performance in one powered electronic level-dependent hearing protector. In *19th International Congress on Sound and Vibration*, Vilnius, Lithuania, July 2012.

[80] B.R. Glasberg and B.C.J. Moore. Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments. *J. Acoust. Soc. Am.*, 79(4):1020–1033, April 1986.

[81] B.R. Glasberg and B.C.J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1-2):103–138, August 1990.

[82] R.L. Goldsworthy and J.E. Greenberg. Analysis of speech-based speech transmission index methods with implications for nonlinear operations. *J. Acoust. Soc. Am.*, 116(6):2395–2405, December 2004.

[83] A.H. Gray Jr. and J.D. Markel. Distance measures for speech processing. *IEEE Trans. Acoust. Speech Signal Process*, 24(5):380–391, October 1976.

[84] D.M. Green and J.A. Swets. Signal detection analysis of the equalization and cancellation model. *J. Acoust. Soc. Am.*, 40(4):833–838, October 1966.

[85] D.M. Green and J.A. Swets. *Signal detection theory and psychophysics*. Peninsula Publishing, Los Altos, 1988.

[86] Y. Gui and H.K. Kwan. Adaptive subband wiener filtering for speech enhancement using critical-band gammatone filterbank. In *48th Midwest Symposium on Circuits and Systems*, pages 732–735, Cincinnati, OH, August 2005.

[87] B. Hagerman and A. Olofsson. A method to measure the effect of noise reduction algorithms using simultaneous speech and noise. *Acust. Acta Acust.*, 90(2):356–361, March/April 2004.

[88] T. Herzke and V. Hohmann. Improved numerical methods for gammatone filterbank analysis and synthesis. *Acust. Acta Acust.*, 93(3):498–500, June 2007.

[89] G. Hilkhuysen, N. Gaubitch, M. Brookes, and M. Huckvale. Effects of noise suppression on intelligibility: dependency on signal-to-noise ratios. *J. Acoust. Soc. Am.*, 131(1):531–539, January 2012.

[90] I.J. Hirsh, H. Davis, S.R. Silverman, E.G. Reynolds, E. Eldert, and R.W. Benson. Development of materials for speech audiometry. *J. Speech Hear. Disord.*, 17(3): 321–337, September 1952.

[91] V. Hohmann. Frequency analysis and synthesis using a gammatone filterbank. *Acust. Acta Acust.*, 88(3):433–442, May/June 2002.

[92] V. Hohmann and B. Kollmeier. The effect of multichannel dynamic compression on speech intelligibility. *J. Acoust. Soc. Am.*, 97(2):1191–1195, February 1995.

[93] I. Holube and B. Kollmeier. Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. *J. Acoust. Soc. Am.*, 100(3):1703–1716, September 1996.

[94] A.S. House, C.E. Williams, M.H.L. Hecker, and K.D. Kryter. Articulation-testing methods: Consonantal differentiation with a closed-response set. *J. Acoust. Soc. Am.*, 37(1):158–166, Jan 1965.

[95] T. Houtgast and H.J.M. Steeneken. Evaluation of speech transmission channels by using artificial signals. *Acustica*, 25:355–367, 1971.

[96] T. Houtgast and H.J.M. Steeneken. The modulation transfer function in room acoustics as a predictor of speech intelligibility. *Acustica*, 28:66–73, 1973.

[97] T. Houtgast, H.J. Steeneken, and A.W. Bronkhorst. Speech communication in noise with strong variations in the spectral or the temporal domain. In *Proceedings of the 14th International Congress on Acoustics*, 2, pages H2–6, Beijing, China, 1992.

[98] G. Hu and D.L. Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Trans. Neural Net.*, 15(5):1135–1150, September 2004.

[99] Y. Hu and P.C. Loizou. Subjective comparison and evaluation of speech enhancement algorithms. *Speech Comm.*, 49(7):588–601, July 2007.

[100] Y. Hu and P.C. Loizou. A comparative intelligibility study of single-microphone noise reduction algorithms. *J. Acoust. Soc. Am.*, 122(3):1777–1786, September 2007.

[101] Y. Hu and P.C. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Audio Speech Lang. Process.*, 16(1):229–238, January 2008.

[102] Pollack I. and J.M. Picket. Masking of speech by noise at high noise levels. *J. Acoust. Soc. Am.*, 30(2):127–130, February 1958.

[103] IEC 60268-16. Sound system equipment - part 16: Objective rating of speech intelligibility by speech transmission index. *International Electrotechnical Commission*, Geneva, Switzerland 1998.

[104] T. Irino and R.D. Patterson. A dynamic compressive gammachirp auditory filterbank. *IEEE trans. audio, speech and language proc.*, 14(2):2222–2232, November 2006.

[105] ISO 226 –2003. Acoustics—normal equal-loudness-level contours. *International Organization for Standardization*, Geneva, Switzerland 2003.

[106] ISO 389-7 –1996. Acoustics—reference zero for the calibration of audiometric equipment. part 7: Reference threshold of hearing under free-field and diffuse-field listening conditions. *International Organization for Standardization*, Geneva, Switzerland 1996.

[107] F. Itakura and S. Saito. A statistical method for estimation of speech spectral density and formant frequencies. *Electronic Communication in Japan*, 53(A):36–43, 1970.

[108] L.A. Jeffress. A place theory of sound localization. *J Comp. Physiol. Psychol.*, 41 (1):35–39, February 1948.

[109] L.A. Jeffress. Binaural signal detection: Vector theory. In J.V. Tobias, editor, *Foundation of modern auditory theory*, pages 351–368. Academic Press, New York, NY, USA, 1972.

[110] L.A. Jeffress and D. McFadden. Differences of inter-aural phase and level in detection and lateralization. *J. Acoust. Soc. Am.*, 49(4B):1169–1179, April 1971.

[111] L.A. Jeffress, H.C. Blodgett, T.T. Sandel, and Wood. C.L. Masking of tonal signals. *J. Acoust. Soc. Am.*, 28(3):416–426, May 1956.

[112] L.M. Jenstad and P.E. Souza. Quantifying the effect of compression hearing aid release time on speech acoustics and intelligibility. *J. Speech Language Hearing Research*, 48(3):651–667, June 2005.

[113] J. Jerger, C. Speaks, and J.L. Trammell. A new approach to speech audiometry. *J. Speech Hear. Disord.*, 33(4):318–328, November 1968.

[114] S. Jørgensen and T. Dau. Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *J. Acoust. Soc. Am.*, 130(3):1475–1487, September 2011.

[115] S. Jørgensen, S.D. Ewert, and T. Dau. A multi-resolution envelope-power based model for speech intelligibility. *J. Acoust. Soc. Am.*, 134(1):436–446, July 2013.

[116] S. Kamath and P.C. Loizou. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, volume 4, pages IV–4164, Orlando, FL, May 2002.

[117] S.D. Kamath. *A multi-band spectral subtraction method for speech enhancement (Thesis).* The University of Texas at Dallas, Dallas, TX, USA, 2001.

[118] J.M. Kates. On using coherence to measure distortion in hearing aids. *J. Acoust. Soc. Am.*, 91(4):2236–2244, April 1992.

[119] J.M. Kates. Cross-correlation procedures for measuring noise and distortion in AGC hearing aids. *J. Acoust. Soc. Am.*, 107(6):3407–3414, June 2000.

[120] J.M. Kates. *Digital hearing aids*. Plural publishing, San Diego, CA, USA, 2008.

[121] J.M. Kates. Understanding compression: modeling the effects of dynamic-range compression in hearing aids. *Int. J. Audiol.*, 49(6):395–409, June 2010.

[122] J.M. Kates. An auditory model for intelligibility and quality predictions. In *Proc. of Meetings on Acoustics*, Montreal, QC, June 2013.

[123] J.M. Kates. Improved estimation of frequency importance functions. *J. Acoust. Soc. Am.*, 134(5):459–464, October 2013.

[124] J.M. Kates and K.H. Arehart. Coherence and the speech intelligibility index. *J. Acoust. Soc. Am.*, 117(4):2224–2236, April 2005.

[125] J.M. Kates and K.H. Arehart. A model of speech intelligibility and quality in hearing aids. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 53–56, New Paltz, NY, October 2005.

[126] J.M. Kates and K.H. Arehart. The hearing-aid speech quality index (HASQI). *J. Aud. Eng. Soc.*, 58(5):363–381, May 2010.

[127] J.M. Kates and K.H. Arehart. The hearing-aid speech perception index (HASPI). *Speech Communication*, 65:75–93, November/December 2014.

[128] M.C. Killion. Revised estimate of minimal audible pressure: Where is the "missing 6 dB"? *J. Acoust. Soc. Am.*, 63(5):1501–1510, May 1978.

[129] M.C. Killion. Equalization filter for eardrum-pressure recording using a kemar manikin. *J. Aud. Eng. Soc.*, 27(1/2):13–16, January/February 1979.

[130] M.C. Killion, P.A. Niquette, G.I. Gudmundsen, L.J. Revit, and S. Banerjee. Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.*, 116(4): 2395–2405, October 2004.

[131] A.B. King and M.C. Martin. Is AGC beneficial in hearing aids? *British J. Audiol.*, 18(1):31–38, February 1984.

[132] D. Klatt. Prediction of perceived phonetic distance from criticalband spectra: A first step. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, Paris, France, May 1982.

[133] G. Kubin and W.B. Kleijin. On speech coding in a perceptual domain. In *Proc. IEEE Conf. Acoust., Speech, and Signal Proc.*, pages 205–208, Phoenix, AZ, March 1999.

[134] T.L. Langford and L.A Jeffress. Effect of noise crosscorrelation on binaural signal detection. *J. Acoust. Soc. Am.*, 36(8):1455–1458, August 1964.

[135] C Laroche, J.G. Roveda, J. Levionnois, C. Giguère, and Vaillancourt. Binaural speech recognition in continous and intermittent noises in people with hearing loss. In *Proc. of Meetings on Acoustics*, Montreal, QC, June 2013.

[136] H. Levitt. Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am.*, 49(2):467–477, February 1971.

[137] H. Levitt and L.R. Rabiner. Binaural release from masking for speech and gain intelligibility. *J. Acoust. Soc. Am.*, 42(3):601–608, September 1967.

[138] H. Levitt and L.R. Rabiner. Predicting binaural gain in intelligibility and release from masking for speech. *J. Acoust. Soc. Am.*, 42(4):820–829, October 1967.

[139] P.H. Lindsay and D.A. Norman. *Human information processing.* Academic Press, New York, NY, USA, 1972.

[140] R.Y. Litovsky. Spatial release from masking. *Acoustics Today*, 8(2):18–25, April 2012.

[141] P.C. Loizou. *Speech enhancement: theory and practice*. CRC Press, Boca Raton, FL, USA, 2007. ISBN 9780849350320.

[142] C. Ludvigsen, C. Elberling, and G. Keisder. Evaluation of a noise reduction method - comparison between observed scores and scores predicted from STI. *Scan. Audiol. Suppl.*, 38:50–55, 1993.

[143] F.J. Lundin. The influence of room reverberation on speech - an acoustical study of speech in a room. *J. Speech Trans. Lab.-QPSR*, 23(2-3):24–59, 1982.

[144] J. Ma, Y. Hu, and P. Loizou. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *J. Acoust. Soc. Am.*, 125(5):3387–3405, May 2009.

[145] R. Martin. Speech enhancement using mmse short time spectral estimation with gamma distributed speech priors. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, volume 1, pages I–253 – I–256, Orlando, FL, May 2002.

[146] D. McFadden. Masking-level differences determined with and without inter-aural disparities in masker intensity. *J. Acoust. Soc. Am.*, 44(1):212–223, April 1968.

[147] M.J. Middelweerd, J.M. Festen, and R. Plomp. Difficulties with speech intelligibility in noise in spite of a normal pure-tone audiogram. *Int. J. Aud.*, 17(1):1–7, January 1990.

[148] B.C.J. Moore. *Perceptual consequences of cochlear damage*. Oxford University Press, Oxford, UK, 1995.

[149] B.C.J. Moore. *Cochlear hearing loss: physiological, psychological and technical issues*. John Wiley & Sons, West Sussex, England, second edition, 2007.

[150] B.C.J. Moore. *Introduction to the psychology of hearing*. Emerald Group Publishing Limited, Wagon Lane, Bingley BD16 1WA, UK, sixth edition, 2012.

[151] H. Müsch. MATLAB implementation of core aspects of the American National Standard 'Methods for calculation of the speech intelligibility index' ANSI S3.5-1997. Downloadable MATLAB Script: `http://www.sii.to`, 2003. Accessed on: April 17, 2010.

[152] G. Naylor, F.M. Rønne, and R.B. Johannesson. Perceptual correlates of the long-term snr change caused by fast-acting compression. In *Proc. Int. Hear. Aid Res. Conf.*, Lake Tahoe, CA, August 2008.

[153] M. Nilsson, S.D. Soli, and J.A. Sullivan. Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *J. Acoust. Soc. Am.*, 95(2):1085–1099, February 1994.

[154] Å. Olofsson and M. Hansen. Objectively measured and subjectively perceived distortion in nonlinear systems. *J. Acoust. Soc. Am.*, 120(6):3759–3769, December 2006.

[155] H.L. Olsen, Å Olofsson, and B. Hagerman. The effect of audibility, signal-to-noise ratio, and temporal speech cues on the benefit from fast-acting compression in modulated noise. *Int. J. Aud.*, 44(7):421–433, July 2005.

[156] A.V. Oppenheim and R.W. Schafer. *Discrete-Time Signal Processing*. Prentice Hall, Upper Saddle River, NJ, USA, third edition, 2010.

[157] E. Osman. A correlation model of binaural masking level differences. *J. Acoust. Soc. Am.*, 50(6B):1494–1511, December 1971.

[158] H. Othman. Hearing aid simulator toolbox (hastb) v.1.4. Technical report, University of Ottawa, April 2009.

[159] D.P. Pascoe. Clinical measurement of the auditory dynamic range and their relation to formulas for hearing aid gain. In J.H. Jensen, editor, *Hearing Aid Fitting*, pages 129–154. Copenhagen : Storgaard Jensen, 1988.

[160] R.D. Patterson. Auditory filter shapes derived with noise stimuli. *J. Acoust. Soc. Am.*, 59(3):640–654, March 1976.

[161] K.L. Payton and M. Shrestha. Evaluation of short-time speech-based intelligibility metrics. In *Proc. Int. Commission on Biol. Effects Noise*, Foxwoods, CT, July 21–25 2008.

[162] K.L. Payton, L.D. Braida, S. Chen, P. Rosengard, and R.L. Goldsworthy. Chapter 11. computing the STI using speech as a probe stimulus. In S.J. van Winjgaarden, editor, *Past, present and future of the Speech Transmission Index*, pages 125–138. TNO Human Factors, Soesterberg, The Netherlands, 2002.

[163] J. Peissig and B. Kollmeier. Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners. *J. Acoust. Soc. Am.*, 101(3):1660–1670, March 1997.

[164] R.W. Peters, B.C.J. Moore, and T. Baer. Speech reception thresholds in noise with and without spectral and temporal dips for hearing impaired and normal hearing people. *J. Acoust. Soc. Am.*, 103(1):577–587, Jan 1998.

[165] H.-J. Platte and H. vom Hövel. Zur deutung der ergebnisse von sprachverständlichkeitsmessungen mit störschall im freifeld. *Acta Acustica united with Acustica*, 45(3):139–150, July 1980.

[166] R. Plomp and A.M. Mimpen. Effect of the orientation of the speaker's head and the azimuth of the noise source on speech-reception threshold for sentences. *Acustica*, 48(4):325–328, August 1981.

[167] D.A. Preves. Expressing hearing aid noise and distortion with coherence measurements. *ASHA*, 32(6-7):56–59, June/July 1990.

[168] R. Rahal, W. Gueaieb, C. Giguère, and H. Othman. Automatic volume settings for environment sensitive hearing aids. In *Canadian Acoustics*, 37(3), pages 130–131, Niagara-on-the-lake, ON, October 2009.

[169] C.M. Reed, L.D. Braida, and P.M. Zurek. Change in prevalence of hearing loss in US adolescents. *Trends in Amplification*, 13(1):4–43, March 2009.

[170] K.S. Rhebergen and N.J. Versfeld. A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *J. Acoust. Soc. Am.*, 117(4):2181–2192, April 2005.

[171] K.S. Rhebergen, Versfeld N.J., and W.A. Dreschler. Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. *J. Acoust. Soc. Am.*, 120(6):3988–3997, December 2006.

[172] K.S. Rhebergen, Versfeld N.J., and W.A. Dreschler. Prediction of the intelligibility for speech in real-life background noises for subjects with normal hearing. *Ear and Hearing*, 29(2):169–175, April 2008.

[173] K.S. Rhebergen, N.J. Versfeld, J.A.P.M. de Laat, and W.A. Dreschler. Modelling the speech reception threshold in non-stationary noise in hearing-impaired listeners as a function of level. *Int. J. Aud.*, 49(11):856–865, November 2010.

[174] JG. Roveda. Effet du type de bruit sur la reconnaissance de la parole mesurée à l'aide du HINT. Technical report, University of Ottawa, August 2011.

[175] A.F. Ryan and P. Dallos. Physiology of the cochlea. In J.L. Northern, editor, *Hearing Disorders*, pages 253–266. Little Brown, Boston, MA, USA, 1984.

[176] A Sarampalis, S. Kalluri, B. Edwards, and Hafter E. Objective measures of listening effort: Effects of background noise and noise reduction. *J. Speech Language Hearing Research*, 52(5):1230–1240, October 2009.

[177] B.M. Sayers and E.C. Cherry. Revised estimate of minimal audible pressure: Where is the "missing 6 dB"? *J. Acoust. Soc. Am.*, 29(9):973–987, September 1957.

[178] P. Scalart and J. Filho. Speech enhancement based on a priori signal to noise estimation. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, volume 2, pages 629–632, Atlanta, GA, May 1996.

[179] J. Shargorodsky, S.G. Curhan, G.C. Curhan, and R. Eavey. Review of the literature on temporal resolution in listeners with cochlear hearing impairment: A critical assessment of the role of suprathreshold deficits. *Journal of the American Medical Association*, 304(7):772–778, August 2010.

[180] E.A.G. Shaw. Transformation of sound pressure level from the free field to the eardrum in the horizontal plane. *J. Acoust. Soc. Am.*, 55(6):1848–1861, December 1974.

[181] B. Sim, Y. Tong, J. Chang, and C. Tan. A parametric formulation of the generalized spectral subtraction method. *IEEE Trans. Acoust. Speech Signal Process.*, 6(4): 328–337, July 1998.

[182] K. Smeds, F. Wolters, S. Båsjö, S. Hertzman, and A. Leijon. Objective measures to quantify the perceptual effects of noise reduction in hearing aids. In *Proc. 38th Int. Conf. Audio Eng. Soc.*, Piteå, Sweden, June 2010.

[183] C. Smits and T. Houtgast. Recognition of digits in different types of noise by normal-hearing and hearing-impaired listeners. *Int. J. Aud.*, 46(3):134–144, March 2007.

[184] S.D. Soli. Use of the extended speech intelligibility index (ESII) and the hearing in noise test (HINT) to quantify functional hearing ability. In *UCLA Visitor Seminar*, Los Angeles, CA, October 07 2012.

[185] S.D. Soli and L.L.N. Wong. Assessment of speech intelligibility in noise with the hearing in noise test. *Int. J. Aud.*, 47(6):356–361, June 2008.

[186] P. Souza, K. Boike, K. Witherell, and K. Tremblay. Prediction of speech recognition from audibility in older listeners with hearing loss: Effects of age, amplification, and background noise. *J. Am. Acad. Audiol.*, 18(1):54–65, January 2006.

[187] P. Souza, Jenstad L., and K. Boike. Measuring the acoustic effects of compression amplification on speech in noise. *J. Acoust. Soc. Am.*, 119(1):41–44, January 2006.

[188] H.J.M. Steeneken and T. Houtgast. A physical method for measuring speech-transmission quality. *J. Acoust. Soc. Am.*, 67(1):318–326, January 1980.

[189] H.J.M. Steeneken and T. Houtgast. Mutual dependence of the octave-band weights in predicting speech intelligibility. *Speech Communication*, 28(2):109–123, June 1999.

[190] H.J.M. Steeneken and T. Houtgast. Phoneme-group specific octave-band weights in predicting speech intelligibility. *Speech Communication*, 38(3–4):399–411, November 2002.

[191] H.J.M. Steeneken and T. Houtgast. Validation of the revised STIr method. *Speech Communication*, 38(3–4):413–425, November 2002.

[192] IEEE Subcommittee. IEEE recommended practice for speech quality measurements. *IEEE Trans. on Audio and Electroacoustics*, 17(3):225–246, September 1969.

[193] D.J.O. Suelzle. *Electroacoustic and Behavioural Evaluation of Hearing Aid Digital Signal Processing Features (Thesis)*. University of Western Ontario, London, ON, Canada, 2013.

[194] Houtgast T., H.J.M. Steeneken, Ahnert W., L.D. Braida, R. Drullman, J.M. Festen, K. Jacob, P. Mapp, S. McManus, K.L. Payton, R. Plomp, J.A. Verhave, and S.J. van Wijngaarden. *Past, present and Future of the Speech Transmission Index*. TNO Human Factors, Soesterberg, The Netherlands, 2002.

[195] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen. Intelligibility prediction of single-channel noise-reduced speech. In *ITG-Fachtagung Sprachkommunikation*, Bochum, Germany, October 06–06 2010.

[196] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. Aud. Sp. Lang. Proc.*, 19(7):2125–2136, September 2011.

[197] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen. An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech. *J. Acoust. Soc. Am.*, 130(5):3013–3027, November 2011.

[198] C.T. Tan and B.C.J. Moore. Perception of nonlinear distortion by hearing-impaired people. *Int. J. Aud.*, 47(5):246–256, January 2008.

[199] THFC.CA - The Hearing Foundation of Canada. Statistics about hearing loss. `http://www.thfc.ca/cms/en/KeyStatistics/KeyStatistics.aspx`. Accessed on: September 16, 2012.

[200] J.M. Tribolet, P. Noll, B.J. McDermott, and R.E. Crochiere. A study of complexity and quality of speech waveform coders. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, volume 3, pages 586–590, Tulsa, OK, April 1978.

[201] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. Jensen. A perceptual model for sinusoidal audio coding based on spectral integration. *EURASIP J. Appl. Sign. Process*, 9:1292–1304, January 2005.

[202] S.J. Van Wijngaarden and R. Drullman. Binaural intelligibility prediction based on the speech transmission index. *J. Acoust. Soc. Am.*, 123(6):4514–4523, June 2008.

[203] S.V. Vaseghi. *Advanced Digital Signal Processing and Noise Reduction*. John Wiley and Sons, Ltd, Chichester, West Sussex, UK, fourth edition, 2008.

[204] A.J. Vermiglio. The american english hearing in noise test. *Int. J. Aud.*, 47(6): 386–387, June 2008.

[205] N.J. Versfeld and W.A. Dreschler. The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners. *J. Acoust. Soc. Am.*, 111(1):401–408, January 2002.

[206] H. vom Hövel. *Zur bedeutung der übertragungseigenschaften des außenohrs sowie des binauralen hörsystems bei gestörter sprach bertragng*. Dissertation, Fakultät für Elektrotechnik, RTWH Aachen, Germany, 1984.

[207] A. Vonlanthen and Arndt H. *Hearing instrument technology for the hearing health care professional*. Thomson Delmar Learning, Kitchener, ON, Canada, third edition, 2007.

[208] K. Wagener, J.L. Josvassen, and R. Ardenkjaer. Design, optimization and evaluation of a danish sentence test in noise. *Int. J. Audiol.*, 42(1):10–17, Jan 2003.

[209] P.D. Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Trans. Audio Electro-acoustics*, 15(2):70–73, June 1967.

[210] R. H. Wilson. Development of a speech-in-multitalker-babble paradigm to assess word-recognition performance. *J. Am. Acad. Audiol.*, 14(9):453–470, November 2003.

[211] R. H. Wilson and C.A. Burks. Use of 35 words for evaluation of hearing loss in signal-to-babble ratio: A clinic protocol. *J. Rehab. Res. Dev.*, 42(6):839–852, November-December 2005.

[212] Y-H. Wu and R.A. Bentler. A method to measure hearing aid directivity index and polar pattern in small and reverberant enclosures. *Int. J. Aud., Early Online*, 50(6):405–416, June 2011.

[213] B. Xie. *Hear-Related Transfer Function and Virtual Auditory Display.* J. Ross Publishing, Plantation, FL, USA, second edition, 2013.

[214] E. Zavarehei. Sample speech enhancement methods. `http://dea.brunel.ac.uk/cmsp/Home_esfandiar/SampleWaveFiles.htm`. Accessed on: June 2, 2012.

[215] C. Zerbs. *Modelling the effective binaural signal processing in the auditory system.* Dissertation, Carl-von-Ossietzky-Universität, Oldenburg, Germany, 2000.

[216] P.M. Zurek. Binaural advantages and directional effects in speech intelligibility. In G.A. Studebaker and I. Hockberg, editors, *Acoustical Factors Affecting Hearing Aid Performance*, chapter 15, pages 255–276. Allyn and Bacon, London, UK, 2nd edition, 1990.

[217] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and Models.* Springer, Berlin Heidelberg, third edition, 2007.