uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Thien-An Vu

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.A.Sc. (Electrical Engineering)

GRADE / DEGREE

School of Information Technology and Engineering

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Double Talk Detection Using a Psychoacoustic Auditory Model

TITRE DE LA THÈSE / TITLE OF THESIS

M. Bouchard

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

H. Ding

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

M. El Tanany

O. Yang

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Double Talk Detection Using a Psychoacoustic Auditory Model

by

Thien-An Vu

B.A.Sc., University of Waterloo, 1998

A thesis presented to

the Faculty of Graduate and Post-Doctoral Studies

in partial fulfillment of the requirement for the degree of

Master of Applied Science

in

Electrical and Computer Engineering

Ottawa-Carleton Institute for Electrical and Computer Engineering

School of Information Technology and Engineering

Faculty of Engineering

University of Ottawa

Ottawa, Ontario, Canada

April 2007

Your file Votre référence
ISBN: 978-0-494-32485-1
Our file Notre référence
ISBN: 978-0-494-32485-1

Abstract

Successful adaptive echo cancellation in telecommunications depends on a control device called a double talk detector. Double talk refers to the situation that signals at both ends of a communication link are simultaneously active. In the presence of a double talk situation, the role of a double talk detector is to assure convergence and to prevent divergence of the adaptive filter in an echo cancellation system.

Following a thorough study of the subject matter, this thesis presents a double talk detection algorithm for a single-channel echo cancellation application using a psychoacoustic auditory model. The model exploits the frequency masking properties of the human auditory system. It performs an analysis of the far-end signal to compute a perceptual threshold, and then removes spectral components below the perceptual threshold to create spectral holes without affecting the perceptual quality of the signal. Double talk conditions are detected by monitoring the spectral levels at the created holes in the near-end input signal.

To evaluate the proposed algorithm, results of simulations with real speeches and performance comparisons with certain other double talk detection algorithms are presented. It is shown with simulation results and analysis that the proposed algorithm outperforms those algorithms in overall.

Acknowledgements

First, I would like to express my sincere gratitude to my supervisors, Dr. Heping Ding and Professor Martin Bouchard, for their invaluable guidance and insightful advice in the world of research, especially to Dr. Ding for inspiring the topic of this thesis. Throughout my graduate study, they have always been prompt and available in giving me constructive feedbacks, careful reviews of papers and the thesis. Without them, this thesis would never have come into being.

I would like to thank Dr. Mike Stinson and his Acoustics and Signal Processing group at the National Research Council for the financial support and for allowing me to use the office and the facility for my graduate study. I would also like to thank Professor Rafik Goubran at the DSP lab of Carleton University, who kindly provided me with the room impulse response used in this thesis.

Finally, I would like to thank my wife and my family for their endless support and encouragement, especially my wife for her patience and personal sacrifice to support me throughout this study, and my sons for being my joy and motivation. Without them, there would be no purpose with this thesis. Thank you sincerely.

Ottawa, April 2007

Thien-An Vu

Table of Contents

List of Figures

List of Tables

List of Acronyms

AEC	Acoustic Echo Cancellation
AP	Affine Projection
ATH	Absolute Threshold of Hearing
AWGN	Additive White Gaussian Noise
CB	Critical Bands
Codec	Coder/decoder
CPU	Central Processing Unit
dB	Decibel
DC	Direct Current (i.e. 0 Hz)
DFT	Discrete Fourier Transform
DSP	Digital Signal Processing/Processor
DT	Double Talk
DTD	Double Talk Detection
EC	Echo Cancellation
EEC	Electrical Echo Cancellation
EERL	Excess Echo Return Loss
ENNR	Echo to Near-end Noise Power Ratio
EPC	Echo Path Change
ERL	Echo Return Loss
ERLE	Echo Return Loss Enhancement
FAP	Fast Affine Projection
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
fs	Sampling Frequency
HAS	Human Auditory System
IDFT	Inverse Discrete Fourier Transform
IFFT	Inverse Fast Fourier Transform
IR	Impulse Response

IIR	Infinite Impulse Response
IP	Internet Protocol
ITU	International Telecommunications Union
LEC	Line Echo Cancellation
LMS	Least-Mean-Square
LS	Least Squares
MAC	Multiply-Accumulate Operation
MIPS	Million Instruction per Second
MMSE	Minimum Mean-Squared Error
MOS	Mean Opinion Score
MSE	Mean-Squared Error
NCC	Normalized Cross Correlation
NEC	Network Echo Cancellation
NER	Near-en Speech to Echo Power Ratio
NLMS	Normalized Least Mean Square
NLP	Non-Linear Processor
P_f	Probability of False Alarm
P_m	Probability of Miss Detection
PSD	Power Spectral Density
PSTN	Public Switched Telephone Network
RLS	Recursive Least Squares
RMS	Root Mean Square
ROC	Receiver Operating Characteristic
SFM	Spectral Flatness Measure
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level
ST	Single Talk

Chapter 1
Introduction

Since the advent of telephony, echo has been a problem preventing comfortable full-duplex conversation in communication systems. So far, the adaptive echo cancellation (EC) has been the best method to remove the undesired echo in those systems. The cancellation can be done by modeling an echo path with an adaptive finite impulse response filter, and then subtracting the echo estimate from the near-end input signal, which consists of the echo and maybe some local signal(s).

One major issue of the EC is that performance of the adaptive filter may deteriorate drastically during double talk (DT) situations, in which speeches from both ends of a communication link are simultaneously active. In the presence of DT condition, the near-end speech signal mixes in with the echo, behaves as an uncorrelated noise to the adaptive filter, and causes the filter to diverge quickly from the optimum value if the adaptation process is not stopped. As a result, the undesired echo can pass through uncancelled. Furthermore, it takes some time for the adaptive filter to get back to a converged state once the double talk condition stops.

It is therefore a very important role of a DT detector to identify the presence of DT condition. Once the condition is detected, a common solution to prevent the divergence of the adaptive filter is to halt the adaptation process for some hold time. To address this problem, there have been many fundamentally different solutions proposed in the open literature. In

this thesis, a novel approach to DT detection (DTD) is investigated and analyzed, with the focus on the single-channel EC system using a finite-impulse-response (FIR) filter structure.

The outline of the remainder of this thesis is organized as follows:

- Chapter 2, "Fundamentals of Echo Cancellation," introduces fundamental problems associated with EC and control, reviews basic concepts and methodologies of EC as the foundations for the proposed DTD algorithm.

- Chapter 3, "Fundamentals of Psychoacoustics," reviews the human auditory system, and the principles of psychoacoustics used in speech processing, such as the frequency masking properties, which are exploited by the proposed DTD algorithm.

- Chapter 4, "A Survey of Double Talk Detection Algorithms," reviews common basics of a generic DTD scheme; summarizes and classifies existing DTD algorithms in the literature. The survey work presented in this chapter has been briefly published in [VU04].

- Chapter 5, "The Proposed Double Talk Detection Algorithm Using a Psychoacoustic Auditory Model," presents a DTD algorithm that exploits the frequency masking principle of psychoacoustics. Performance of the proposed algorithm is evaluated and compared with certain other typical DTD algorithms under various conditions via simulations. The computational complexity of the

proposed algorithm is analyzed and compared with that of the other DTD algorithms. This proposed algorithm has been published in [VU06].

- Chapter 6, "Conclusions and Future Works," provides conclusions of the thesis, and offers some recommendations for future work.

Chapter 2
Fundamentals of Echo Cancellation

Echo can significantly degrade quality of signal in a communication network. EC is therefore an essential part in the network. The cancellation relies on an adaptive filter to estimate the echo path of the network, and then use this estimate to reduce the echo to an acceptable level. This chapter introduces fundamental problems associated with echo control, and reviews basic concepts and methodologies of EC as the foundations for the development of the proposed DTD algorithm.

2.1 Echo in Communication Networks

Echo is an effect in which delayed, attenuated, and distorted versions of an original signal are reflected back to the source. It can significantly degrade the quality of speech depending on the associated roundtrip delay and the echo intensity. However, not all echoes reduce the quality of speech. For a natural phone conversation, a short-delay echo, called side-tone, is deliberately inserted so that a talker can hear himself/herself speaking. For roundtrip delays greater than 30 milliseconds, echo is obvious to most people; and for delays greater than 50 milliseconds, it becomes very disruptive. In addition, echo must be relatively loud enough to be heard. Echo with intensity less than 30 dB below its source is unlikely to be annoying. For intensity levels above that and roundtrip delays greater than 30 milliseconds, echo can be considered objectionable [KOS02] and [BENE01]. A few sources of echo can occur in communication networks, each with different causes and characteristics. Echoes can be

classified, based on the media where they are generated, into two main types: electrical echo and acoustic echo.

Electrical echo is primarily caused by the impedance mismatches between juncture network elements. In the public-switched telephone network (PSTN), it is generated by reflection from a hybrid, which is located at the juncture connecting the 2-wire subscriber loop and the 4-wire trunk at the local central office. The hybrid splits the 2-wire loop into two pairs of wires, i.e. 4-wire. One pair is used for transmission and the other for reception. If the impedance of the balancing network at the hybrid does not precisely match the impedance of the 2-wire loop, a small part of the signal passing through the hybrid in the 4-wire to the 2-wire direction is reflected back to the 4-wire circuit and then to the original source, thus creating echo. For local calls with relatively short distance or the roundtrip delay of less than 30 milliseconds, echo is usually not a problem because it is not heard by, or annoying to, the original speaker. It appears as a normal side-tone to the speaker in this case. As the roundtrip delay exceeds 30 milliseconds, echo can have some degree of annoying effect. For long distance calls, where the delay approaches 50 milliseconds or greater, distinctive echo can be heard [BENE01].

Acoustic echo is typically associated with use of electro-acoustic transducers, such as in hands-free systems. It is caused by acoustic coupling between a hands-free terminal's loudspeaker and microphone. A sound comes out of the speaker, bounces off objects in the acoustic environment, and reflects back to the microphone. Acoustic echo is therefore characterized by multiple reflections that create a very complex and sometimes non-linear

echo signal due to distortions in the electro-acoustic system, e.g., in electronics, the loudspeaker, and the microphone. Unlike the electrical case, acoustic echo path usually has a longer delay, which is determined by the size of the acoustic environment. The bigger the size, the longer the echo path delay can be. In addition, since the objects can move around in the environment, the acoustic echo path can be time varying.

In today's digital communication networks, the echo problem is even more significant because the processing delay is considerably higher than in the traditional circuit-switched long distance networks. In addition, the problem would be particularly more challenging because of nonlinearities associated with codec operations, variable delays associated with packet routing, and mixed network topologies.

2.2 Echo Control

The interest for echo control began in the late 1950s with the development of echo suppression devices [GOU64]. These echo suppressors were first used to manage the electrical echo generated in satellite circuits. They consisted essentially of voice-activated switches, which transmitted a voice path and then turned off the path to block any echo signal. Thus, the main idea behind this is that when a person speaks, signals from the other end are attenuated significantly, preventing the audibility of echo. Although the echo suppressors could remove the echo in the networks successfully, they also resulted in choppy first syllables and artificial volume fluctuations. Furthermore, they made full duplex communication impossible. A more sophisticated technique is to cancel the echo, rather than to suppress it. EC was first introduced by Sondhi and colleagues in the 1960s [SON66, 67],

[BEC66]; and it has been extensively studied in the literature since then. The basic principle of EC is to estimate a replica of the actual echo, and then remove the echo by simple subtraction. The echo estimation is accomplished by using an adaptive filter to model the impulse response of the echo path. EC leads to a capability of full duplex communication, which allows conversations that are more natural. These days, echo cancellers are widely used in all modern communication networks. In the following sub-sections, the basic concepts and components of an echo canceller are presented.

2.2.1 Echo Cancellation and Control

Inside an echo canceller, it is essentially an adaptive filter that is used for modeling the echo path. An adaptive filter consists of two fundamental elements: a filter and an adaptive algorithm. Figure 2-1 shows a basic configuration of EC. The notation of the diagram will be consistently maintained throughout the thesis and is summarized as below.

$\underline{\mathbf{h}}(n)$ vector consisting of samples of impulse response of the echo path;

$\underline{\hat{\mathbf{h}}}(n)$ vector consisting of weights, or coefficients, of the adaptive filter;

$x(n)$ the far-end speech, input to the adaptive filter;

$y(n)$ the actual echo signal, output of the echo path;

$\hat{y}(n)$ the echo estimate, output of the adaptive filter;

$v(n)$ the near-end speech;

w(n) the near-end background noise;

d(n) the near-end input signal;

e(n) the residual error signal, to be processed and sent to the far-end.

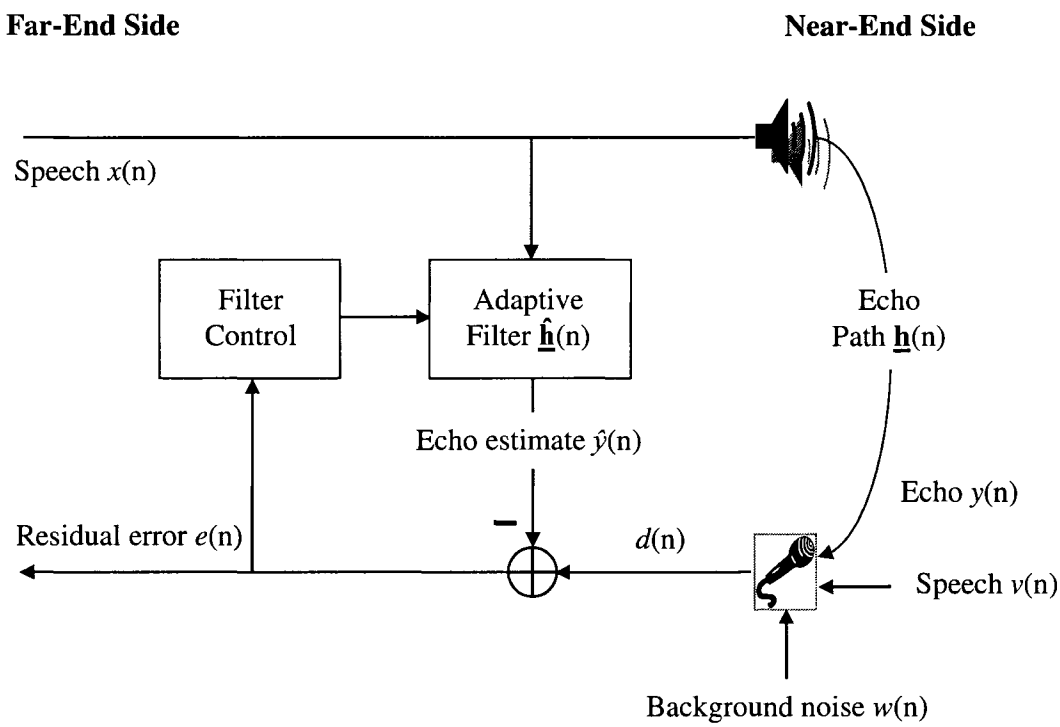

Figure 2-1: Basic Configuration of Echo Cancellation.

It can be seen from the diagram that EC is similar to a system identification problem. The adaptive filter aims to model the unknown system, i.e. the echo path, by adjusting its weights to replicate the impulse response of the system. The far-end speech x(n) is passed through both the filter and the echo path, with the output of the echo path being the actual echo signal

y(n) and the output of the filter being the echo estimate signal $\hat{y}$(n). The working principle of the adaptive filter is to minimize the difference between the echo and the echo estimate, known as the residual error signal e(n). The error signal is used by the adaptive algorithm, which, based on a cost function, updates the filter weights to typically minimize the mean square error (MSE) of e(n). The residual error signal is also passed along back to the far-end and heard as an echo if the adaptive filter does not cancel the echo effectively. Once the MSE of e(n) is minimized, the echo appearing at the far-end is minimized, and the adaptive filter is said to have converged. In practice, the residual error does not reach zero because $\hat{\underline{\mathbf{h}}}$(n) cannot be made long enough and e(n) is limited by the level of background noise w(n).

The above described a simple case of single talk (ST). Typical phone conversations involve DT periods, when speeches from both ends are simultaneously active. During DT condition, the near-end input signal d(n) consists of the near-end speech v(n), the echo y(n), and the near-end background noise w(n). Since v(n) and w(n) can be assumed uncorrelated with x(n) and therefore uncorrelated with y(n), the adaptive filter sees these signals as a noise, and tries to adapt its weights based on this noise. This leads to degradation of the adaptive filter so more echo will pass through uncancelled as a result. A DT detector is required to ensure proper adaptation of the adaptive filter. The role of a DT detector is to identify the presence of the DT condition. Once a DT condition is detected, the adaptation process can be halted or slowed down to prevent divergence of the adaptive filter. This seems like an easy problem but it is a very challenging task in practice. Moreover, the background noise should not be detected as a DT condition. In addition, a change in the echo path should not be detected as DT, as in this case the filter should try to adapt to the change.

The described EC scheme is a linear process and therefore it is very sensitive to the nonlinearity of the echo path such as that in the electronic circuit, the loudspeaker, and the microphone of a handsfree system. In addition, EC alone is unable to reduce the residual error to an acceptable level. This necessitates the usage of a nonlinear processor (NLP) in an echo canceller. The NLP acts as a center clipper and a comfort noise generator. It evaluates the residual error signal and substantially suppresses the residual echo, which remains at the output of an echo canceller. The NLP attenuates low-level signals below a threshold, which are assumed the residual echo, and replaces them with a simulated background noise, called comfort noise, which sounds like the original background noise without the echo. It passes the high-level signal, assuming that the signal corresponds to the desirable near-end speech $v(n)$. Note that it is important for the NLP to remove the residual echo precisely as fading parts of speech can be erroneously replaced with background noise, thus making a phone conversation choppy. In addition, the NLP has to handle the generation of the comfort noise properly so as not to make a phone conversation unnatural or distracting.

2.2.2 Performance Measurements

The performance of an echo canceller can be objectively measured using the echo return loss enhancement (ERLE). In the simulation environment where one can have full access to the impulse response of the echo path, the actual echo, and the background noise, other useful performance measurements can be also made, such as the misalignment and the excess echo return loss enhancement (EERLE).

2.2.2.1 ERLE

The performance of an echo canceller during ST situation is usually measured using the ERLE. It is defined as the power ratio of the echo signal y(n) to the residual error e(n). In ST situation, the near-end input signal d(n) can be assumed to contain only the echo signal y(n) provided that the background noise is very insignificant. The ERLE is therefore calculated as follows.

$$\text{ERLE}=10\log_{10}\left[\frac{\hat{\text{E}}\{d^2(n)\}}{\hat{\text{E}}\{e^2(n)\}}\right] \quad (dB), \tag{2.1}$$

where the operator $\hat{\text{E}}\{\bullet\}$ returns the argument's expected value, which is usually estimated via a time averaging process. The ERLE is therefore the measurement of the attenuation of the echo signal achieved by the echo canceller. The higher the value of ERLE, the better the performance of the echo canceller is. The ERLE however does not include any further reduction in the residual echo by the NLP. The plot of ERLE versus time shows the convergence and the steady-state residual echo, so it is a good indication of the performance of the echo canceller during ST situation.

2.2.2.2 EERLE

During DT situation, i.e. the near-end speech v(n) is present, the EERLE is a more useful performance measure than the ERLE because it is a ratio of the actual echo to the residual error after cancellation. The EERLE is calculated as follows,

$$\text{EERLE}=10\log_{10}\left[\frac{\hat{\text{E}}\left\{\left[d(n)-v(n)-w(n)\right]^2\right\}}{\hat{\text{E}}\left\{\left[e(n)-v(n)-w(n)\right]^2\right\}}\right]$$
$$=10\log_{10}\left[\frac{\hat{\text{E}}\left\{\left[y(n)\right]^2\right\}}{\hat{\text{E}}\left\{\left[y(n)-\hat{y}(n)\right]^2\right\}}\right] \quad \text{(dB)}. \tag{2.2}$$

The EERLE is therefore a measure of the attenuation of the echo signal achieved by the echo canceller so the highest possible value of EERLE is desired. Since the near-end speech signal and the background noise are not taken into consideration in the measurement, the EERLE plot is a good indication of the performance of the echo canceller during both ST and DT situations.

2.2.2.3 Misalignment

The misalignment is a measurement of how well the adaptive filter estimates the actual impulse response of the echo path. It is defined as the normalized Euclidean distance between the vector consisting of the coefficients of the adaptive filter and the actual echo path's impulse response truncated to the same length of the adaptive filter.

$$\text{Misalignment} = 20\log_{10}\left[\frac{\left\|\underline{\text{h}}(n)-\hat{\underline{\text{h}}}(n)\right\|}{\left\|\underline{\text{h}}(n)\right\|}\right] \quad \text{(dB)}. \tag{2.3}$$
$$\text{Misalignment} = 20\log_{10}$$

A small misalignment is therefore desired. Note that for input speech signals, a low value of misalignment does not necessarily mean a large ERLE. That is because the spectrum of a short portion of speech often has very distinct bands of frequencies. If the echo canceller

does not model the echo path very well in these bands, the ERLE is poor even if the overall weight error can be low. Therefore, the misalignment must be used with caution.

2.2.3 Acoustic versus Electrical Echo Cancellation

Both acoustic EC (AEC) and electrical EC (EEC) address similar problems, and rely on adaptive filters to remove echo. However, there are some differences in the characteristics of the acoustic echo and the electrical echo. Therefore, their requirements and expected performance are different from each other. In general, the AEC is a more difficult problem because of the underlying facts, discussed in the following.

The acoustic echo path is generally much longer than the electrical echo path. Typically, the delay of the electrical echo path in the network is in the range of 8-128 milliseconds whereas the delay of the acoustic echo path is in the range 60-200 milliseconds or even more. As a result, the adaptive filter requires a much larger number of weights in the AEC than in the NEC. This is mainly attributed to the lower speed that sound waves travel through air, compared to the speed at which electrical signals travel over circuits. In addition, multiple reflections off objects in the acoustic environment serve to lengthen the impulse response; whereas the electrical echo path in a network generally has one or two reflections from the impedance mismatches.

Furthermore, the acoustic echo path can be time varying depending on the environmental conditions such as ambient temperature, pressure and humidity; and movements of objects such as doors and humans. Therefore, variations of echo path must be tracked not only while

the first establishment of a connection but also over the duration of the connection. On the other hand, the echo path of a network can often be different from one connection to the next because of different impedances of the subscribed loops, but it changes very little during a single connection.

In AEC, the echo level before cancellation is generally much higher than in EEC. In EEC, an echo return loss of 6 dB is the typical hybrid attenuation although in some cases it can be as high as 11-15 dB. However, the magnitude of the echo signal $y(n)$ in AEC is usually higher than the near-end speech $v(n)$ because of the proximity of the loudspeaker to the microphone in a hands-free terminal and the presence of amplifiers in the echo path. In general, a larger quantity of echo cancellation is required in the AEC.

2.3 Adaptive Filtering

The theoretical basis for echo cancellers is in the field of adaptive filtering [MUR90]. Indeed, an echo canceller essentially consists of an adaptive filter for modeling the echo path. An adaptive filter consists of two fundamental elements: a filter with adjustable parameters for performing a desired processing function; and an adaptive algorithm for adjusting the parameters of the filter. The following sub-sections present an overview of adaptive filtering theory that can be applied to EC, and will be used as a foundation for the development of the proposed DTD algorithm. The materials presented in the followings are referenced from [HAY96], [TRE01], [WID99], [AHG04], [LIN89], and [CRE99]. One can refer to them for detailed treatment on the topic.

2.3.1 Filter Structure

Many types of filter structures can be used for the adaptive filtering. For EC, the most commonly used structure is the transversal FIR filter, also known as a tapped delay line, as shown in Figure 2-2. The FIR filter consists of a series of one-sample delays denoted as z^{-1}, multipliers, and adders. The input x(n) is fed into the tapped delay line of length L, spaced equally with time interval T. Note that $1/T$ is the sampling rate of the discrete time system. Each tap input is weighted by a filter coefficient $\hat{h}_i$(n). The output of the FIR filter, the convolution of the input and the filter weights, is expressed as a linear combination of the weighted tap inputs as follows.

$$\hat{y}(n) = x(n) * \hat{h}(n) = \sum_{i=0}^{L-1} \hat{h}_i(n) x(n-i). \tag{2.4}$$

Equation 2.4 can also be expressed more compactly in vector notation as $\hat{y}(n) = \hat{\underline{\mathbf{h}}}^{\mathrm{T}}(n)\underline{\mathbf{x}}(n)$, where the weight vector is defined as $\underline{\mathbf{h}}^{\mathrm{T}}(n) = \begin{bmatrix} h_0(n) & h_1(n) & \cdots & h_{L-1}(n) \end{bmatrix}$ and the input vector consists of the L latest inputs $\underline{\mathbf{x}}(n) = \begin{bmatrix} x(n) & x(n-1) & \cdots & x(n-L+1) \end{bmatrix}^{\mathrm{T}}$. It can be seen that the FIR structure contains only zeros in its transfer function. This makes the FIR filter inherently more stable than other structures such as the infinite impulse response (IIR) filter. The IIR filter's transfer function generally contains both zeros and poles so its output would oscillate indefinitely if the poles move outside the unit circle. One disadvantage of the FIR filter is that its impulse response is limited in time by the number of tap weights. For AEC applications with very long echo paths, the FIR filter requires a large number of weights.

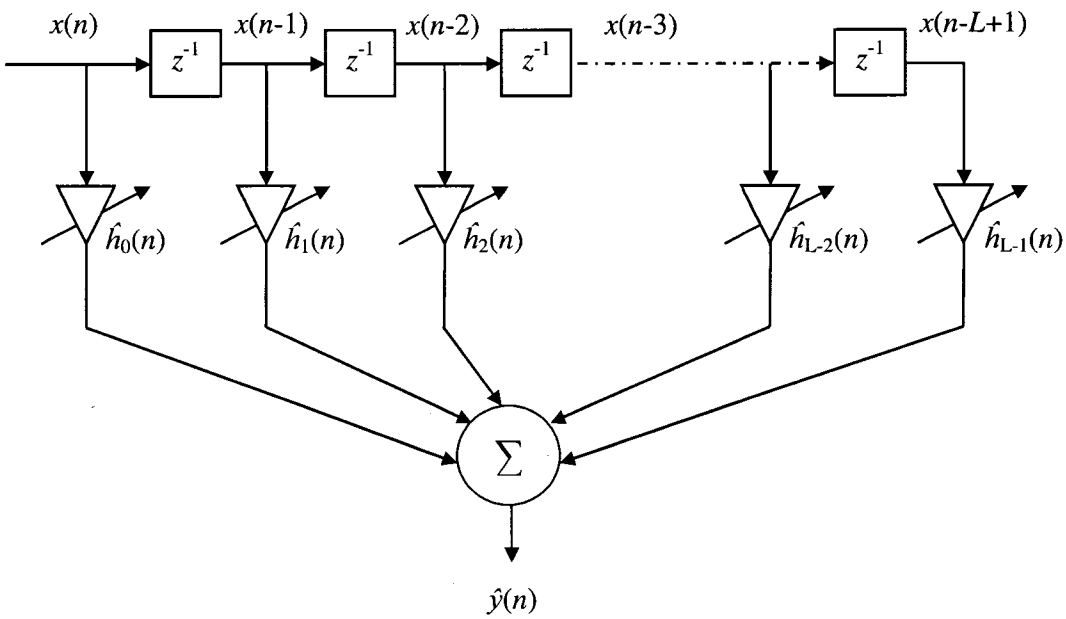

Figure 2-2: FIR Filter Structure.

2.3.2 Adaptive Algorithms

For a time varying system that a filter tries to model, the filter must adapt to the system variations. This is done by adjusting the filter weights using an adaptive algorithm. A goal for the adaptive algorithm is therefore to achieve an optimal performance level specified by a cost function. The cost function can be defined based on some measure of the estimation error defined as the difference between the desired and estimated responses. There are many different adaptive algorithms with different properties. In selecting an adaptive algorithm, one should consider the following characteristics:

- *Misadjustment* measures how much the adaptive filter differs from the true solution. A low misadjustment is desired.

- *Convergence rate* measures how fast the adaptive filter converges to a certain solution. A fast convergence rate is generally desired.

- *Tracking* concerns how the algorithm responds to changes in the desired system. A fast tracking is generally desired.

- *Computational complexity* measures the number of arithmetic operations and memory storage required to implement the algorithm. A low computational complexity is desired.

- *Robustness* concerns how sensitive the adaptive filter is to disturbances such as noise.

In the following, some adaptive algorithms based on the two main families are presented. The two are different mainly in the cost function. The first family is based on the steepest decent method, which defines a cost function as the mean square of estimation error. In the other family, the Least Squares (LS) method defines a cost function as the cumulative squared error. In addition, approaches used by adaptive algorithms to reduce computational complexity are briefly reviewed.

2.3.2.1 Steepest Descent Algorithms

The famous least mean square (LMS) algorithm, developed by Widrow and Hoff in the 1960 [HAY96], is an important member of the steepest descent family. The steepest descent method is an iterative optimization technique that seeks the minimum value of the cost function. The cost function, denoted as $J_{MSE}(n)$, of the steepest descent method is defined as the mean-square value of the estimation error, $e(n) = d(n) - \underline{\hat{\mathbf{h}}}(n)^T \underline{\mathbf{x}}(n)$. The cost function is therefore equal to $J_{MSE}(n) = \mathrm{E}\{e^2(n)\}$, where $\mathrm{E}\{\bullet\}$ denotes the statistical expectation operator. It can be seen from the above equations that the cost function is a quadratic form of the filter weights with a unique minimum value. The solution can be found by taking the partial derivatives with respect to the individual weights to find the gradient of the cost function. The minimum value can be found where the gradient is equal to zero.

$$\nabla J_{MSE}(n) = \frac{\partial_{MSE}(n)}{\partial \underline{\hat{\mathbf{h}}}(n)} = -2\mathrm{E}\{e(n)\underline{\mathbf{x}}(n)\} = 0. \tag{2.5}$$

The calculation of the optimum would require advance knowledge of the input statistics and the response of the modeled system. Since this information is not available, an iterative technique is used to search for the minimum solution. For each iteration, the steepest descent method updates the filter weights in the direction of the negative gradient as follows.

$$\underline{\hat{\mathbf{h}}}(n+1) = \underline{\hat{\mathbf{h}}}(n) + \Delta\underline{\hat{\mathbf{h}}}(n) = \underline{\hat{\mathbf{h}}}(n) - \mu\nabla J_{MSE}(n) = \underline{\hat{\mathbf{h}}}(n) + \mu\mathrm{E}\{e(n)\underline{\mathbf{x}}(n)\}. \tag{2.6}$$

The parameter μ in equation 2.6 is referred to as the step size. The LMS algorithm can be derived directly from the steepest descent method by replacing the expectation in equation

2.6 with an instantaneous value. The weight update of the LMS algorithm is calculated at each iteration as follows.

$$\underline{\hat{\mathbf{h}}}(n+1) = \underline{\hat{\mathbf{h}}}(n) + \mu e(n)\underline{\mathbf{x}}(n). \tag{2.7}$$

The selection of the step size for the LMS algorithm can critically affect the stability, convergence rate, and the steady state MSE of the adaptive filter. Since the LMS algorithm employs a noisy estimate of the gradient, the weight vector will not converge exactly to the minimum solution. This produces a steady state MSE that is always greater than the minimum MSE by an amount called the excess MSE. The ratio of excess MSE to the minimum MSE is defined as the misadjustment, which is directly proportional to the step size. Therefore, a small step size would produce a better steady state solution. However, it would take a long time for a filter with a small step size to converge because the convergence speed is proportional to the step size. In addition, a too large step size will cause the divergence of the adaptive filter. Furthermore, the convergence time of the LMS algorithm is approximately proportional to the eigenvalue spread in the correlation matrix of the filter input signal x(n), or the ratio of the largest to the smallest eigenvalues. This explains why the LMS algorithm converges very slowly for speech input because of its large eigenvalue spread. In general, the LMS is popular for its simplicity and robustness in both concept and implementation.

It can be seen from equation 2.7 that when the input speech signal x(n) is large, the weight update of the LMS algorithm is subject to the gradient noise amplification effect. The weights jump around with large instantaneous offsets, effectively undoing any fine-tuning

achieved when the input speech signal is smaller. This problem can be solved with the normalized LMS (NLMS) algorithm. The NLMS algorithm uses the following equation to update its weights.

$$\hat{\underline{\mathbf{h}}}(n+1) = \hat{\underline{\mathbf{h}}}(n) + \frac{\mu e(n)\underline{\mathbf{x}}(n)}{\underline{\mathbf{x}}^{\mathrm{T}}(n)\underline{\mathbf{x}}(n) + \delta}. \tag{2.8}$$

The regulation parameter δ is used to prevent the second term in equation 2.8 from getting very large when the input energy is very small. In addition, the necessary and sufficient condition for the convergence of the NLMS algorithm is $0 < \mu < 2$.

2.3.2.2 Least Squares Algorithms

The LS family of adaptive algorithms differs from the steepest descent method in the cost function used to update its filter weights. The cost function to be minimized is defined as the time-averaged squares of the estimation error as follows.

$$J_{\mathrm{LS}}(n) = \sum_{i=0}^{n} \beta^{n-i} \mathrm{e}^2(i). \tag{2.9}$$

The exponential weighting factor is defined as $0 < \beta < 1$, which emphasizes recent estimation errors and exponentially forgets the past ones. The LS algorithms process the input data acquired up to time n for the optimal weight vector on the n^{th} iteration. Therefore, the LS method is deterministic, unlike the heuristic method of the steepest descent family. From the principle of orthogonality based on time averaging, the optimal weights at time n can found as

$$\hat{\underline{\mathbf{h}}}(n) = \mathbf{\Phi}^{-1}(n)\underline{\mathbf{z}}(n). \tag{2.10}$$

$\mathbf{\Phi}(n)$ is defined as a time-averaged autocorrelation matrix of the tap inputs, and $\underline{\mathbf{z}}(n)$ is defined as a time-averaged cross-correlation vector between the tap input x(n) and the desired response d(n):

$$\begin{aligned} \mathbf{\Phi}(n) &= \sum_{i=1}^{n} \beta^{n-i} \underline{\mathbf{x}}(i)\underline{\mathbf{x}}^{*}(i) \\ \underline{\mathbf{z}}(n) &= \sum_{i=1}^{n} \beta^{n-i} \underline{\mathbf{x}}(i)d^{*}(i) \end{aligned} \tag{2.11}$$

The major advantage of the LS method is that its convergence rate does not depend on the eigenvalue spread of the input autocorrelation matrix. In addition, its steady state MSE can approach the minimum MSE for the forgetting factor $\beta = 1$.

It can be seen that direct computation of the matrix inversion in equation 2.10 is computationally costly. To reduce the computational complexity, the recursive least squares (RLS) algorithm uses the matrix inversion lemma to derive a recursive form for inverting the autocorrelation matrix $\mathbf{\Phi}(n)$. The RLS algorithm typically converges faster than the LMS or the NLMS algorithm but it is still too computationally expensive for application with high-order filter, i.e. large L. To reduce the computational complexity, several fast versions of the RLS algorithm are proposed such as the fast Kalman and the fast transversal filter algorithms. However, these algorithms suffer from stability problems, especially in fixed-point implementations.

2.3.2.3 Affine Projection Algorithm

The affine projection (AP) algorithm provides an intermediate solution in terms of complexity and performance between the NLMS and the RLS algorithms. The AP algorithm updates its weights in the directions that are orthogonal to the last P input vectors, and thus decorrelates the input signal. The convergence rate is therefore improved for speech signal that is highly correlated. The equations for the AP algorithm are shown in the followings.

$$\begin{aligned}
\hat{\underline{\mathbf{h}}}(n+1) &= \hat{\underline{\mathbf{h}}}(n) + \mu \mathbf{X}(n)\underline{\varepsilon}(n) \\
\underline{\varepsilon}(n) &= \left[\mathbf{X}^{\mathrm{T}}(n)\mathbf{X}(n) + \delta \mathbf{I}\right]^{-1} \underline{\mathbf{e}}(n). \\
\underline{\mathbf{e}}(n) &= \underline{\mathbf{d}}(n) - \mathbf{X}^{\mathrm{T}}(n)\hat{\underline{\mathbf{h}}}(n)
\end{aligned} \tag{2.12}$$

The L-by-P input matrix is defined as $\mathbf{X}(n) = \left[\underline{\mathbf{x}}(n) \quad \underline{\mathbf{x}}(n-1) \quad \cdots \quad \underline{\mathbf{x}}(n-P+1)\right]$, the desired response $\underline{\mathbf{d}}(n) = \left[d(n) \quad d(n-1) \quad \cdots \quad d(n-P+1)\right]^{\mathrm{T}}$ is a P-by-1 vector of the previous samples, and the regularization parameter δ keeps the matrix from becoming ill-conditioned. It can be seen that the AP algorithm reduces to the NLMS algorithm when $P = 1$.

2.3.2.4 Block Processing Approach

When modeling a system with a very long impulse response, one way to reduce the computational complexity is to use block processing. A block-based adaptive algorithm processes data in sequential blocks instead of one sample at a time. In doing so, the adaptive algorithm collects B input and output samples before computing new filter weights. A main drawback with block processing algorithm is the delay from when the data is collected to when the new filter weights are computed. However, if the delay is allowed, the algorithm

offers a reduction in computational complexity. In addition, block processing can be performed in the frequency domain. Thus, the algorithm can make use of more efficient computational methods such as fast convolution using the fast Fourier transform.

2.3.2.5 Frequency Domain Approach

Frequency domain adaptive filtering has two big advantages over the time domain one. First, it can reduce computational complexity because of the block processing nature. In addition, a convolution in the time domain can be replaced by multiplications in the frequency domain, resulting in a big saving in computations. Second, the discrete Fourier transform, which transforms data into the frequency domain, produces spectral signals that are approximately uncorrelated. As a result, faster convergence of filter weights can be obtained because different step sizes can be used for the different filter weights to be adapted. Again, the delay in signal is a cost for the reduction in computational complexity.

2.3.2.6 Sub-band Approach

A sub-band adaptive filter can alleviate the problems with the required filter length and the convergence speed. The approach is to split both the input and the desired signals into N frequency bands using analysis filter-banks, of which the outputs are down-sampled by a factor of R. Adaptive filtering is performed separately in these bands to generate sub-band error signals. The sub-band errors are up-sampled by the same factor R, passed through a set of synthesis filter-bank, and then combined to give the full band error. The main advantages of sub-band filtering are two-fold: a reduction in filter length and a gain in convergence speed. Assuming that all the sub-bands are down-sampled by the same factor, the length of

the filter for each down-sampled sub-band is $1/R$ of the full band filter. The computational complexity of an adaptive filter such as the LMS-type depends directly on the filter length and the sampling rate. For each sub-band, the computational complexity is $1/R^2$. Hence, the overall gain in computational efficiency for a sub-band system is R^2/N of the full band system. The convergence speed depends on both the filter length and the eigenvalue spread of the signal. For a shorter sub-band filter, the convergence speed increases because of the increased step size allowed. More importantly, since the sub-band spectrum of the input signal is likely flatter than that of the full band signal, the spread of its eigenvalues decreases. The convergence speed of the adaptive filter therefore increases. However, the sub-band adaptive filtering inherits the group delay of the filter-banks. In addition, undesired aliasing effects may occur due to non-ideal band-pass filtering at the edges of sub-bands.

Chapter 3
Fundamentals of Psychoacoustics

Psychoacoustics is the study of how sound is subjectively perceived by a human listener. Alternatively, it can be described as the science that examines the relationship between sounds and their psychological effects. The psychology and acoustics fields form the basis of psychoacoustics. One of the most significant psychoacoustic phenomena utilized in speech processing is the masking property. Masking reflects the limited frequency selectivity of the human auditory system. An approximate measure of the amount of masking can be obtained by evaluating a masking threshold. It indicates the sound pressure level at which a test sound is just audible in the presence of a masking sound or masker. This chapter gives a brief overview of the human auditory system, and reviews some important principles of psychoacoustics used in speech processing such as the absolute threshold of hearing, the critical bands, and the masking properties. These principles are basics for the development of the proposed DTD algorithm.

3.1 The Human Auditory System

The human auditory system (HAS) can be divided into two main parts. The peripheral part or the human ear, shown in Figure 4.1 of [SHA00], consists of the outer ear, the middle ear, and the inner ear. In addition, the central part consists of the auditory nerve fibers and the auditory sensation area of the brain. The purpose of the HAS is to capture sound waves and convert theses into mechanical vibrations. The HAS then converts the processed mechanical

oscillation into nerve impulses. Finally, the nerve fibers convey the information to the auditory sensation of the brain in which it is perceived as sounds. In the followings, the functions of the peripheral auditory system are reviewed.

3.1.1 The Outer Ear

The outer ear consists of the pinna (the visible ear lobe), the meatus (ear canal), and the tympanic membrane (eardrum). The sound wave enters the ear via the pinna and propagates through the meatus. The difference in air pressure makes the eardrum vibrate. The pinna and the outer ear influence strongly on the incoming sounds, process the inlet angle of sound, and thus provide cues to the localization of the sound. For high frequencies, the ear is less sensitive to sounds coming from behind a listener. That is because of the shadowing effect caused by the pinna, which attenuates the sounds coming from behind. The meatus, which is structured as an open tube with approximate length of 2 cm behaves like a resonator with centre frequency at 4 kHz, so it principally accounts for the hearing around this range. The eardrum is an enclosed membrane whose function is to convert the acoustic sound pressure to mechanical vibration to the middle ear. In addition to the modification of incoming sound by the outer ear, the head and shoulder contribute to the effect of shadowing and reflecting.

3.1.2 The Middle Ear

The middle ear is an air-filled chamber that consists of the ossicles. The ossicles consist of three small bones namely the malleus (hammer), the incus (anvil), and the stapes (stirrup). The malleus attaches to the eardrum of the outer ear, and the stapes touches the oval window

of the inner ear. The middle ear primarily functions as an impedance matching between the air medium of the outer ear and the fluid of the inner ear to ensure efficient transfer of acoustical energy, and to avoid large reflections. The impedance matching is performed by the lever system constructed by the ossicles and the ratio of the area of the eardrum to that of the oval window. The other function of the middle ear is the amplitude limiting. This is done by the tiny inner ear muscles attached to the ossicles. These muscles contract and therefore limit the transmission of high-intense sound through the ossicles. This operation is also called the middle ear reflex, which helps to protect the vulnerable structure of the inner ear at low frequencies, and decreases the audibility of self-generated sounds.

3.1.3 The Inner Ear

The inner ear, the most complicated part of the ear, consists of the cochlea and the semicircular canals of the vestibula. The vestibula helps balancing the body and has no apparent role in the hearing process. The cochlea transforms mechanical oscillations at the oval window into electrical pulses, which is detected by the auditory nerves. The cochlea is a spiral-shaped tube, filled with nearly incompressible lymphatic fluid, and surrounded by hard bones. The region close to the oval window is recognized as the base, whereas the inner tip of the coil is known as the apex. The cochlea has two membranes, the basilar membrane (BM) and the Reissner's membrane, which divide along the cochlea into three channels, namely the scala vestibule, the scala media, and the scala tympani. The cochlea's most important role is to analyze the frequency content of the incoming sounds. The mechanical vibrations move the oval window; result in the longitudinal displacement of the fluid inside

the cochlea and the vertical displacement of the BM. The maximum displacement of the BM depends on the frequency of the stimulus, with the low frequencies (stronger vibrations) near the apex and high frequencies near the oval window. The characteristic frequency (CF) of a stimulus is the frequency that causes the maximum response at a given point of the BM. However, it gets more complicated for the stimulus with two frequency components sufficiently close to each other. In this case, the BM fails to discriminate between them, and only the one with stronger intensity can be observed by the BM. This phenomenon can certainly explain the masking effects in the HAS.

The vibrations of the BM are picked up by the sensory hair cells of the organ of Corti, which lie in the BM. About 30,000 sensory hair cells attach to the auditory nerves. There are two different kinds of the sensory hair cells, the inner hair cell (IHC) and the outer hair cells (OHC), with their own functions. About 90% of afferent (ascending) nerve fibers that carry information from the cochlea to the brain terminate at the IHC. The OHC receives messages from the brain through several efferent (descending) nerve fibers. These messages are most likely used for active processes that improve the frequency selectivity of the auditory system.

3.2 Absolute Threshold of Hearing

The absolute threshold of hearing (ATH) or the threshold in quiet characterizes the amount of energy in dB SPL for a pure tone to be audible by a listener in the absence of any interfering sounds, i.e. in quiet environment. The ATH is unique from person to person and furthermore changes with ages. Figure 2 of [PAI00] shows a typical ATH curve. The curve is usually obtained by averaging many individual absolute thresholds measured in many

subjects. It can be seen that the ATH is a function of frequency, having a minimum of around 0 dB SPL between the frequency ranges of 2-5 kHz. The minimum indicates the region with high sensitivity of hearing and the high susceptibility to hearing impairment. In signal processing, the ATH is often used in combination with masking thresholds to calculate which spectral components are inaudible and may thus be ignored in processing; any part of a spectrum which has amplitude below the ATH may be removed from the process without any audible change to the signal. Based on experimental test results by many researchers over the years [TER79 and 82], the ATH can usually be approximated by the following equation.

$$T_q(f) = 3.64\left(\frac{f}{1000}\right)^{-0.8} - 6.5e^{-0.6\left(\frac{f}{1000}-3.3\right)^2} + 10^{-3}\left(\frac{f}{1000}\right)^4 \quad (dB \quad SPL). \qquad (3.1)$$

The frequency f is expressed in Hz. The unit dB SPL is the dB level relative to the reference intensity level of 20 μPa. The ATH expression represents the combined effects of the outer and middle ear frequency responses along with the internal noise of the inner ear.

3.3 Critical Bands and Bark Scale

A frequency-to-place conversion takes place in the cochlea along the BM that affects the frequency selectivity of the HAS. The frequency selectivity is crucial to perception as it determines the ability of the HAS to resolve frequency components. Each point on the BM responds to only a certain range of frequencies. This behavior leads to the theory that different points on the BM correspond to auditory filters with different centre frequencies. The concept of critical bands (CB) is therefore introduced to define frequency ranges, in which changes in stimuli can greatly affect perception, i.e. for sounds in the same critical

band, the one with the highest energy level would dominate the perception. Fletcher first introduced the CB concepts [FLE40]; and suggested that the HAS perceives sounds non-uniformly in frequency, and behaves like a bank of bandpass filters with bandwidth corresponding to the critical bandwidth. According to [TOB70], Scharf measured the bandwidth of CB as a function of their centre frequency. It was suggested, while representing the inner ear as a discrete set of non-overlapping auditory filters, that 25 CB were sufficient to represent the audible frequency range of the HAS. The bandwidths of the CB are listed in the Appendix A with centre frequencies spanning from 0 to 19 kHz. It was evident from Scharf's results that the auditory bandwidths are larger at lower centre frequencies than at higher; and that CB are roughly constant below 500 Hz, and then increase approximately linearly with higher frequencies. In addition, the majority of CB lies below 5 kHz. That means the ability of the HAS to resolve sound components is superior at low frequencies than at high frequencies.

The CB is commonly expressed in the Bark scale unit, where an increment of one Bark corresponds to one critical band. Figure 3-1 illustrates the relationship between the linear frequency in kHz and the critical band in Bark. [ZWI99] suggested an analytical expression to characterize the relationship of the critical band in Bark and the frequency in Hz.

$$CB(f) = 13\arctan(0.00076f) + 3.5\arctan\left[\left(\frac{f}{7500}\right)^2\right] \quad \text{(Bark)}. \qquad (3.2)$$

The bandwidth of each critical band as a function of its centre frequency is approximately given in the following equation.

$$BW(f) = 25 + 75\left(1 + 1400 f^2\right)^{0.69} \quad \text{(Hz)}. \tag{3.3}$$

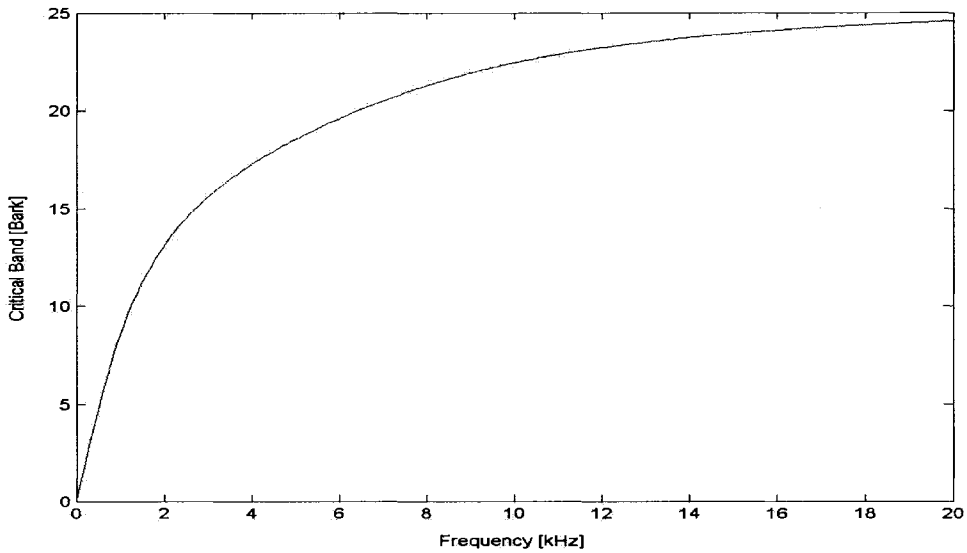

Figure 3-1: Mapping Linear Frequency (kHz) into Critical Bands (Bark).

3.4 Auditory Masking Effects

Masking effects can be experienced in everyday life. For example, it can be very difficult or even impossible to listen to someone if loud music is playing in the background. This phenomenon is called masking. Auditory masking is the process by which the perception of one sound is suppressed by another. The masking phenomenon is charaterized by an increase of the ATH or the audibility threshold making an otherwise perceptible sound inaudible because of the presence of a stronger sound. The stronger sound is commonly referred to as the masker and the sound being masked is referred to as the maskee. Masking effects are generally classified into two types: the simultaneous and temporal masking. Simultaneous

masking occurs when the masker and maskee are presented concurrently. This effect is also called the frequency masking. On the other hand, the temporal masking, also known as non-simutaneous masking, occurs when the masker and maskee have a temporal offset with respect to each other. These masking effects are discussed in more detail below.

3.4.1 Temporal Masking

The temporal masking occurs when the masker and maskee are presented consecutively in non-overlapping time. There are two different types of temporal masking: backward and forward masking. In the backward masking, also termed as prestimulatory or pre-masking, the maskee emerges before the masker. In the forward masking, also termed as poststimulatory or post-masking, the maskee appears after the masker. Figure 10 of [PAI00] illustrates an example of the temporal masking. It shows the different temporal effects and their relevant time scales for the masker of 200 ms duration. It can be seen that the effect of backward masking is minor, and that the forward masking takes effect much longer than the backward masking. It was also shown in [ZWI99] that the forward masking depends strongly on the duration of the masker with a shorter duration resulting in the masking threshold to drop faster. Apart from the masker duration, the level and the bandwidth of the masker and the frequencies of both the maskee and masker can influence the forward masking effect.

3.4.2 Simultaneous Masking

The simultaneous masking effect occurs when the masker and the maskee are present concurrently to an auditory system. The effect can be explained by the fact that the basilar

membrane of the ear, a frequency analyzer, has limits of resolution so it fails to discriminate the two sounds. Only the stronger sound therefore is perceivable. Figures 8 and 9 of [PAI00] illustrate the simultaneous masking effect with the masker and the maskee shown in frequency domain. In addition, it shows an overall masking threshold under which frequency components would be masked and deemed inaudible.

The amount of masking depends on the levels, frequencies, and duration of the masker and maskee. There are two major types of simultaneous masking: the noise-masking-tone (NMT) and the tone-masking-noise (TMN). In the NMT scenario, a narrow-band noise can mask a tone whose frequency is in the frequency band of the noise, if the intensity of the masked tone is below a predictable threshold directly related to the intensity and the centre frequency of the masking noise. In the TMN case, a pure tone can mask a narrow-band noise, if the frequency of the tone is in the frequency band of the noise, and the noise spectrum is below a predictable threshold directly related to the strength and the centre frequency of the masking tone. In addition, the NMT and TMN masking effects are not band-limited to within the boundaries of a single critical band; the inter-band masking also occurs. This effect, also known as the spread of masking, means that a masker within a critical band can have some predictable effect on other critical bands.

Chapter 4
A Survey of Double Talk Detection Algorithms

Adaptive EC requires a control component to prevent the adaptive filter from deteriorating drastically during the DT situation, in which active speeches are present simultaneously at both ends of a communication link. There are commonly two fundamentally different strategies in dealing with the DT problem. One is to detect the DT situation directly, and then freeze the adaptation of the EC filter once a DT condition is detected. The other is to detect and track for the echo path changes (EPC). In the later case, assuming that the adaptive filter is in a convergent state, its weights are frozen, and will be enabled for adaptation only when an EPC event occurs. In this thesis, the emphasis is more on the former approach that is presented in the next chapter. This chapter presents common basics of a generic DTD scheme; then classifies some basic techniques of DTD, and reviews briefly some existing DTD algorithms. Note that the same notations of signals as in Figure 2-1 are used in the review of this chapter.

4.1 Basis of a Generic DTD Scheme

The common basic operation in DTD algorithms involves the computation of a decision variable ξ from available signals, and then the comparison of the decision variable with a preset constant threshold T. Depending on whether the decision variable is above or below the threshold, a decision is made on whether a DT condition is present. Once a DT condition is declared, the adaptation process of the filter is halted for some hold time T_{hold}. In addition,

the DT decision can trigger other control components such as the variable loss and the non-linear processor in the EC system to function. After the hold time, the adaptation process resumes until the next DT situation arises. It was observed in [GAY00] that a hold time is necessary to suppress detection dropouts because of the noisy behavior of the decision variable.

A DT detector may struggle with a change in the echo path. It can falsely detect the change as DT situation. It is desirable in this case for the adaptive filter to adapt quickly to the change, not to turn off the adaptation process by freezing its weights because of false DTD. Thus, an effective DT detector should be insensitive to the echo path variations. In addition, a robust DT detector should not detect the near-end background noise as DT. Furthermore, it is desirable for a DT detector to make a decision with a minimal delay because a slow response to DT can seriously affect the convergence of the adaptive filter.

There are two types of errors associated with a DT detector: miss detection and false detection. Depending on the current state of the adaptive filter, the errors can greatly reduce the overall performance of the echo canceller. When the adaptive filter is in the convergent state, a false detection has no effect on the cancellation of echo. However, when false detection occurs during the non-convergent state of the adaptive filter, the filter would remain in this inferior state for the period of the false alarm. As a result, echo is not cancelled properly and passed through to the other end. In addition, when a true DT condition is missed during the convergent state of the adaptive filter, the filter would diverge because the near-

end speech signal causes an increase in the residual error signal. Consequently, distorted signal and echo can pass through to the other end.

Most DTD algorithms share the common basics as presented above; they differ, however, in the techniques used to compute the decision variable. In the following sections, a survey of several DTD algorithms existing in the literature is presented. The DTD algorithms are classified into level comparison techniques and cross correlation techniques. In addition, DTD algorithms, which have techniques other than the two mentioned above, and which are combinations of the two, are grouped into the third category.

4.2 Survey of DTD Algorithms

4.2.1 Energy Level

A simple approach in this category is the well-known Geigel algorithm shown in [DUT78]. The Geigel algorithm compares the magnitude of the near-end input signal d(n) with the maximum magnitude of the L most recent samples of the far-end speech signal x(n), where L is the length of the adaptive filter. The L past samples are used because of the possible delay of x(n) through the echo path. The Geigel assumes that the echo path typically dampens x(n), and therefore the magnitude of d(n). Therefore d(n) containing only the echo y(n) in ST case would be smaller than d(n) containing both the echo y(n) and the near-end speech signal v(n) in DT case. The Geigel algorithm computes its decision variable as follows.

$$\xi = \frac{|d(n)|}{\max\{x(n) \quad x(n-1) \quad \cdots \quad x(n-L+1)\}} > T \,. \tag{4.1}$$

For EEC applications, the threshold T for the Geigel algorithm is set to 0.5 because the hybrid attenuation is assumed no less than 6dB. This assumption would not be valid for the AEC as the echo path is time varying and the echo path attenuation is much lower The performance of the Geigel algorithm would therefore suffer because the acoustic echo path is changing and the echo level is much higher than of the near-end speech.

The authors in [GAN00] combined the Geigel algorithm with a novel adaptation strategy to improve performance of EEC considerably during DT situation. The proposal employs a method of robust statistics to make the proportionate NLMS adaptive algorithm more robust to DT conditions. The robust statistics involves the computation of a scaled nonlinearity, which operates in the same manner as varying the step size μ. What differentiates the proposal is that traditional variable step size method tries to detect DT and then take action, while the robust technique uses the signal before DT in order to prepare for it. As a result, the robust technique is faster and more efficient.

A more sophisticated modification to the Geigel algorithm, suggested in [DYB04], uses short-term power estimates rather than the instantaneous values as follows.

$$\xi = \frac{P_d(n)}{P_x(n)} > T . \tag{4.2}$$

An example of the short-term power estimate of a signal s(n) at time n using a sliding window of W samples is shown in the following.

$$P_s(n) = \sum_{i=0}^{W-1} |s(n-i)| . \tag{4.3}$$

In the above case, it is not necessary to make the comparison for every sample, instead for an interval of W samples. This would reduce the computation complexity but at a cost of detection latency and missed detection. Other similar modifications to the Geigel algorithm, for example that proposed in [CHU96], uses single-pole IIR filters or arithmetic average using FIR filters for estimating the short-term average energy or power of the signal d(n) and x(n).

The authors in [PAR90] proposed a method to estimate the echo return loss (ERL) for determining a DT condition in a real implementation of an acoustic echo canceller. A DT condition is declared when the following condition is met:

$$\xi = ERL(n) = 10\log_{10}\left[\frac{\hat{\mathrm{E}}\left[x^2(n)\right]}{\hat{\mathrm{E}}\left[d^2(n)\right]}\right] < ERL_{\min}, \tag{4.4}$$

where the short-term average power $\hat{\mathrm{E}}[\bullet]$ is calculated using the first-order recursive method, i.e. single-pole IIR filter. The $ERL_{\min}$ is the minimum threshold value for echo loss, and generally set between -24 dB to -30 dB, which is 0 to 6 dB larger than the expected performance of AEC. It was claimed that the DTD algorithm gives a stable and acceptable performance with AEC.

The authors in [ENE98] suggested a more advanced energy-based DTD algorithm in determining the decision variable as follows.

$$\xi = \frac{\mathrm{E}\left[x^2(n)\right]\mathrm{E}\left[e^2(n)\right]}{\mathrm{E}\left[x^2(n)\right] + \mathrm{E}\left[\hat{y}^2(n)\right]} > T. \tag{4.5}$$

In the absence of DT, the value of the decision variable ξ would be smaller than one. When DT starts to occur, the decision variable would rise, so the threshold T can be selected at about one.

A novel energy-based DTD algorithm for AEC application, proposed by authors in [DIN04], exploits the difference in bandwidths of the narrow-band echo signal y(n) (in the telephone band of 300-3400 Hz) and the wideband DT voice v(n). By monitoring the energy in the out-of-telephone-band of the microphone signal, a DT condition can be effectively detected.

4.2.2 Cross Correlation

The authors in [YE91] first exploited the orthogonality principle in dealing with DT situation in EC. The basic idea of the principle is shown in the following equations.

$$\begin{aligned} \mathrm{E}\{e(n)\underline{\mathrm{x}}(n)\} &= \mathrm{E}\left\{\left(d(n)-\hat{y}(n)\right)\underline{\mathrm{x}}(n)\right\} \\ &= \mathrm{E}\left\{\left(y(n)+v(n)+w(n)-\hat{y}(n)\right)\underline{\mathrm{x}}(n)\right\} \\ &= \mathrm{E}\left\{\left(y(n)-\hat{y}(n)\right)\underline{\mathrm{x}}(n)\right\}+\mathrm{E}\{v(n)\underline{\mathrm{x}}(n)\}+\mathrm{E}\{w(n)\underline{\mathrm{x}}(n)\} \\ &\approx \mathrm{E}\left\{\left(y(n)-\hat{y}(n)\right)\underline{\mathrm{x}}(n)\right\} \end{aligned} \quad (4.6)$$

Basically, when the adaptive filter has converged to its optimal solution, the orthogonality between the far-end signal vector $\underline{\mathbf{x}}$(n) and the residual error e(n) is satisfied, i.e. $\mathrm{E}\{e(n)\underline{\mathrm{x}}(n)\} \approx \mathrm{E}\left\{\left(y(n)-\hat{y}(n)\right)\underline{\mathrm{x}}(n)\right\} \approx 0$, where $\mathrm{E}\{\bullet\}$ denotes the statistical expectation. In a DT situation, the local input signal d(n) and therefore the residual echo e(n) get larger

because of the presence of the near-end signal v(n). However, as long as the signal v(n) and the noise w(n) are uncorrelated with the signal x(n), the orthogonality still holds. However, the orthogonality cannot be satisfied anymore with an echo path changes. In general, the algorithm does not detect DT condition explicitly. Instead, it decides whether the adaptive filter has converged or not. If the adaptive filter has converged, the adaptation is stopped to protect the filter from being disturbed by DT interference. Otherwise, the filter has not converged or the echo path has changed, and the adaptive filter will keep adapting.

Typical examples of the cross-correlation techniques for DTD are presented in [BENE00] and [CHO99]. The cross-correlation vector between the far-end signal vector $\underline{\mathbf{x}}$(n) and the input signal d(n) is used for computing the decision variable. The cross correlation is defined as follows

$$\underline{\mathbf{c}}_{xd} = \frac{\mathrm{E}\{\underline{\mathbf{x}}(n)d(n)\}}{\sqrt{\mathrm{E}\{x^2(n)\}\mathrm{E}\{d^2(n)\}}} = [c_{xd}(0) \quad c_{xd}(1) \quad \cdots \quad c_{xd}(L-1)]^{\mathrm{T}}. \tag{4.7}$$

Where the $c_{xd}(i)$ is the cross-correlation coefficient between x(n-i) and d(n). The decision variable is calculated as follows.

$$\xi = \max|\underline{\mathbf{c}}_{xd}| < T. \tag{4.8}$$

With some preset threshold value, the DT can be detected when the maximum magnitude of the cross-correlation coefficients is smaller than the threshold. Since the cross-correlation value depends on the statistics of the signals and the echo path, the threshold value T would vary from one experiment to another. To resolve this problem, the authors proposed a

normalized cross-correlation (NCC) algorithm. The idea is to normalize the cross-correlation vector so that the decision variable is equal to one when the near-end signal $v(n)$ is zero and less than one when $v(n)$ is not, i.e. DT situation. Without going into details of the derivation process, the decision variable for the NCC algorithm is defined as follows.

$$\xi = \frac{\sqrt{\underline{\mathbf{h}}^{\mathrm{T}}\mathbf{R}_x\underline{\mathbf{h}}}}{\sqrt{\underline{\mathbf{h}}^{\mathrm{T}}\mathbf{R}_x\underline{\mathbf{h}} + \sigma_v^2(n)}} = \sqrt{\underline{\mathbf{r}}_{xd}^{\mathrm{T}}\left(\sigma_d^2\mathbf{R}_x\right)^{-1}\underline{\mathbf{r}}_{xd}^{\mathrm{T}}} \,. \tag{4.9}$$

Where σ_d^2 is the variance of $d(n)$, $\mathbf{R}_x$ is the autocorrelation matrix of the vector $\underline{\mathbf{x}}(n)$, and $\underline{\mathbf{r}}_{xd}$ is the cross-correlation vector between vector $\underline{\mathbf{x}}(n)$ and signal $d(n)$. When the decision variable $\xi < T$, DT is declared; when $\xi \geq T$, DT is not declared. In practice, the computation of the decision variable based on the above equation is very computationally expensive. One way to simplify the equation is to assume that $\underline{\hat{\mathbf{h}}} \approx \underline{\mathbf{h}} = \mathbf{R}_x^{-1}\underline{\mathbf{r}}_{xd}$ when the adaptive filter has converged. Therefore, the decision variable of the NCC method can be written as follows.

$$\xi = \sqrt{\underline{\mathbf{r}}_{xd}^{\mathrm{T}}\,\underline{\hat{\mathbf{h}}}\,\sigma_d^{-2}} \,. \tag{4.10}$$

Where σ_d^2 and $\underline{\mathbf{r}}_{xd}$ can be practically estimated by a time-averaging method using a sliding window of W samples. The NCC algorithm in general shows better performance for AEC applications than the level-based techniques because it has very low sensitivity to the attenuation of the echo path. However, the technique involves matrix/vector operations, which are computationally expensive. In addition, the NCC technique in general has a longer response time to the onset of DT. That is because the technique needs to process more data to arrive at a DT decision.

There are other variations of the above NCC technique, in which the authors try to improve a certain aspect of the original NCC algorithm. The authors in [GOR05] proposed a novel low-complexity version of the NCC algorithm for IP-enabled telephone applications employing LPC-based speech coders. The algorithm, in computing the decision variable, can obtain a decorrelated input signal and decorrelation filter coefficients directly from the speech decoder. As a result, the proposed NCC algorithm still performs as well with a reduction in computational complexity.

The authors in [PAR02] proposed a DTD algorithm that employs two cross-correlation coefficients for hands-free applications. The first coefficient measures the cross correlation between the near-end input $d(n)$ and the echo estimate $\hat{y}(n)$. The second coefficient measures the cross correlation between the signal $d(n)$ and the residual error $e(n)$. When the adaptive filter is converged and the near-end speech $v(n)$ does not exist, the first coefficient is close to 1 because $d(n)$ is highly correlated to $\hat{y}(n)$. On the contrary, the second coefficient is close to 0 in this case because $e(n)$ is uncorrelated to $d(n)$. The algorithm uses these properties to detect DT condition. The decision flow of DTD consists of two stages. In the first stage, when the first coefficient is less than a threshold, a ST condition is declared. Otherwise, when the second coefficient is greater than another threshold, a DT condition is declared. It was claimed that the implementation of the algorithm performs detection accurately and is robust under adverse conditions.

The authors in [SUG05] proposed a noise-robust version of the original NCC DTD algorithm. It was observed that in deriving an expression for computing the decision variable,

the background noise w(n) in the near-end input d(n) was neglected and assumed insignificant. A method is proposed to estimate the noise offset influenced by the noise signal in the original form of the decision variable. By incorporating the noise offset in the decision variable, it was shown that the proposed algorithm gives superior detection performance in comparison with the conventional NCC algorithm.

In [GAN01], the authors derived a frequency-domain computation scheme of the original NCC DTD algorithm for multi-channel EC. The frequency-domain approach is chosen because of its desirable properties such as low computational complexity and inherent stability. The same authors, in later work [BEN03], proposed a multidelay version of the frequency domain NCC DTD algorithm, which combines efficiently with the multidelay block frequency domain adaptive filter.

The authors in [JUN02] proposed a DT detector that computes the decision variable based on echo path estimation. The actual echo path is estimated using the far-end speech x(n) and the near-end input signal d(n) as follows.

$$\hat{\underline{\mathbf{h}}}(n) = \mathbf{R}_x^{-1}\underline{\mathbf{r}}_{xd}(n). \tag{4.11}$$

The proposed DTD algorithm consists of two decision stages. In the first stage, the ST condition is differentiated from the DT condition and EPC based on the gain of the echo path estimate. The second stage makes distinction between DT condition and EPC based on the derivative of the gain. The normalized decision variable, which is the gain of the echo path, is time varying and computed as follows.

$$Gain \triangleq \xi(n) = \left\|\hat{\underline{\mathbf{h}}}(n)\right\|^2 \frac{L\sigma_x^2}{\sigma_d^2} = \frac{\sigma_x^2}{\sigma_d^2}\sum_{i=0}^{L-1}\hat{h}^2(n-i). \quad (4.12)$$

The parameter L is the length of the adaptive filter. In the first stage, when $\xi(n)$ is less than a threshold T_1, it is not a ST situation, i.e. DT or EPC. In the second stage, when $\nabla\xi(n)$ is greater than a threshold T_2, it is a DT situation. It was claimed that the proposed DTD algorithm is effective in detecting DT, and the required decision delay is shorter than that of the conventional methods.

4.2.3 Others

This section reviews briefly other DTD techniques, which do not fall into one of the two categories discussed above. In some cases, they are hybrid approaches of both categories.

The authors in [DIN04] proposed a DTD algorithm with a low complexity, which combines the orthogonality principle between e(n) and y(n), and an energy level technique to be able to distinguish between DT situation and echo path change. The proposal DTD scheme was incorporated into commercial products for both AEC and NEC applications.

The authors in [GAN96] proposed a method to measure the similarity between the far-end speech and the near-end input signal for DTD based on their spectral shapes in full-band. The approach is to estimate the squared magnitude coherence between the two signals on a block-to-block basis. The averaged coherence value is close to one for no DT, and close to zero for DT situation. The DT decision variable is calculated on average over a few frequency intervals. The coherence technique shows very good DTD performance for NEC in

simulations, however, computing of the coherence function in full-band is still computational costly. The authors in [JIA03] proposed computational efficient versions of the coherence method by employing sub-band DT detectors instead. One approach is to select a sub-band with the best detection performance and use its local decision variable as the global decision. The other is to weigh sub-band decision variables according to the signal power of their bands to compute the global decision. The sub-band structure reduces computational complexity, and improves performance of DTD.

The authors in [DOH97] proposed a spread spectrum technique to estimate the impulse response of a room echo path. A low-level, wide-band pseudo-noise training signal is superimposed on the far-end speech signal $x(n)$. The near-end signal $d(n)$ is correlated with the training sequence to give an estimate of the room impulse response, even during DT conditions. It was claimed that the training is harmless and unaffected by the presence of the speech signal $v(n)$ because these signals are assumed uncorrelated.

The authors in [GHO00] proposed a method using a sub-band adaptive filter for modeling the echo path in AEC applications. The technique measures the angle between the two vectors of samples, the near-end input signal $d(n)$ and the echo estimate $\hat{y}(n)$ at the output of the adaptive filter. In the presence of DT situation, the angle significantly deviates from zero; otherwise, when the adaptive filter is converged, the angle is very close to zero. The decision variable is calculated using the deviation of the angle from zero. It was shown with simulations that the drop in ERLE during DT is only marginal. With a similar approach, the authors in [CRE00] proposed to measure the angle of trajectory of the filter weights to

control the variable step size of the adaptive filter. In DT situation, the trajectory direction is more random than that under the condition of EPC. The method shows good performance with speech in AEC simulations. However, it involves many optimizations of many parameters. In addition, the author of [AHG04] proposed a similar technique with very low computational complexity. It considers the near-end speech v(n) as noise corrupting the near-end input signal d(n). A filter that adapts to a noise would fluctuate dramatically. By measuring the variance of the largest value of the weights of the adaptive filter over time, DT is declared when this measure exceeds a threshold value. The performance of the proposed DTD method using an objective evaluation method shows very good overall results in comparing with other energy and correlation methods.

The authors in [OCH77] proposed a novel structure with two adaptive filters to model a room echo path. The background filter operates as in the conventional EC. The foreground filter, which is not adaptive, cancels the echo. Whenever the background filter performs better than the foreground filter based on a set of conditions, its coefficients are copied to the foreground. The set of conditions being met are:

i) The estimated power of the residual error of the background filter is smaller than that of the foreground by some threshold.

ii) The estimated power of the near-end input signal d(n), mainly echo, is greater than that of the background residual error by another threshold.

iii) The estimated power of the near-end input signal $d(n)$ is smaller than that of the far-end speech $x(n)$. This is actually a modified version of the Geigel algorithm for DTD with the threshold being set at one.

The third condition prevents the transfer during a clear DT condition. A great advantage with the structure is that it is not sensitive to echo path changes since the background filter is allowed to track changes freely and its coefficients are not copied to the foreground one until it performs better than the latter. In addition, simulations show a very good DT performance of the structure with a very low rate of miss detection. However, the structure doubles the computational and memory requirements.

The authors in [HIR03] proposed another dual-filter acoustic echo canceller that is DT resistant. An echo path filter is used for estimation of the room echo path; and a cancellation filter is used for echo cancellation. The cross correlation between the echo estimate, which is the output of the echo path filter, and the near-end input signal is used to control a variable step size of the filter. The filter coefficients are copied from the echo-path filter to the cancellation filter when the cross correlation value is small enough. It was shown that the method successfully reduces echo during DT periods. A very similar approach proposed in [LIU01] uses the NLMS adaptive algorithm and the Geigel algorithm in the dual-filter structure that could discriminate an EPC, a DT situation, and the initial convergence and adjust the step size accordingly.

4.3 Summary

This chapter presented several DTD algorithms existing in the literature and classified them based on the techniques of computing the DT decision variable. Most DTD algorithms employ at least one of the two basic techniques: energy or power level techniques and cross-correlation techniques. The level-based techniques in general have the benefit of being computationally simple, needing little memory requirement. However, these techniques do not always provide reliable performance with time varying echo path such as the acoustic environment. On the other hand, the cross-correlation techniques exhibit desirable properties for such echo path condition, mainly because of their low sensitivity to the attenuation of the echo path. However, these techniques require relatively high computation complexity and memory storage due to vector and/or vector based operations. In general, an abundance of so many different proposals addressing the DTD problem in the literature demonstrates difficulties in solving the problem. Designing and implementing a simple but efficient practical DT detector is therefore a very challenging task; and one would consider the following properties:

- Insensitive to echo path variations;
- Equal performance regardless of the convergence state of the adaptive filter;
- Quick response time; not slowing down the convergence rate of the adaptive filter;
- Sensitive to DT condition; very low probabilities of miss and false alarm;
- Low computational complexity and memory storage; and simple implementation.

Chapter 5
The Proposed DTD Algorithm Using a Psychoacoustic Auditory Model

The preceding chapters reviewed fundamental concepts of the EC and control, and of the perceptual masking properties in psychoacoustics as foundations for the proposed DTD algorithm. The necessity of a simple but efficient DTD algorithm was also identified and discussed. In this chapter, a proposed DTD algorithm using a psychoacoustic auditory model is presented and analyzed for its feasibility. Through simulations, its performance is evaluated and compared with typical DTD algorithms such as the Geigel and the NCC algorithms using real speech and an actual room impulse response.

5.1 Basic Echo Cancellation Equations

A typical configuration of an echo canceller and a DT detector for a single-channel communication system is shown in Figure 5-1. The far-end speech $x(n)$ passes through an echo path represented by $\underline{\mathbf{h}}(n)$ to produce a near-end echo $y(n)$, which is picked up at the near-end input port together with the near-end speech $v(n)$ and the background noise $w(n)$. The signal $x(n)$ also passes through an adaptive filter with coefficients or weights $\underline{\hat{\mathbf{h}}}(n)$ to produce an echo estimate $\hat{y}(n)$.

During a DT condition, the near-end input signal consists of the echo mixed with the near-end speech and the background noise.

$$d(n) = \underline{\mathbf{h}}(n)^{\mathrm{T}} \underline{\mathbf{x}}(n) + v(n) + w(n). \tag{5.1}$$

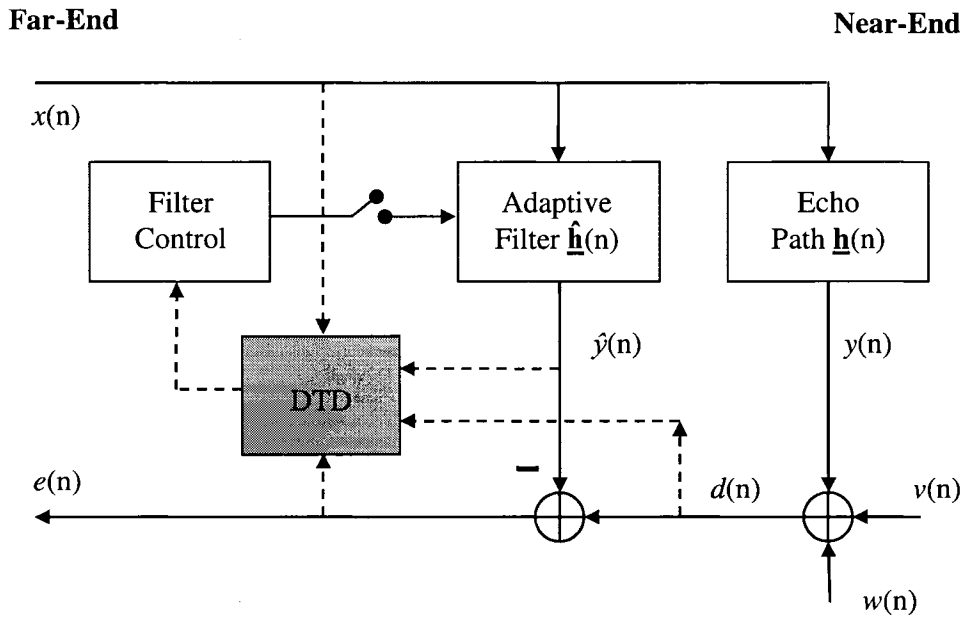

Figure 5-1: Typical Configuration of Echo Canceller and DT Detector.

In equation 5.1, the actual echo path and the far-end speech $\underline{\mathbf{x}}(n)$ are represented by the vectors $\underline{\mathbf{h}}(n) = [h_0(n) \quad h_1(n) \quad \cdots \quad h_{L-1}(n)]^{\mathrm{T}}$ and $\underline{\mathbf{x}}(n) = [x(n) \quad x(n-1) \quad \cdots \quad x(n-L+1)]^{\mathrm{T}}$ respectively, where L is the estimated length of the echo path. The echo path is assumed to be time-variant, and modeled using an adaptive FIR filter with coefficients $\underline{\hat{\mathbf{h}}}(n)$ defined similarly to $\underline{\mathbf{h}}(n)$. For simplicity in the following discussion, $\underline{\hat{\mathbf{h}}}(n)$ and $\underline{\mathbf{h}}(n)$ are also assumed to have the same length. In practice, the length of the adaptive filter is usually shorter than that of the echo path because the computation resources may be limited and/or it is not possible to define the exact echo path length. To remove the undesired echo $y(n)$, the echo

estimate $\hat{y}$(n) output from the adaptive filter is subtracted from the near-end input signal d(n) to yield a residual error e(n). The residual error is also sent to the far-end, so ideally it should consist of only v(n) and w(n) when the near-end speech is active.

$$\begin{aligned} e(n) &= d(n) - \hat{y}(n) \\ &= \underline{\mathbf{h}}(n)^{\mathrm{T}}\underline{\mathbf{x}}(n) + v(n) + w(n) - \hat{\underline{\mathbf{h}}}(n)^{\mathrm{T}}\underline{\mathbf{x}}(n). \\ &= \left[\underline{\mathbf{h}}(n) - \hat{\underline{\mathbf{h}}}(n)\right]^{\mathrm{T}}\underline{\mathbf{x}}(n) + v(n) + w(n) \end{aligned} \tag{5.2}$$

Assuming that the near-end speech v(n) is not present and the background noise w(n) is insignificant, the residual error signal in equation 5.2 becomes

$$e(n) = \left[\underline{\mathbf{h}}(n) - \hat{\underline{\mathbf{h}}}(n)\right]^{\mathrm{T}}\underline{\mathbf{x}}(n). \tag{5.3}$$

It can be seen from equation 5.3 that this is essentially a system identification problem. The goal for the adaptive filter $\hat{\underline{\mathbf{h}}}$(n) is to predict or estimate the actual echo path $\underline{\mathbf{h}}$(n) by minimizing the residual error based on some cost function. This can be done using an adaptive algorithm. Assuming that the near-end speech v(n) is present and the background noise w(n) is again insignificant, i.e. DT condition, the residual error signal in equation 5.2 becomes

$$e(n) = \left[\underline{\mathbf{h}}(n) - \hat{\underline{\mathbf{h}}}(n)\right]^{\mathrm{T}}\underline{\mathbf{x}}(n) + v(n). \tag{5.4}$$

It can be seen from equation 5.4 that the residual error would become large because of the near-end speech v(n). This would cause the adaptive filter to diverge from the true estimate of the actual echo path if the adaptation process is allowed to continue in the presence of the near-end speech. As a result, undesired echo or distorted near-end speech can pass through to

the far-end side. A DTD algorithm would detect such condition with available signals, and then stop the adaptation process of the adaptive filter at the onset of the DT condition.

The masking properties of the human auditory system have been widely exploited in many areas of research such as audio coding, digital watermarking, speech enhancement etc. Masking refers to a psychoacoustic phenomenon when two sounds are close in frequency or time, and largely different in amplitudes; the weaker sound is rendered inaudible because of the stronger one. When a sound is masked by another separated in time, it is called temporal masking; and when two sounds are produced at the same time, it is called simultaneous or frequency masking. In the following sections, a proposed DTD algorithm, which exploits the frequency masking properties of the human auditory system, is presented and evaluated.

5.2 The Proposed DTD Algorithm

5.2.1 Algorithm Overview

The configuration of an echo canceller with the proposed DT detector is shown in Figure 5-2. The novel idea of the proposed DTD algorithm is to use a psychoacoustic auditory model to determine perceptual masking thresholds of the far-end speech $x(n)$. Spectral holes in the far-end speech can then be created by removing spectral components under the perceptual masking thresholds - without affecting the perceptual quality of the far-end speech. Assuming that the echo path is a linear process, the echo $y(n)$ should ideally contain the same spectral holes as in the far-end speech. Since the near-end input signal $d(n)$ consists of the echo, the near-end speech, and the near-end background noise, the presence of the

near-end signal v(n), i.e. DT condition, can be detected by monitoring the spectrum level of d(n) at locations of the created spectral holes.

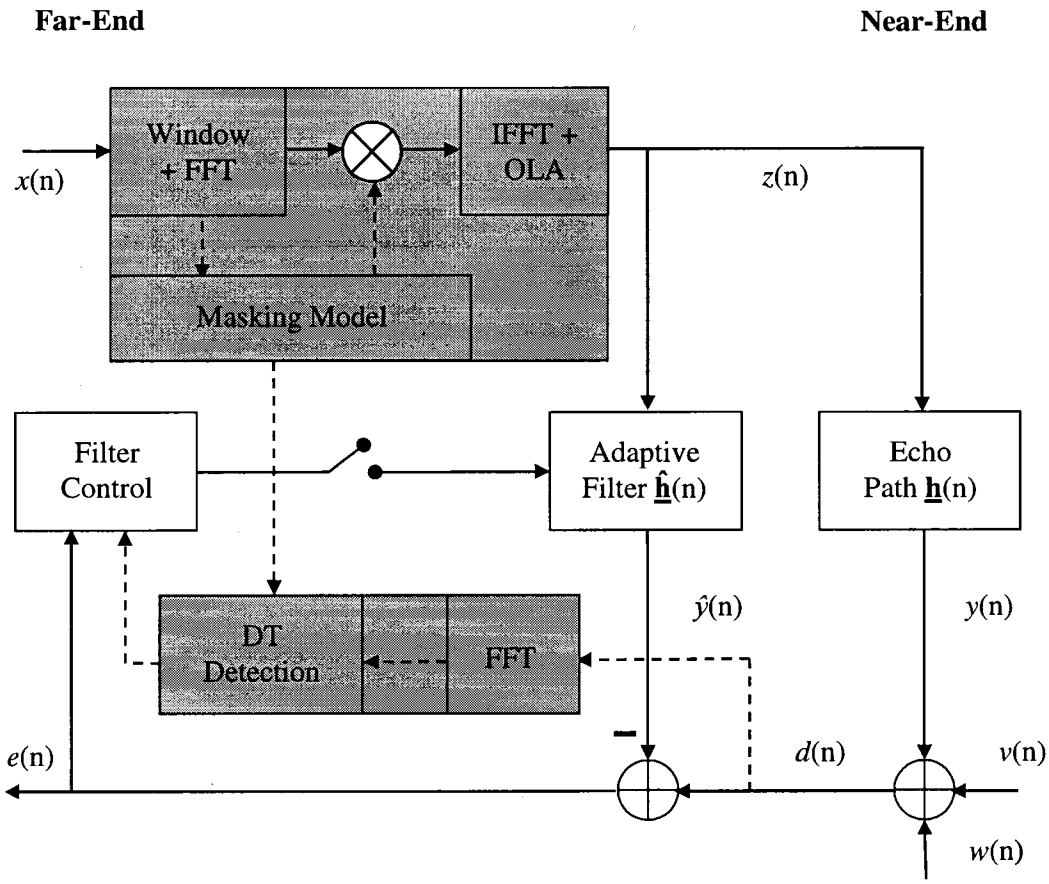

Figure 5-2: Configuration of Echo Canceller and the Proposed DT Detector.

The main functional parts of the proposed DTD algorithm are also summarized in the block diagram of Figure 5-3. The far-end signal x(n) is first segmented into overlapping frames. Each frame is then transformed into the frequency domain using the Fourier transform, and then fed into a psychoacoustic auditory model to determine a perceptual

masking threshold. The threshold of each signal frame is computed based on the principles of the absolute threshold of hearing and of the frequency masking. A binary masking template is then generated by comparing the power spectrum of a frame with its corresponding masking threshold. The masking template is set with a value of zero at frequency components below the masking threshold, and a value of one elsewhere. Spectral holes in a frame are created by zeroing out the complex spectrum components with their power spectrum less than the corresponding masking threshold values. This operation is also equivalent to multiplying the complex spectrum of a frame with its corresponding masking template for each frequency component. A masked signal z(n), which is perceptually the same as the far-end speech x(n), is generated by transforming the masked spectra of frames back to the time domain, and by adding back together overlapped sections between consecutive frames. The masked far-end signal z(n) is then passed through the echo path and sent to the adaptive filter. The near-end input signal is therefore generated as follows,

$$d(n) = \underline{\mathbf{h}}(n)^{\mathrm{T}} \underline{\mathbf{z}}(n) + v(n) + w(n) . \tag{5.5}$$

It can be seen from equation 5.5 that, during DT condition, the goal is to detect or discriminate the near-end speech v(n) from the other signals. For the proposed DTD algorithm, the near-end signal d(n) is segmented into frames and transformed into the frequency domain in a way similar to that done to x(n) during the frequency masking process. Since the spectrum of the masked speech z(n), and therefore the echo y(n), consists of holes with ideally zero spectral components, the proposed method is to monitor the presence of the near-end speech v(n) at these spectral holes using the masking template created in the

frequency masking process. Assuming that the near-end noise $w(n)$ is a white noise, with a flat spectrum, a DT condition can be deemed to have occurred when a spectral hole has energy filled up above a predefined threshold, because of the presence of the near-end speech. Since it can be assumed that the near-end speech $v(n)$ and the far-end speech $x(n)$, and therefore the echo signal $y(n)$, are independent processes, it is very likely that the non-zero spectrum of the near-end speech would be aligned with the created spectral holes in the masked signal $z(n)$. In the following sections, the detailed processes of spectral holes creation and of DT detection are described. The algorithm also features a method to combat smearing effects resulting from the frequency masking process and the filtering effect of the echo path.

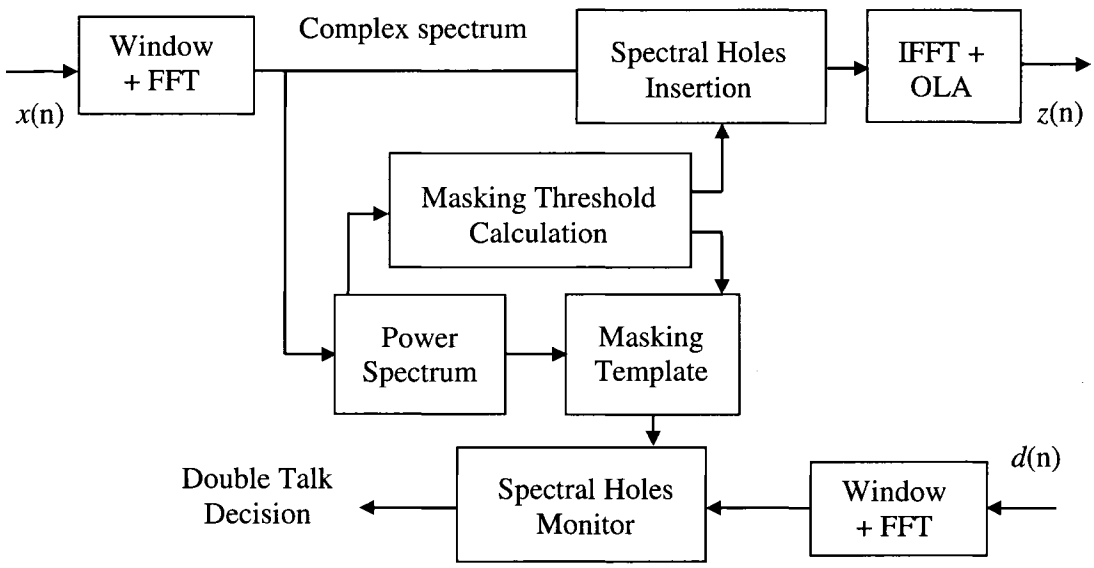

Figure 5-3: Block Diagram of the Proposed DTD Algorithm.

5.2.2 Spectral Holes Insertion

The process of spectral hole insertion is mainly based on the frequency or simultaneous masking properties of the human auditory system. Therefore, the focus for the proposed algorithm is restricted to a psychoacoustics auditory model that exploits the frequency masking properties for the calculation of a masking threshold under which spectral holes are created. In general, there have been many proposals for modeling the human auditory system with variable accuracy. Thus, selecting a model should be done from different angles and the choice is made according to a particular situation, sometimes ending up in compromise. For the proposed algorithm, the model proposed by Johnston [JOH88] is chosen because of the following underlying facts. Based mainly on the speech quality aspect, the model has already been applied with success, such as in [MA03-06] and [LAH03], and would be readily applicable to narrowband speech such as the telephone signal. In addition, Johnston's model has a simpler complexity than most others do, as it is based solely on the frequency masking properties. A block diagram of the masking threshold calculation is shown in Figure 5-4 and the steps for inserting spectral holes in the far-end signal is described in details next.

1. The far-end speech is segmented into overlapping frames of length N with a frame being defined as $\underline{\mathbf{x}}(n) = [x(n) \quad x(n-1) \quad \cdots \quad x(n-N+1)]^{\mathrm{T}}$. Segmenting a speech stream into short frames allows treating speech frame as a relatively quasi-stationary signal.

2. Each input frame $\underline{\mathbf{x}}(n)$ is weighted by a Hanning window of length N as follows.

$$\underline{\mathbf{x}}_w(n) = \underline{\mathbf{x}}(n) \bullet^* \underline{\mathbf{w}}_H(N) . \tag{5.6}$$

Where the Hanning coefficients are defined as $\underline{\mathbf{w}}_H(N) = [w(1) \quad \cdots \quad w(k) \quad \cdots \quad w(N)]^T$ and $w(k) = 0.5\left(1 - \cos\left(2\pi \frac{k}{N+1}\right)\right)$.

3. The weighted frame $\underline{\mathbf{x}}_w(n)$ is transformed into the frequency domain by the fast Fourier transform (FFT) of length M, yielding the complex spectrum in the following. Since the FFT length, M is usually greater than or equal to the frame length N, the signal $\underline{\mathbf{x}}_w(n)$ may be appended with $(M\text{-}N)$ zeroes before the FFT operation.

$$X(k) = \mathrm{FFT}_M\{\underline{\mathbf{x}}_w(n)\}, \quad k = 0, \quad \cdots \quad M-1. \tag{5.7}$$

4. The power spectrum is obtained as follows,

$$P_x(k) = |X(k)|^2, \quad k = 0, \quad \cdots, \quad \frac{M}{2} - 1. \tag{5.8}$$

5. Assuming discrete non-overlapping critical bands, the power spectrum is converted into Bark scale by adding up energies in each critical band.

$$P_b(i) = \sum_{k=F_L(i)}^{F_U(i)} P_x(k). \tag{5.9}$$

Where $i = 0, \quad \cdots, \quad I_{\max}$, and $F_L(i)$ and $F_U(i)$ are indices corresponding to the lower and upper frequencies, respectively, of the i^{th} critical band, as determined in Appendix A. The number of critical bands $(I_{\max} + 1)$ depends on the sampling frequency. For example, the narrowband speech, with 8000 Hz sampling rate, has 18 critical bands.

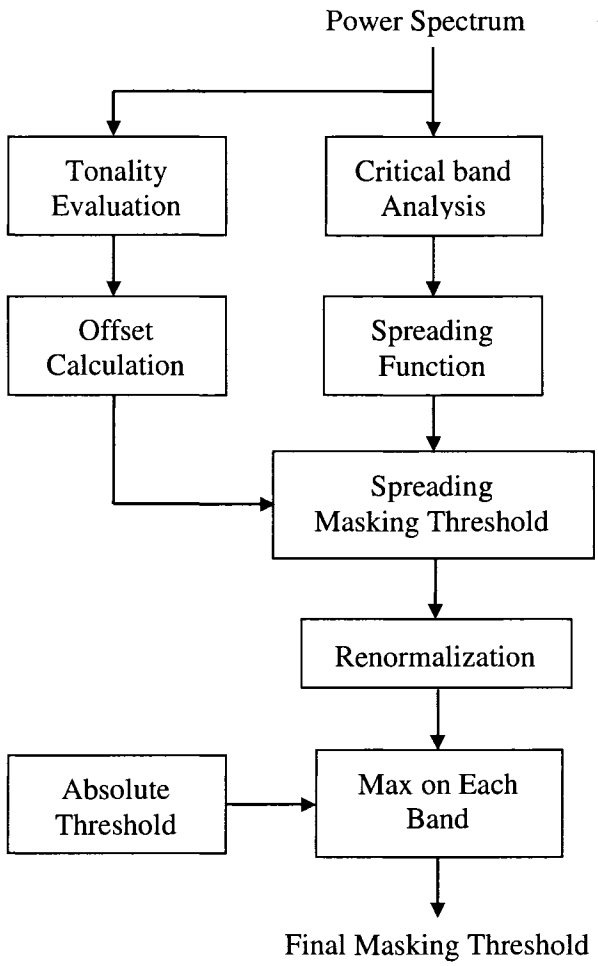

Figure 5-4: The Process of Masking Threshold Calculation

6. The spread of frequency masking is modeled by convolving the critical band energies $P_b(i)$ with a spreading function $B(z)$ as proposed in [SCH79]. The spreading function has lower and upper skirts with slopes of about +25 dB/Bark and -10 dB/Bark, respectively, and is analytically expressed as $B_{dB}(z) = 15.91 + 7.5(z + 0.474) - 17.5\sqrt{1 + (z + 0.474)^2}$, where $z = |i - j|$, and i and j represent the Bark frequencies of the masked and masking

components, respectively. The spreading function is first converted from Decibels into linear scale by using $B(z)=10^{B_{dB}(z)/10}$, and then arranged into a symmetric Toeplitz matrix **B** as the following

$$\mathbf{B}=\begin{bmatrix} B(0) & B(1) & B(2) & \cdots & B(I_{\max}) \\ B(1) & B(0) & B(1) & & \vdots \\ B(2) & B(1) & B(0) & & \\ \vdots & & & \ddots & \vdots \\ B(I_{\max}) & \cdots & & \cdots & B(0) \end{bmatrix}.$$

The critical band energies can also be written in a vector form as $\underline{\mathbf{p}}=[P_b(0) \quad \cdots \quad P_b(I_{\max})]^{\mathrm{T}}$. Therefore, the effects of masking across critical bands can be implemented as a matrix-vector multiplication in the following equation, where $\underline{\mathbf{s}}=[s(0) \quad \cdots \quad s(I_{\max})]$ denotes the spread critical band spectrum.

$$\underline{\mathbf{s}}=\mathbf{B}\underline{\mathbf{p}}. \tag{5.10}$$

7. There are two different types of maskers in Johnston's model. The tone masking noise is estimated as $(14.5+i)$ dB below the corresponding spread critical band spectrum $s(i)$. The noise masking tone is estimated as 5.5 dB below $s(i)$, independent of the Bark frequency. To differentiate the tone-like or noise-like nature of a signal, a spectral flatness measure (SFM) is used and is defined as the ratio of the geometric mean to the arithmetic mean of the power spectrum.

$$SFM_{dB}=10\log_{10}\left(\frac{\text{Geometric Mean}}{\text{Arithmetic Mean}}\right)=10\log_{10}\left(\frac{\sqrt[M/2]{\prod_{k=0}^{M/2-1}P_x(k)}}{\frac{2}{M}\sum_{k=0}^{M/2-1}P_x(k)}\right). \tag{5.11}$$

A coefficient of tonality α is then derived from the spectral flatness measure as follows.

$$\alpha = \min\left(\frac{SFM_{dB}}{SFM_{dB\min} = -60dB}, 1\right). \tag{5.12}$$

Therefore, for a frame with SFM value smaller or equal to -60dB, the coefficient of tonality is $\alpha = 1$, which means the frame is tone-like. For a frame with SFM value of 0 dB, the coefficient is $\alpha = 0$, which means the frame is noise-like. The threshold offset for masking energy in each critical band is then calculated as follows.

$$o_{dB}(i) = \alpha(14.5 + i) + 5.5(1 - \alpha). \tag{5.13}$$

8. The threshold offset is then subtracted from the spread critical band spectrum to yield the spread threshold estimate in linear scale for each critical band.

$$T_s(i) = 10^{\left(\log_{10}(s(i)) - o_{dB}(i)/10\right)}. \tag{5.14}$$

9. The spreading function, because of its shape, increases the energy estimates in each critical band due to the effects of spreading [JOH88]. The effects must be undone, i.e. the spread threshold estimate should be deconvoluted. The rationale underlying this process is related to the excitation pattern model of masking. A target signal is audible if the difference between the excitation patterns produced by a masker individually and the combination of the masker and the target is less than a critical value. Assuming linear summation of excitation patterns in the spectral domain, the spread threshold estimate represents the maximum acceptable excitation produced by the target signal [CAV02]. It was suggested by Johnston that the deconvolution is a very unstable process because of

the shape of the spreading function, and a normalization process is therefore used in place of the deconvolution to return the spread threshold $T_s(i)$ to a desired level. This is simply done by dividing a critical band component of the spread threshold by the number of frequency indices in the corresponding critical band as follows.

$$T_n(i) = \frac{T_s(i)}{M_i}. \tag{5.15}$$

Where M_i is the number of frequency indices in critical band i ($i = 0, \cdots, I_{\max}$).

10. Taking into account the absolute threshold of hearing, the final masking threshold is calculated by comparing the normalized threshold with the normalized absolute threshold for each critical band. The normalization of the absolute threshold of hearing is done by equating its lowest point, i.e. near 4 kHz, to the energy in ±1 bit of the signal amplitude.

$$T_f(i) = \max\left[T_n(i), \quad T_{abs}(i)\right]. \tag{5.16}$$

11. The final masking thresholds are then converted from the Bark scale back to the linear frequency scale. Spectral holes are created by comparing the spectral components with the final masking thresholds. Components with energy above their corresponding thresholds are retained; otherwise, they are set to zero.

$$X_m(k) = \begin{cases} X(k) & P_x(k) \geq T_f(k) \\ 0 & \text{otherwise} \end{cases}. \tag{5.17}$$

A binary masking template $t_m(k)$ is also created with values of zero at spectral components below the corresponding masking thresholds, and values of one elsewhere.

$$t_m(k) = \begin{cases} 1 & P_x(k) \geq T_f(k) \\ 0 & \text{otherwise} \end{cases} . \tag{5.18}$$

The masking template will be later used to locate spectral holes for detecting the presence of the near-end speech in the near-end input.

12. The masked complex spectrum of a current frame is then converted back to the time domain using an M-point inverse FFT to yield the current output frame. The reconstruction of the masked far-end speech z(n) is done by adding the overlapped portions of the consecutive output frames as illustrated in Figure 5-5. By adding the later half of the previous output frame to the first half of the current output frame, the effect of the Hanning windowing process in Step 2 can be approximately cancelled.

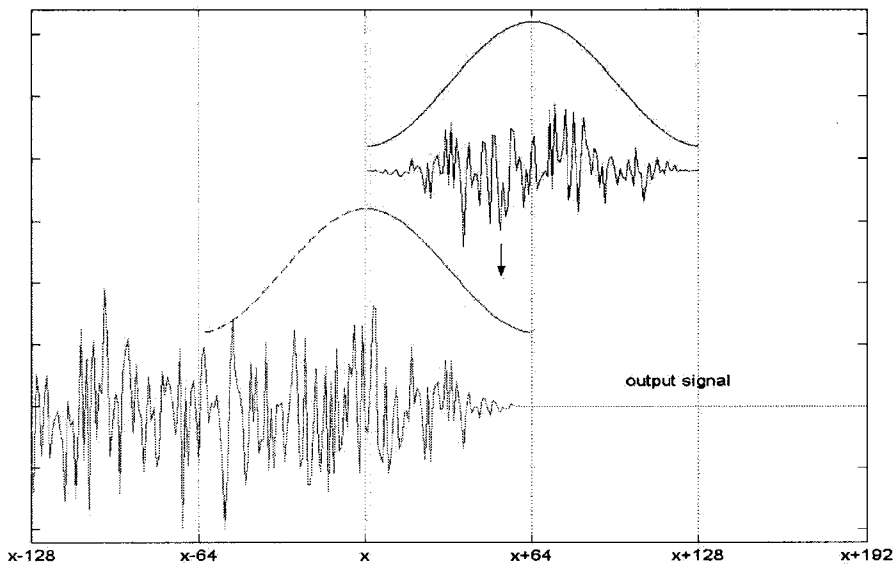

Figure 5-5: Construction of Output Signal by Overlap-Add.

5.2.3 Spectral Holes Monitor

The near-end signal d(n), being the sum of the echo y(n), the speech v(n), and the additive noise w(n), is segmented into frames in the same manner as that for the far-end signal x(n). The masking template $t_m(k)$ generated in the Step 11 of the above section is used to keep track of the locations of spectral holes in the corresponding frame of d(n). It is therefore important to properly synchronize a masking template with a d(n) frame so that the created spectral holes in d(n) align with those of the template. The alignment can be done by simply monitoring the magnitude peak of the adaptive filter coefficients, or by a signal cross-correlation method between x(n) and d(n). The detection of v(n), i.e. a DT condition, is performed by comparing the spectral level of the d(n) frame at the spectral holes with a threshold T just above a monitored noise floor level. If the spectral level is greater than the threshold T, then a DT condition is declared. The decision variable of the proposed algorithm is therefore defined as following.

$$\xi = \begin{cases} 1 & Any\ P_d(k) > T \\ 0 & Otherwise \end{cases}, \qquad k \in \{k : t_m(k) = 0, k = 0, \cdots, M/2-1\}, \tag{5.19}$$

where the power spectrum and complex spectrum of signal d(n) are calculated as follows.

$$P_d(k) = |D(k)|^2 \text{ and } D(k) = \text{FFT}_M\left\{\underline{\mathbf{d}}(n)\underline{\mathbf{w}}_H^{\mathrm{T}}(N)\right\}. \tag{5.20}$$

5.2.4 Smearing Effects

This section describes the smearing effects resulting from the frequency masking process and the filtering of the echo path on the created spectral holes and proposes a remedy to combat the problem.

The smearing effects refer to leakage or spilling of energy from the unmasked frequency components into the created spectral holes. First, the masking threshold of a signal frame can be very different from frame to frame because of the non-stationary properties of speeches. As a result, locations of spectral holes, which are determined by a masking template, can vary between consecutive frames. Therefore, the overlap-add process as described in the Step 12 of the above section may add non-zero unmasked spectral components of the previous frame to the created spectral holes in the current frame. Second, the smearing effect can be caused by the filtering process of the echo path $\underline{\mathbf{h}}(n)$. The actual echo $y(n)$ is a weighted sum of signals originating from consecutive frames with possibly different spectral hole locations. As a result, some "would be" spectral holes of the frame of $y(n)$ may be filled up by signal components from previous frames.

Since the proposed algorithm depends on the integrity of the created spectral holes for detection of a DT condition, any excessive level of the smearing effects at the spectral holes can cause the algorithm to detect the condition falsely as a DT. Since the smearing effects on spectral holes are certainly unavoidable, the detection of near-end speech $v(n)$, i.e. a DT condition, should be done only at intact locations of spectral holes. It is therefore very important to identify places in the spectral holes where level of smearing is excessively high

so that the detection process can be avoided at these places. A remedy to the problem is to use the output of the adaptive filter $\hat{y}$(n) to identify these places because it is a reasonable estimate of y(n) once the adaptive filter has converged. A smearing template, similar to the masking template, is therefore created by comparing the spectral level of $\hat{y}$(n) at the created spectral holes with a threshold to locate smearing effects at these holes. The smearing template is defined as follows.

$$t_s(k) = \begin{cases} 1 & P_{\hat{y}}(k) \geq T_s \quad k \in \{k : t_m(k) = 0, k = 0, \cdots, M/2-1\} \\ 0 & \text{Otherwise} \end{cases} \tag{5.21}$$

$$\text{Where } P_{\hat{y}}(k) = \left|\hat{Y}(k)\right|^2 \text{ and } \hat{Y}(k) = \text{FFT}_M\left\{\underline{\hat{\mathbf{y}}}(n)\underline{\hat{\mathbf{w}}}_H^{\mathrm{T}}(N)\right\}. \tag{5.22}$$

Any spectral component of $\hat{y}$(n) in locations of the created spectral holes with a power level above a preset smearing threshold T_s, is excluded from the DT detection process. A final template is created by taking a union of the masking and the smearing templates as $t_f(k) = t_m(k) \cup t_s(k)$. The final decision variable is therefore modified as follows.

$$\xi = \begin{cases} 1 & \textit{Any } P_d(k) > T \\ 0 & \textit{Otherwise} \end{cases}, \quad k \in \{k : t_f(k) = 0, k = 0, \cdots, M/2-1\}. \tag{5.23}$$

Figure 5-6 demonstrates the described idea of combating the smearing effects. Figure 5-6a shows the log-scaled spectrum of an input frame of x(n) with its corresponding masking threshold. Figure 5-6b shows the masked, or holed, spectrum with a generated masking template in linear scale. Figure 5-6c to Figure 5-6e show a zoomed section of the same signal frame. Figure 5-6c illustrates the smearing effect at spectral holes of the echo signal y(n). Figure 5-6d shows the corresponding spectral holes in the echo estimated signal $\hat{y}$(n), and a

smearing template, which identifies the spectral components at the holes where smearing level is high. Note that the misadjustment level for the echo estimate shown in the Figure 5-6d is at around -10 dB. Figure 5-6e shows the spectral holes of y(n) and the modified detection masking template, which is a union of the masking and the smearing templates.

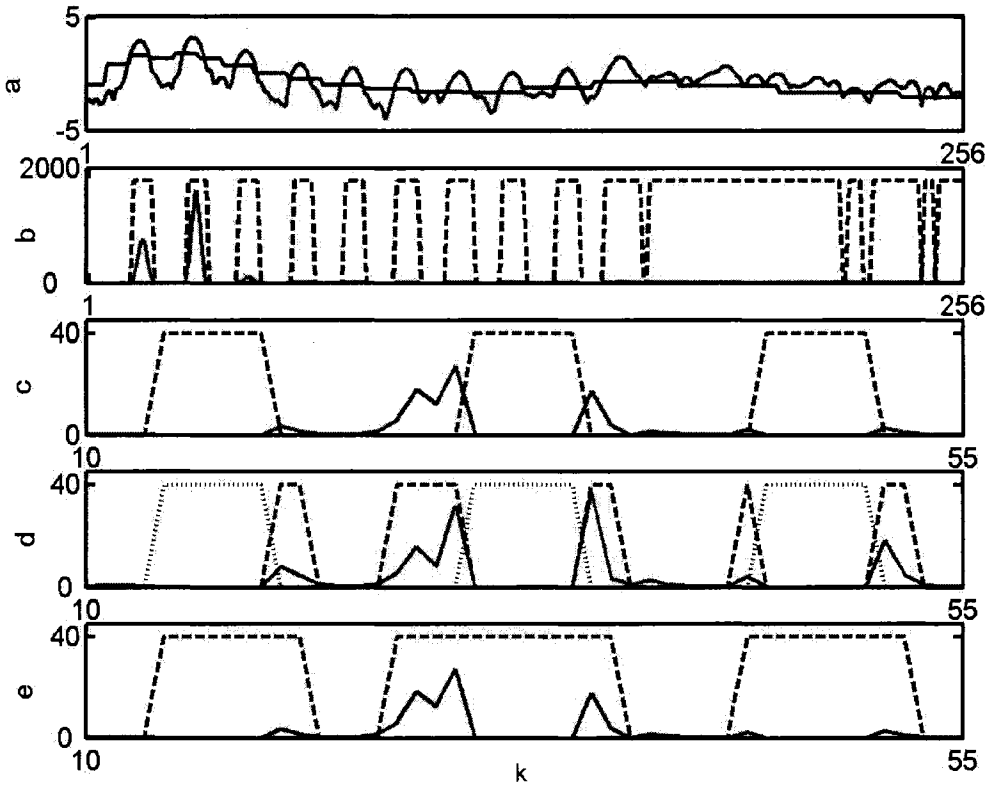

Figure 5-6: Spectral Holes and Smearing Effects of a Signal Frame: a) spectrum and masking threshold of $x(n)$ in log scale; b) holed spectrum and masking template of x(n) in linear scale; c) zoomed spectrum of $y(n)$ showing masking template and smearing effects at spectral holes; d) zoomed spectrum of $\hat{y}(n)$ showing smearing template; e) final masking template for detection as an union of the masking and the smearing templates.

5.3 Simulations and Results

5.3.1 Evaluation Methodology

The proposed DTD algorithm is evaluated and compared with other algorithms using the objective technique proposed by [CHO99]. This evaluation technique is based on the standard statistical methods of binary detection theory developed for radar and communication applications. It proposes a systematic approach for selecting a threshold T for the decision variable, evaluating, and comparing different DTD algorithms objectively. The performance of a DT detector is measured by means of the characteristics defined in the following.

- *Probability of False Alarm* (P_f): Rate of samples where a DT condition is falsely claimed.
- *Probability of Detection* (P_d): Rate of samples where a true DT condition is detected.
- *Probability of Miss* (P_m=1-P_d): Rate of samples where a true DT condition is missed.

The goal in designing a DT detector is therefore to maximize P_d and minimize P_f and P_m. In general, lower P_m (or higher P_d) means higher P_f. In addition, the performance of a DT detector should be measured together with an echo canceller as a joint system because it can seriously affect the convergence rate of the echo canceller. A common approach to evaluate or characterize a DT and EC system is to represent detection characteristic P_d or P_m as a function of a system parameter, such as speech to noise ratio, for a given constraint of false alarm rate P_f. A measurement using those previously defined characteristics are known in

applications of the binary detection theory as a receiver operating characteristic (ROC) curve. The constraint of P_f can be set up by varying the threshold of decision variable T for typical far-end signals when there is no presence of near-end speech. The level of setup P_f would depend on, for example, the maximum tolerable false alarm rate of an EC system or the sensitivity of the adaptive filter to missed detection.

Though the ROC representation provides a useful performance metric for DTD, it still lacks an important piece of information on the curve. That is the response time of a DT detector because it can indicate how well the adaptive filter would stay converged during DT condition. Ideally, a DT detector's response should be as quick as the onset of DT. That means that the DT detector should arrive at a DT decision faster than the adaptive filter starts to diverge from its optimum estimate. Performance measurement of an EC and DTD system using the ROC curves should therefore be accompanied or complemented by another measurement, such as the echo return loss between the actual echo path and its estimation, which shows the convergence rate of the whole system. In addition, the performance of a DT detector should be evaluated under special conditions such as echo path changes to show how well the system handles such cases.

It was shown by the authors in [BEN01] that the performance of an adaptive filter and a DT detector as an EC system is governed by the echo-to-near-end-signal (i.e. near-end speech plus noise) power ratio (ENER). In setting parameters for a DT detector, it is therefore important to consider ENER. For a low ENER, the near-end speech is more dominant than the echo; the adaptive filter is therefore more sensitive to a DT condition. As a

result, a low P_m is important. This means that one has to tolerate more false alarms (higher P_f), which will often stop the adaptation unnecessarily and thus slow down the overall convergence. On the other hand, a high ENER would mean that the adaptive filter is less sensitive to a DT condition. However, while the divergence of the adaptive filter would be less likely with a higher ENER, the performance of the DT detector would degrade, so these properties somewhat balance each other.

In general, there is a trade-off between P_m and P_f. A lower P_m can be achieved by a higher P_f and vice versa. In the case of DT condition, a low P_m rate is more important because it prevents divergence of the adaptive filter due to DT. The cost for higher P_f means a lower convergence rate because a false alarm simply stops the adaptation for some hold time. On the other hand, a high P_f rate becomes more critical in the case of EPC because of slower recovery from EPC. In practice, one would consider the trade-off in performance depending on a cost function of the false alarm and miss detection.

5.3.2 Simulations

The proposed DTD algorithm was evaluated and compared against the Geigel and the NCC DTD algorithms. The DTD algorithms were simulated with a common single-channel EC platform using the Matlab software. Figure 5-7 shows the block diagram of the simulation setup for evaluating the EC system with the DTD algorithms.

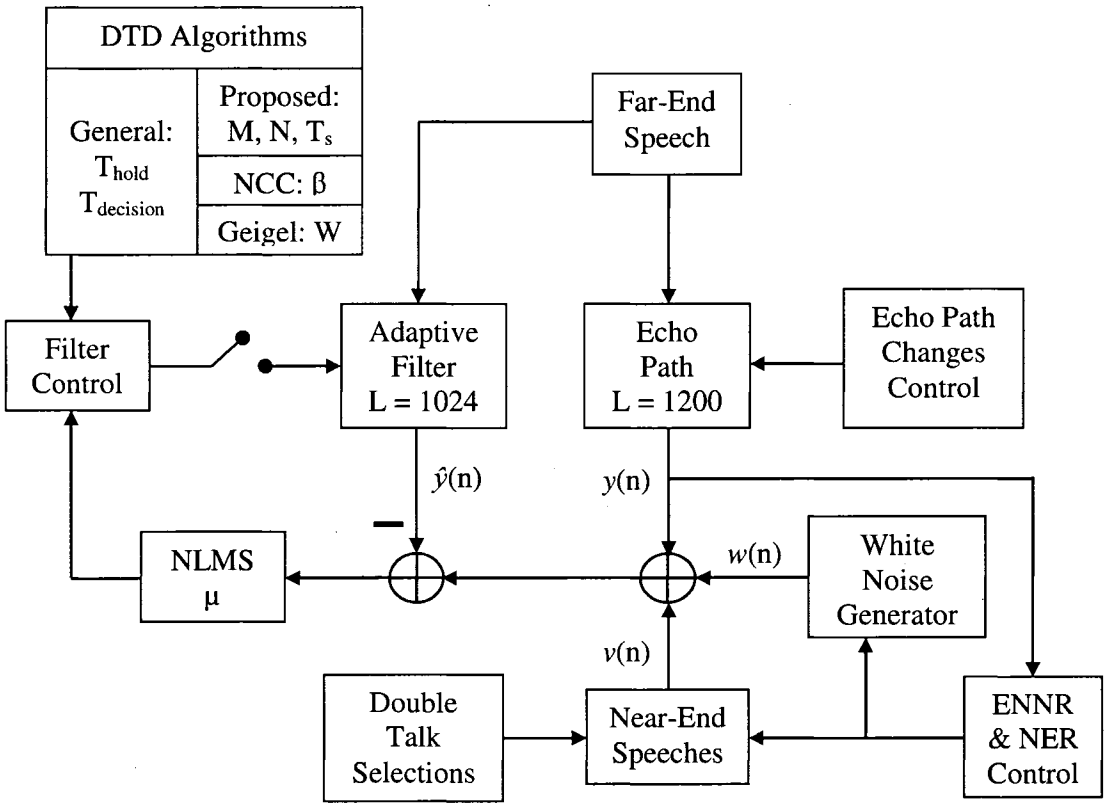

Figure 5-7: Simulation Setup.

Simulations in this chapter were conducted using an actual room impulse response with length of 1200 samples as the echo path. The length of the echo path is dictated by the length of the filter used in measuring the room impulse response. Figure 5-8 shows the first 200 samples of the impulse response with its characteristics shown in more details in Appendix B. The impulse response was also normalized so that the power ratio of the echo $y(n)$ to the far-end speech $x(n)$ is 0 dB. This is just for convenience but not a necessity.

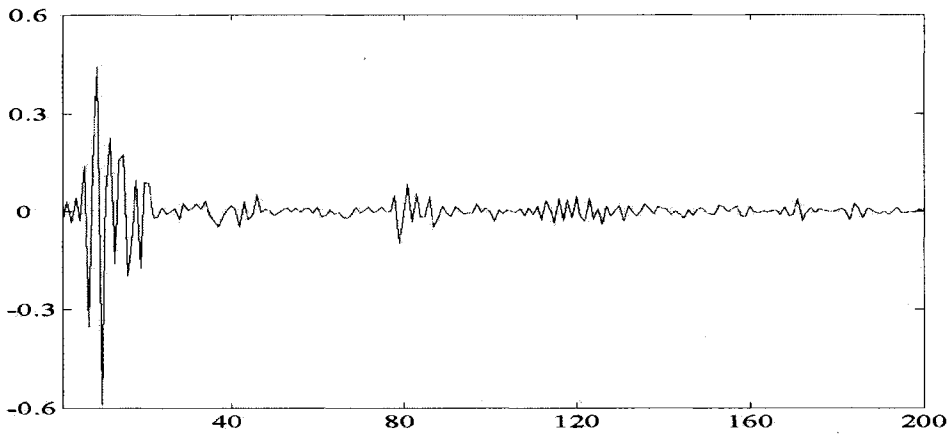

Figure 5-8: Room Impulse Response (First 200 of 1200 samples)

The input far-end and near-end signals, at 8 kHz sampling rate, were taken from the Harvard sentences speech database prepared by the Acoustics and Signal Processing group at the NRC lab. The input far-end signal $x(n)$, from a female talker, is about 10 seconds long. There are six input near-end speeches for simulating a DT condition from three different females and three different males; each is about 2 seconds long. The levels of the near-end speeches are varied with respect to the echo level based on a specified Near-end speech $v(n)$ to Echo $y(n)$ power Ratio (NER). The background noises $w(n)$ used in the simulations are white Gaussian noises, with zero mean, randomly generated using a built-in Matlab function, and varied at different levels with respect to the echo level based on a specified Echo $y(n)$ to Near-end Noise $w(n)$ power Ratio (ENNR),

The adaptive filter is an FIR transversal structure with length L = 1024 coefficients, or weights. The filter length is a little smaller than the actual echo path (1200), making it more

realistic because the echo path is usually under-modeled by the adaptive filter. The normalized least mean square (NLMS) is used for all DTD algorithms as the adaptive algorithm with its step size set at 1.0 for the first 20000 samples (2.5 seconds) of the far-end signal, and then reduced to 0.1. The DT detectors are also turned on from the sample number 20000. The larger step size set in the first 2.5 seconds allows a faster initial convergence of the adaptive filter. The smaller step size is then set to prevent the adaptive filter from diverging too fast because of a missed detection and/or the slow response time of a DTD algorithm. In addition, the hold time for the DTD algorithms is set at 30 milliseconds.

For the proposed algorithm, an FFT length $M = 256$ and a frame length $N = 128$ samples are used; consecutive frames are 50% overlapped. A short frame is preferred as it reduces the delay of the EC system, and that of the DTD decision. Short frame also allows speech to be treated as a relatively quasi-stationary signal, of which the spectrum can be more alike from frame to frame. As a result, locations of the created spectral holes would not change very drastically from frame to frame, which may help to reduce the smearing effects resulting from the masking process. Note that the selection of the M and N values is based on the perceptual quality of output speech from the Johnston's model used in the proposed DTD algorithm and on the overall performance of the EC system. In addition, the proposed DTD algorithm processes a frame of data samples to arrive at a DT decision whereas the NLMS algorithm updates the filter weights on a sample-by-sample basic. Due to the block processing nature, it is therefore possible that the response of the proposed algorithm is later than the actual onset of a DT condition, by a maximum delay of half a frame of samples. A simple solution to accommodate the slow response time of the proposed algorithm is to

maintain a buffer of recently good filter weights. Once a DT condition is detected, the system can revert to using the previously saved filter weights. The cost for this solution is an increase in memory requirement.

Due to limited time and resources, the performance of the psychoacoustic auditory model used in the proposed DTD algorithm was first evaluated using informal listening tests. The tests were conducted in the following manner. Male and female speeches, with lengths between 2-8 seconds, were used as inputs to the model. Different speech outputs created by different parameter settings of the model were evaluated and compared by two expert listeners. A decision was made on what setting would produce more spectral holes with minimal degradation in speech quality.

There are two thresholds used by the proposed DTD algorithm. First, the decision variable threshold is selected following the procedure described in the evaluation methodology section for a given initial constraint of P_f when there is no near-end speech. The other is the smearing threshold that determines the smearing effects in the created spectral holes. This threshold can be preset with a value depending on the computational precision of the application program because ideally the spectral holes should have zero components. In addition, the smearing threshold should be a smaller value than the decision variable threshold in order to prevent the smearing effects from being detected as DT conditions. For all simulations in this section, the smearing threshold is set at 0.0033.

For comparison, the Geigel algorithm uses the decision variable computed by Equation 5.18, with the window size equal to the adaptive filter length $N = 1024$.

$$\xi = \frac{|d(n)|}{\max\{|x(n),| \quad ..., \quad |x(n-N+1)|\}}. \tag{5.24}$$

For the NCC algorithm, the decision variable is estimated using the following computationally simplified equation.

$$\xi = \frac{\sqrt{\hat{\underline{\mathbf{r}}}_{xd}^{\mathrm{T}} \hat{\underline{\mathbf{h}}}}}{\hat{\sigma}_d}. \tag{5.25}$$

Where $\hat{\underline{\mathbf{r}}}_{xd}$ is an estimate of the cross correlation vector between signals $\underline{\mathbf{x}}$(n) and d(n); $\hat{\sigma}_d$ is an estimate of the variance of signal d(n). In addition, $\hat{\underline{\mathbf{r}}}_{xd}$ and $\hat{\sigma}_d$ can be practically computed using a forgetting factor $(\beta = 0.99)$ as

$$\hat{\underline{\mathbf{r}}}_{xd}(n) = \beta \hat{\underline{\mathbf{r}}}_{xd}(n-1) + (1-\beta)\underline{\mathbf{x}}(n)d(n), \tag{5.26}$$

$$\hat{\sigma}_d(n) = \beta \hat{\sigma}_d(n-1) + (1-\beta)d^2(n). \tag{5.27}$$

For the ROC curves in this chapter, the probability of miss P_{m} is used as a performance measure versus other variable parameters such as the NER, the ENNR, and the probability of false alarm P_{f}. The P_{m} is calculated as the percentage of samples where a true DT condition is not detected. The P_{f} is calculated as the percentage of samples where a DT is falsely claimed. The P_{m} and P_{f} rates are counted only during the active portions of the far-end and near-end speeches. For inactive far-end signal, the adaptation process is stopped; the near-end signal is passed to the far-end without correction. If the adaptation process is not stopped in this case, the adaptive filter could drift away due to the activity of the background noise at the near-end. On the other hand, the inactive near-end speech does not cause the adaptive

filter to diverge. In addition, the speech activity detector proposed in [CHO99] is used to detect the active portion of speech, and shown in Figure 5-9. The threshold used in the activity detector is chosen as -30 dB with respect to the normalized input speech; the constant α is heuristically selected as $\alpha = e^{-20/L}$, with L the length of the adaptive filter; and $\bar{g}(n)$ is the binary output of the activity detector.

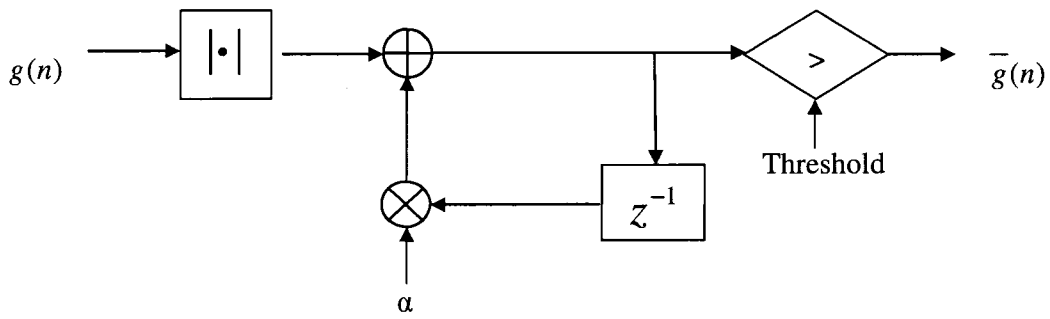

Figure 5-9: Speech Activity Detector.

The following procedure describes the steps to obtain the ROC curves in the results section.

1. Select one of the parameters NER, ENNR, or P_f as a variable and set the others to fixed values.

2. With the near-end speech signal v(n) = 0, select a value of the decision variable threshold T for each algorithm that results in a given probability of false alarm P_f. The P_f is calculated using the following equation.

$$P_f = \frac{\sum \phi(n)\bar{x}(n)}{L_{far-end}}, \tag{5.28}$$

where $\phi(n)$ is the binary output of a DT detector, $\bar{x}(n)$ is the activity detector output of the far-end signal x(n) (z(n) for the proposed DTD algorithm), and $L_{far\text{-}end}$ is the length of the far-end signal.

3. Select a value for the variable selected in the Step 1.

4. Select one of the six DT signals.

5. Select one of six random positions within the last 7.5 seconds of the far-end speech.

6. Compute the probability of P_m as follows.

$$P_m = 1 - \frac{\sum \phi(n)\bar{x}(n)\bar{v}(n)}{\sum \bar{x}(n)\bar{v}(n)}, \tag{5.29}$$

with $\bar{v}(n)$ being the activity detector output of the near-end signal v(n).

7. Repeat Steps 4 to 6 over all thirty-six conditions (six near-end signals times six random locations).

8. Average the P_m values over the thirty-six realizations obtained in the Step 7.

9. Plot the averaged P_m as a function of the variable selected in the Step 1.

Additionally, the performance of the EC and DTD system during a DT condition can be measured by means of the excess echo return loss (EERL). The EERL is defined as the normalized mean square error (MSE) between the actual echo path and its estimation by the echo canceller. Since an EERL plot shows the convergence rate of the adaptive filter, it is used as a complement measurement to the ROC curves, as the convergence rate in the plot indicates the response time of a DTD algorithm. In addition, the EERL measurement is used so that the effects of the near-end signal and the background noise are not taken into account from the plot. To fairly compare the convergence rates, the DTD algorithms are tuned to achieve the same asymptotic minimum mean square error as when there is no near-end signal, i.e. $v(n) = 0$ for a given constraint of false alarm rate P_f. In the simulation environment, the EERL measurements for the DTD algorithms can be made with full knowledge of the impulse response of the actual echo path, which is not applicable in real-time measurement of an EC implementation. The EERL is computed as the normalized MSE between the echo and its estimate as follows.

$$\mathrm{EERL}(n) = 10\log_{10}\left(\frac{\hat{\mathbf{E}}\left[\left[y(n)-\hat{y}(n)\right]^2\right]}{\hat{\mathbf{E}}\left[\left[y(n)\right]^2\right]}\right), \tag{5.30}$$

where the operator $\hat{\mathbf{E}}[\bullet]$ is defined as a time average process using the first-order IIR low pass filter with a single pole at $\lambda = 0.999$ (time constant is 1/8 s at 8 kHz sampling rate). For the EERL simulations in the results section, the far-end signal and a near-end signal from the ROC data set are used to measure the normalized MSE with a very similar set-up with the

ROC results. In addition, since speech has a time varying power level, simulations with a set of speeches may give less consistent estimate of the MSE measurement than in the case of stationary signal. Nevertheless, for comparative studies of the DTD algorithms, it is still a very useful performance metric to use speech signals.

The EERL plots are also used to evaluate how well the DTD algorithms handle conditions of EPC. In reality, there are infinite possibilities of EPCs. For the evaluations in this chapter, four typical cases of EPCs are simulated and investigated. They are

1. Time delay: $h(n) \rightarrow h(n-\delta)$.
2. Time advance: $h(n) \rightarrow h(n+\delta)$.
3. Echo path gain: $h(n) \rightarrow \alpha h(n)$, where $\alpha > 1$.
4. Echo path loss: $h(n) \rightarrow \alpha h(n)$, where $\alpha < 1$.

When an EPC occurs, it is important that the adaptive filter can keep track of the change and converges quickly to the changed echo path. Therefore, a major concern with EPC is the false alarm rate P_f of a DT detector. If a DT detector is too sensitive to a condition of EPC, the false alarm rate P_f may increase; the convergence rate of the adaptive filter therefore slows down. In general, slower convergence rate of an EC system during EPC means a better immunity to DT conditions. For simulations under EPCs, the same female far-end signal is used to evaluate the four cases of EPCs. To compare fairly the convergence rates, the DTD algorithms are set up for a given constraint of missed detection rate P_m in the presence of a

DT signal. The DT signal is then removed, an EPC is simulated during the length of the far-end signal, and the normalized MSE and the increase/decrease in false alarm rate P_f are measured.

5.3.3 Simulation Results

The following subsections present and discuss results of the simulations described in the above section. In addition, the computational complexity of the DTD algorithms are compared and discussed.

5.3.3.1 Receiver Operating Characteristic Performance

Figure 5-10 shows a plot of the probability of miss P_m as a function of the near end speech to echo power ratio (NER) for a range from -20dB to 10dB, for the echo to near-end noise power ratio ENNR of 30 dB, and for the given constraint of false alarm rate $P_f = 0.1$. The ROC plot was obtained following the procedure described in section 5.3.2. Each point on the plot is therefore an average of 36 realizations. In addition, the standard deviation of each point was also calculated and shown on the plot. For all algorithms, it can be seen from the plot that as the power of the near-end speech $v(n)$ increases in comparison with the echo signal, or with the far-end signal x(n) because of the normalized impulse response of the echo path, the probability of miss P_m decreases. Therefore, the lower the curve, the better the performance of an algorithm is. The Geigel algorithm performs much worse than the other algorithms, while the proposed algorithm performs the best for the whole range of NER.

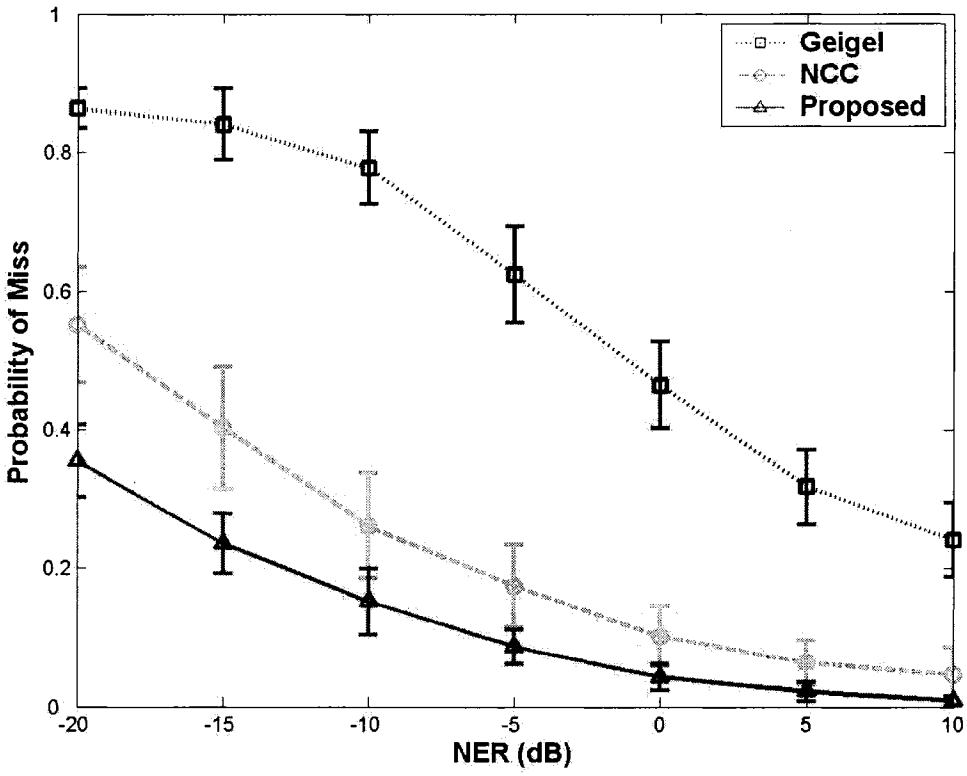

Figure 5-10: P_m versus NER for ENNR = 30 dB and $P_f = 0.1$.

Figure 5-11 shows a plot of P_m as a function of P_f in a range from 0.0 to 0.3, with the near-end speech to echo power ratio NER set at -10 dB and the echo to near-end noise ENNR set at 30 dB. The ROC plot was obtained following the procedure described in section 5.3.2. Each point on the plot is therefore an average of 36 realizations. In addition, the standard deviation of each point was also calculated and shown on the plot. It can be seen from the plot that there is a trade off between P_m and P_f. Setting a DT detector at a low rate of P_m would increase the false alarm rate P_f, so the lower the ROC curve is in this case the better

the performance of a DT detector is. In overall, the proposed algorithm performs the best for the whole range of probability of false alarm P_f.

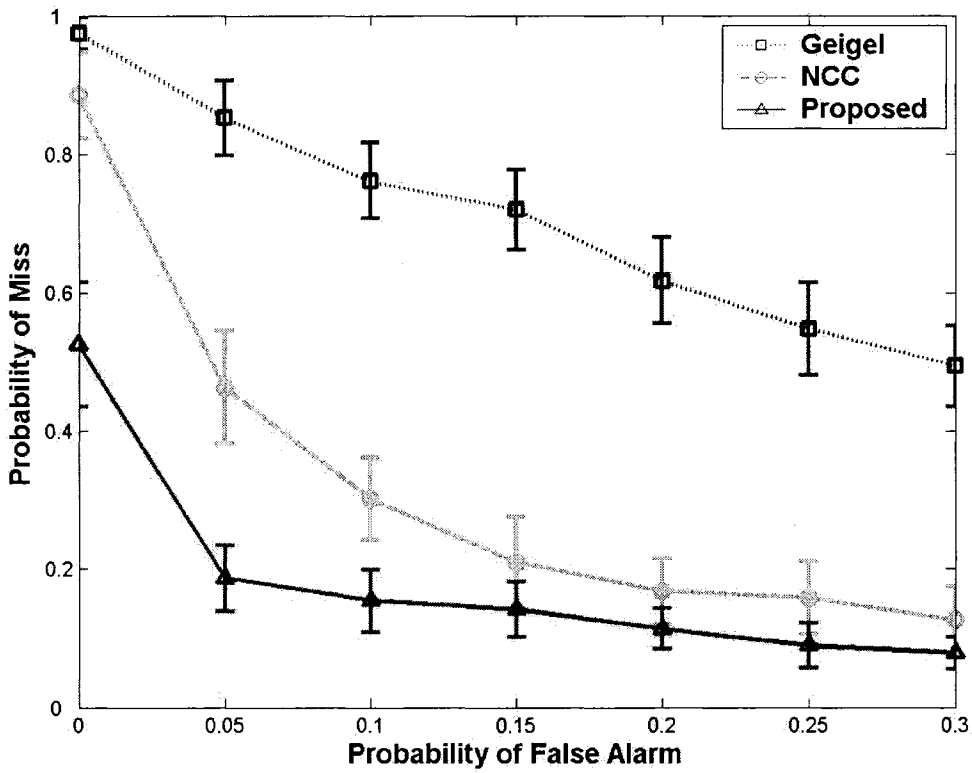

Figure 5-11: P_m versus P_f for NER = -10 dB and ENNR = 30 dB.

Figure 5-12 shows a plot of P_m as a function of the echo to near-end noise power ratio ENNR for a range from 0 dB to 50 dB, with the near-end speech to echo power ratio NER set at 0 dB, and with the constraint of false alarm rate P_f set at 0.1. The ROC plot was obtained following the procedure described in section 5.3.2. Each point on the plot is therefore an average of 36 realizations. In addition, the standard deviation of each point was also

calculated and shown on the plot. As the power of near-end noise $w(n)$ increases or ENNR decreases, P_f increases. Thus, the threshold decision variable T has to be adjusted to keep the given constraint of P_f at 0.1. This increases P_m as a result. Therefore, the lower the curve, the better the performance of an algorithm is. From the plot, no DTD algorithms perform very well when the echo level, which is the same as the near-end signal level, is in the range 0-10 dB with respect to the near-end noise level. However, the proposed algorithm still performs better than the other algorithms overall.

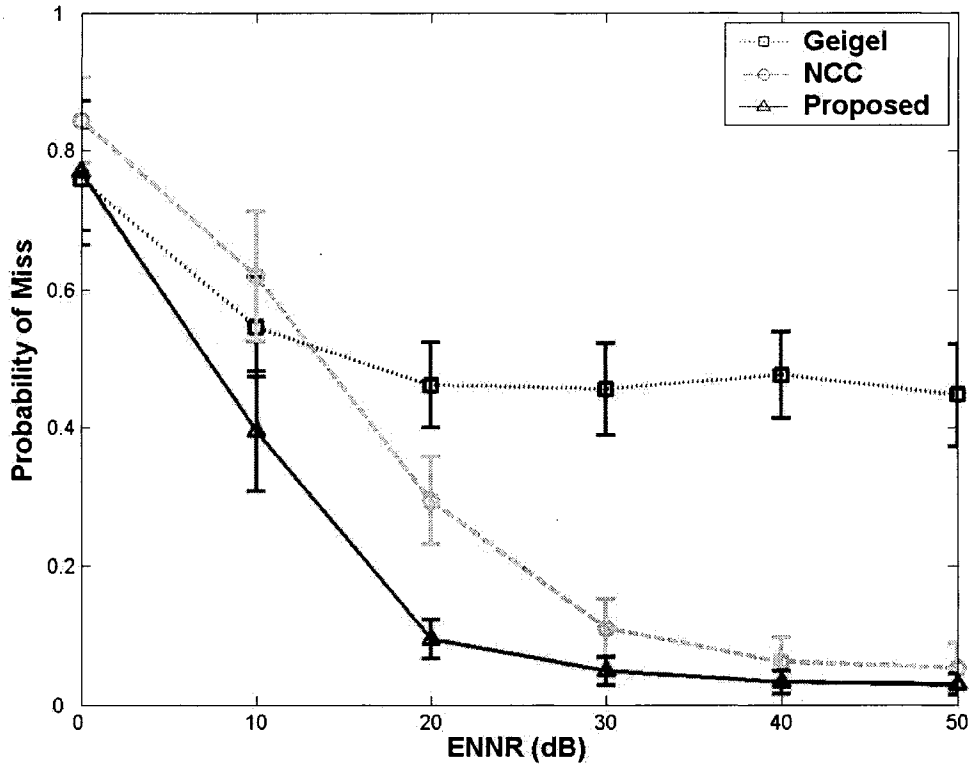

Figure 5-12: P_m versus ENNR for NER = 0 dB and P_f = 0.1.

5.3.3.2 Excess Echo Return Loss Performance

Figure 5-13 shows the plots of DT decisions and normalized MSE performance of the DTD algorithms during a DT condition. The constraints of false alarm rate P_f is set at approximately 0.05 with the near-end speech to echo power ratio NER = 0 dB, and with the echo to near-end noise power ratio ENNR = 30 dB. The DT detectors are turned on after the first second of simulations. In the top plot, the far-end signal is about ten seconds long, and the near-end speech or the DT signal starts four seconds after the far-end signal does and lasts for about two seconds. The top plot also compares the detection performance of the DTD algorithms. It can be seen from the plot that the proposed DTD algorithm has the lowest miss detection rate. The Geigel, as expected, performs the worst. The miss rates P_m of the DTD algorithms are measured and given in Table 5-1.

	Proposed	**NCC**	**Geigel**
P_m	0.07	0.37	0.64

Table 5-1: Measured P_m for P_f = 0.05, NER = 0 dB, and ENNR = 30 dB.

The bottom plot compares the EERL performance of the DTD algorithms. It can be seen from the plot that the normalized MSE curve of the proposed algorithm does not diverge significantly during the entire length of the DT signal. This is as expected because the proposed algorithm has a very low rate of miss detection. As a result, the adaptive filter weights are frozen mostly during the DT conditions. On the other hand, the other MSE curves diverge greatly during the DT condition, and can reach as high as 0 dB, because of higher rates of miss detection.

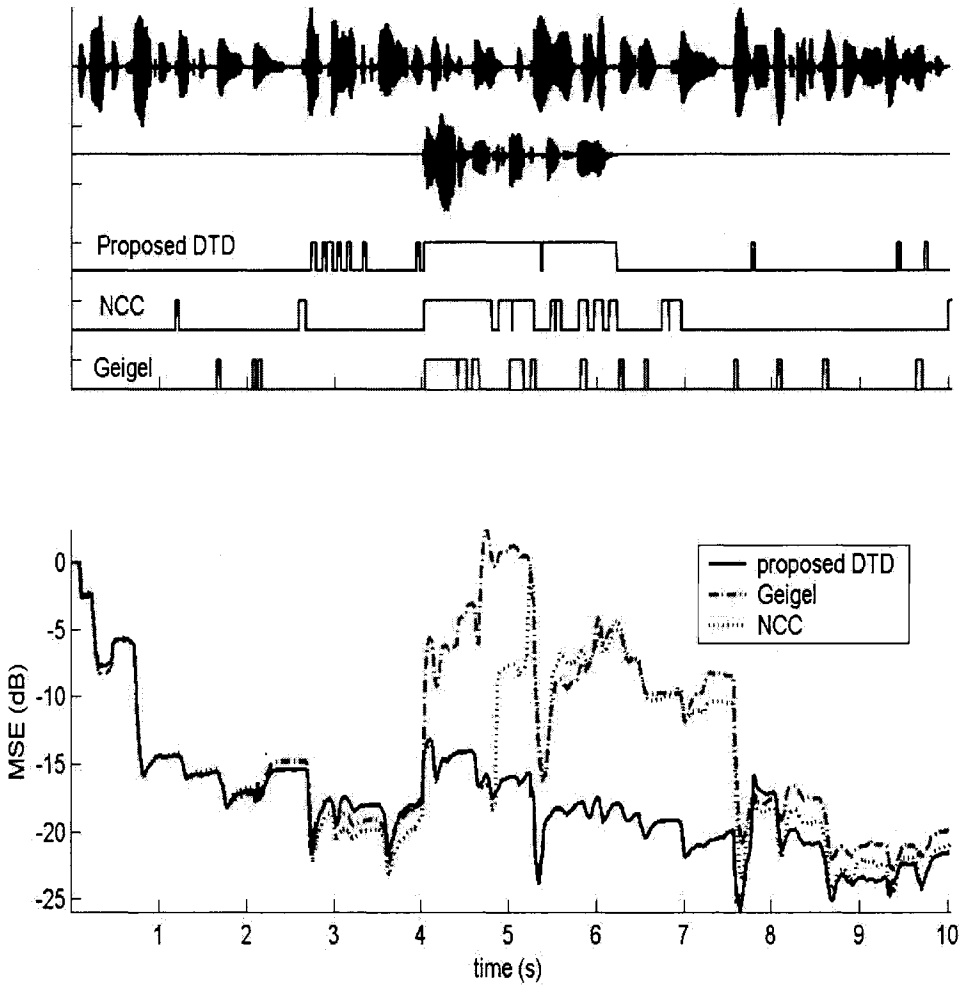

Figure 5-13: EERL and DTD Performance during DT Condition for $P_f = 0.05$, NER = 0 dB and ENNR = 30 dB: Top) far-end signal, DT signal, and DT decisions of the DTD algorithms; Bottom) normalized MSE performance of the DTD algorithms.

5.3.3.3 Echo Path Change Performance

This section presents and compares simulation results of the DTD algorithms under different conditions of EPCs. To compare the overall performance of the EC and DTD systems fairly in the presence of EPCs, all DTD algorithms are set up for the same given constraint of the P_m rate. This is done by applying a DT signal at the fourth second of the far-end signal, then the thresholds of the DTD algorithms are adjusted so that the measured miss detection rates are all approximately $P_m = 0.3$. The other conditions are set up the same as in the EERL results. In practice, a lower P_m rate should probably be used for a DTD algorithm because it would be able to allow better convergence of the adaptive filter. However, for simulations of the Geigel algorithm, setting a very small P_m rate would increase the false alarm rate so high that the adaptation of the filter is mostly halted. After setting the algorithms, the DT signal is then removed, and then an EPC is applied four seconds after the far-end signal starts. The normalized MSEs and the increases in false alarm are measured and compared.

Figure 5-14 shows the plots of the EERL and the false alarms of the algorithms under no EPC for a given constraint of $P_m = 0.3$, and for fixed NER = 0 dB and ENNR = 30 dB. In overall, the MSE performance of the proposed algorithm is a bit better than that of the NCC and the Geigel algorithms because of its lowest false alarm rate. For the given constraint of P_m, the initial false alarm rates for the algorithms under no EPC are recorded in Table 5-2.

	Proposed	**NCC**	**Geigel**
P_f	0.000	0.048	0.266

Table 5-2: Measured P_f for no EPC ($P_m = 0.3$, NER = 0 dB, and ENNR = 30 dB).

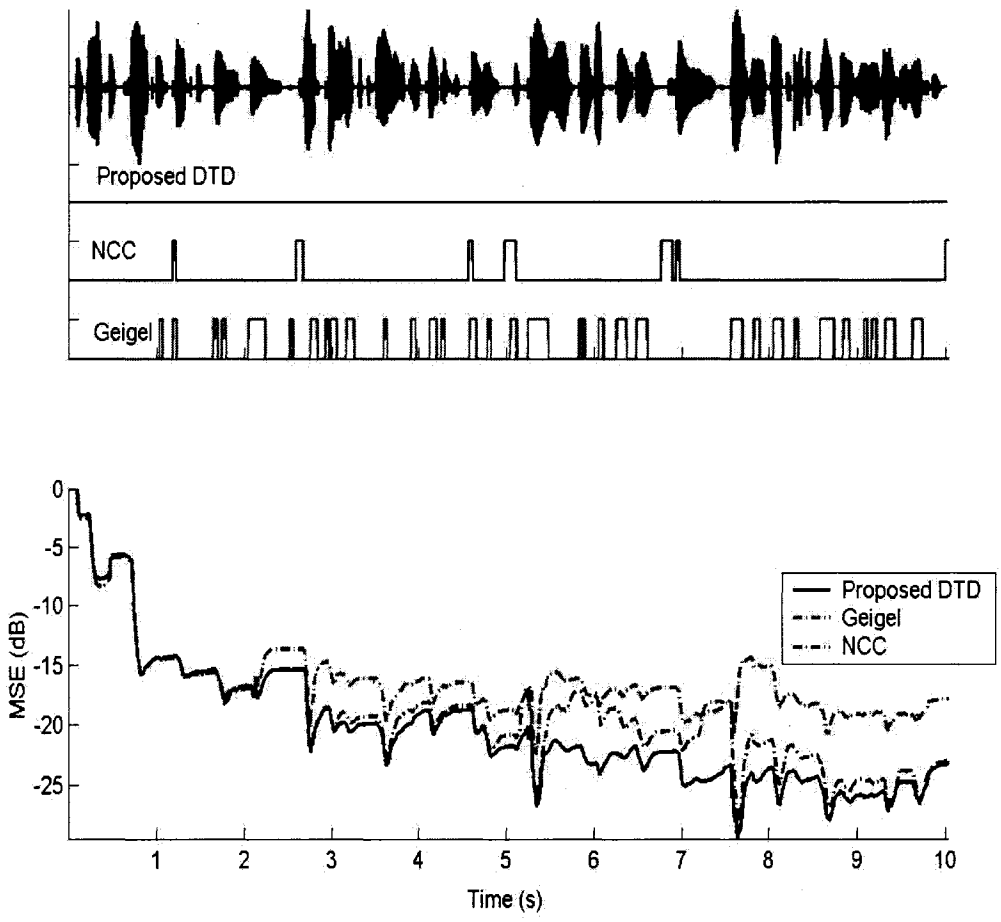

Figure 5-14: EERL and False Alarms under no EPC for P_m = 0.3: Top) far-end signal, false alarms of the algorithms; Bottom) normalized MSE performance.

Figure 5-15 shows the plots of the EERL and the false alarms of the algorithms under time-delayed EPC of 10 samples, i.e. $\delta = 10$. The measured false alarms of the algorithms are given in Table 5-3.

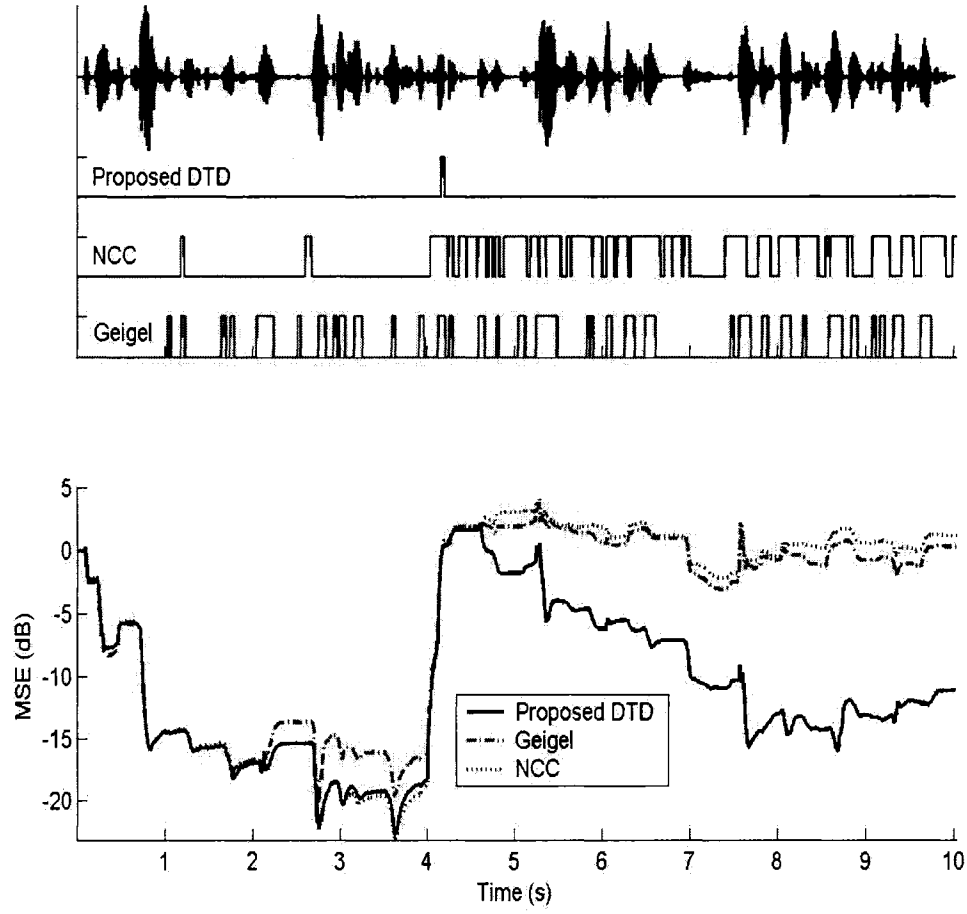

Figure 5-15: EERL and False Alarms under Time-Delayed EPC for P_m = 0.3: Top) far-end signal, false alarms of the algorithms; Bottom) normalized MSE performance.

	Proposed	NCC	Geigel
P_f	0.003	0.461	0.269

Table 5-3: Measured P_f for Time-Delayed EPC (P_m = 0.3, NER = 0 dB, ENNR = 30 dB).

It can be seen that the NCC algorithm is the most sensitive to the time delayed EPC, indicated by a significant increase in false alarms shown in Table 5-3. This can be explained by that, in computing its decision variable, the NCC algorithm uses the adaptive filter weights as an estimate for the echo path. When the EPC occurs at the end of the fourth second, the filter weights are therefore far from a good estimate of the echo path. As a result, there is a so big increase in false alarms that its adaptive filter halts and stops converging. On the other hand, the proposed algorithm has a very small increase in false alarms and the fastest convergence rate because of its lowest false alarm rate. The Geigel algorithm also has a small increase in false alarms although the convergence rate of the Geigel is still very low because of its high false alarm rate. In overall, the proposed algorithm, because of its low false alarm rate, still allows the adaptive filter to keep track of the EPC efficiently.

Figure 5-16 presents the results of the algorithms under time-advanced EPC of 10 samples, i.e. $\delta = 10$. The plots show similar behaviors to that with the time-delayed case although the false alarm rates for the Geigel and the proposed algorithms are a little bit higher. In overall, the proposed algorithm has the best MSE performance. The measured false alarm rates for the algorithms are recorded in Table 5-4.

	Proposed	**NCC**	**Geigel**
P_f	0.011	0.456	0.387

Table 5-4: Measured P_f for Time-Advanced EPC (P_m = 0.3, NER = 0 dB, and ENNR = 30 dB).

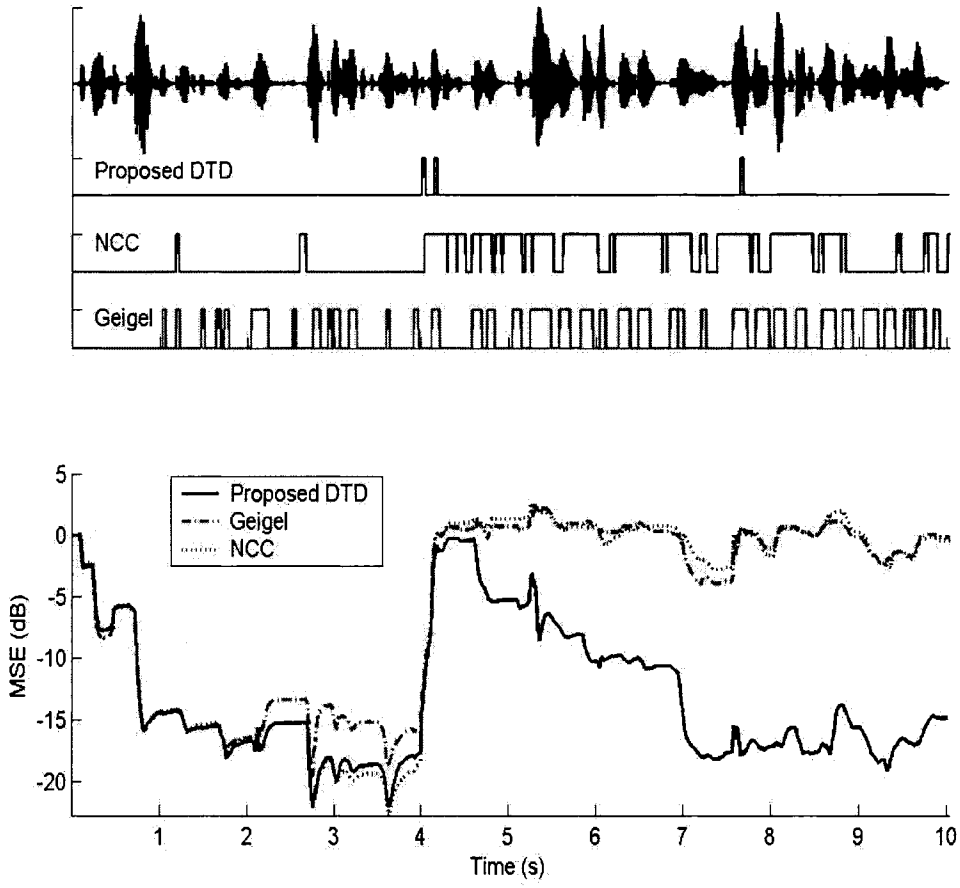

Figure 5-16: EERL and False Alarms under Time-Advanced EPC for P_m = 0.3: Top) far-end signal, false alarms of the algorithms; Bottom) normalized MSE performance.

Figure 5-17 presents the results of the algorithms under the echo path gain of 6 dB, i.e. $\alpha = 2$. The measured false alarm rates for the algorithms are presented in Table 5-5.

	Proposed	NCC	Geigel
P_f	0.007	0.043	0.444

Table 5-5: Measured P_f under 6 dB EP Gain (P_m = 0.3, NER = 0 dB, ENNR = 30 dB).

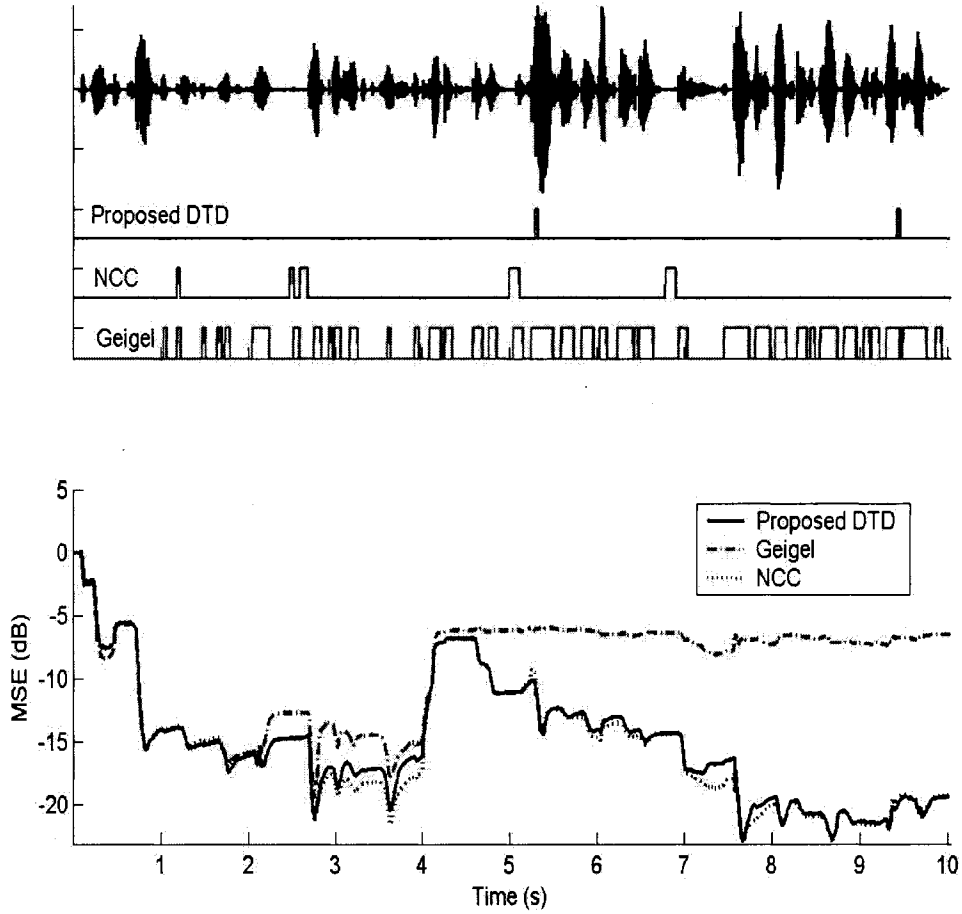

Figure 5-17: EERL and False Alarms under EP Gain for P_m = 0.3: Top) far-end signal, false alarms of the algorithms; Bottom) normalized MSE performance.

It can be seen that the Geigel algorithm is the most sensitive to the echo path gain, indicated by a significant increase in false alarms. This can be understood by that, in computing its decision variable, the Geigel algorithm compares the magnitude of the near-end input signal d(n), which consists of the echo and the background noise in this case, with

the magnitude of the far-end speech $x(n)$. When the echo path gain of 6 dB occurs at the end of the fourth second, the magnitude of the echo signal also increases by the same amount. As a result, there is an increase in false alarms; the adaptation of the filter therefore halts and stops converging. On the other hand, the proposed and the NCC algorithms have a very small or almost no increase in false alarms over the entire time after the EPC. In overall, the proposed and the NCC algorithms, because of their lower alarm rates, allow the adaptive filter to keep track of the EPC.

Figure 5-18 presents the results of the algorithms under echo path loss of 6 dB, i.e. $\alpha = 0.5$. The measured false alarm rates for the algorithms are given in Table 5-6.

	Proposed	**NCC**	**Geigel**
P_f	0.000	0.042	0.117

Table 5-6: Measured P_f for 6 dB EP Loss (P_m = 0.3, NER = 0 dB, and ENNR = 30 dB).

It can be observed that there are no increases in false alarms for the NCC and the proposed algorithms. On the other hand, there is a decrease in false alarms for the Geigel algorithm. This can be explained by the fact that, in computing its decision variable, the Geigel algorithm compares the magnitude of the near-end input signal $d(n)$, which consists of the echo and the background noise in this case, with the magnitude of the far-end speech $x(n)$. When the echo path loss of 6 dB occurs at the end of the fourth second, the magnitude of the echo signal also decreases by the same amount. As a result, there is a decrease in false alarms. The convergence rates of the NCC and the proposed algorithm are about the same for the condition of echo path loss, and better than that of the Geigel algorithm. The proposed

algorithm, however, still has the lowest amount of false alarms. In addition, a reduction in echo level improves the convergence rate of the Geigel algorithm because the false alarms are reduced under the echo path loss.

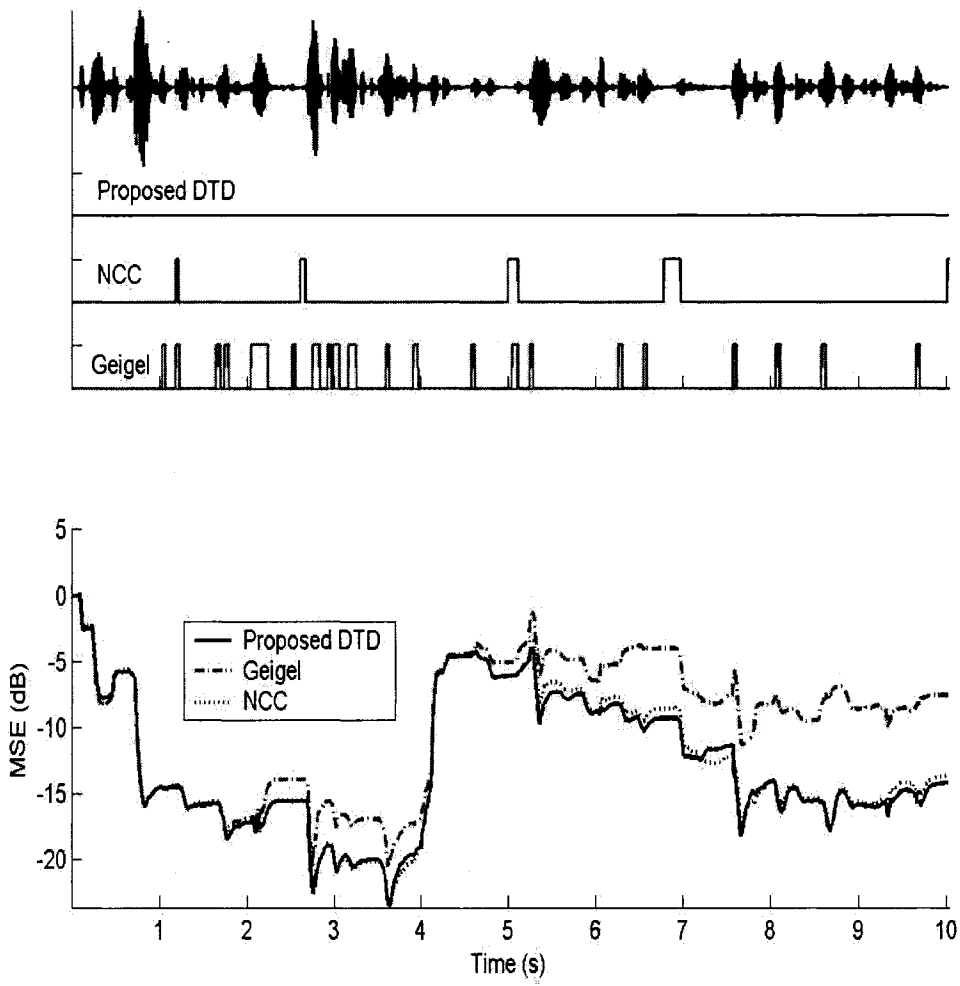

Figure 5-18: EERL and False Alarms under EP Loss for P_m = 0.3: Top) far-end signal, false alarms of the algorithms; Bottom) normalized MSE performance.

When comparing the DTD algorithms under different EPCs, the constraint of $P_m = 0.3$ was initially set up for the algorithms. This value was selected in favor of the Geigel algorithm because a lower rate of P_m would result in a much higher false alarm rate for the algorithm that its adaptive filter would freeze mostly during the entire length of simulations. However, setting up the constraint P_m as high as 0.3 does not take advantage of the very low P_f characteristic of the proposed DTD algorithm. That means that the proposed algorithm can be set with a lower P_m rate, such as 0.135, for which the false alarm rate $P_f = 0.0$ can still be achieved. Figure 5-19 shows the results of the proposed algorithm under the same EPCs and with the same setups as that with the previous EPC simulations, except that the initial P_m is set at 0.135. The measured false alarm rates of the proposed algorithm under all four conditions of EPC are given in the Table 5-7. In general, the same behaviors of the proposed algorithm can be observed as that with the other settings previously.

Proposed Algorithm	Time Delay	Time Advance	Echo Path Gain	Echo Path Loss
P_f	0.004	0.011	0.011	0.000

Table 5-7: Measured P_f of the Proposed Algorithm under Four EPCs (P_m = 0.135, NER = 0 dB, and ENNR = 30 dB).

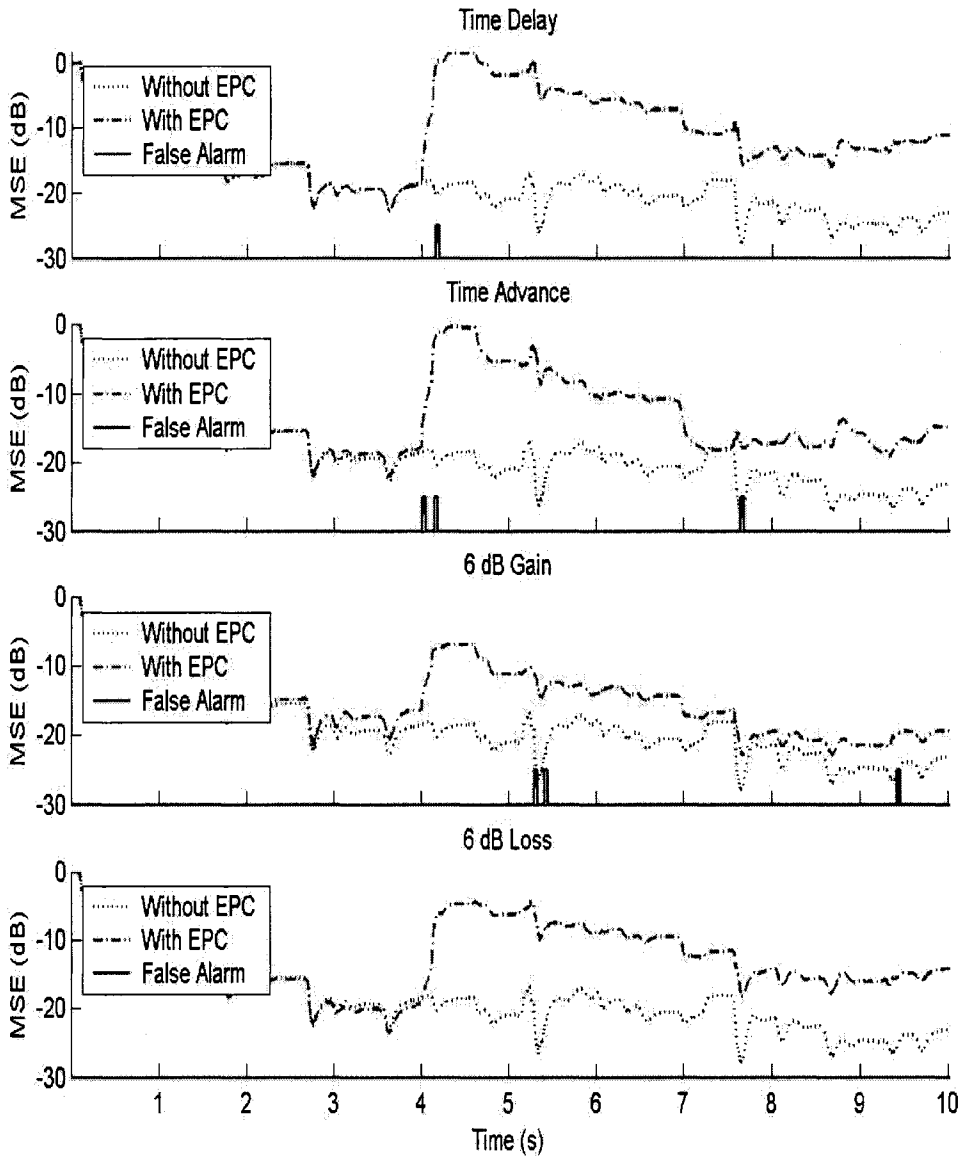

Figure 5-19: EERL and False Alarms of the Proposed Algorithms under Four EPCs for $P_m = 0.135$.

5.3.3.4 Effects of Frame Length

Due to the block processing nature, the proposed algorithm can possibly have a response that is slower than the actual onset of a DT condition. A shorter frame length is therefore preferred as it reduces the delay of the DTD decision. On the other hand, a longer frame length would reduce the overhead computations per sample for the proposed algorithm. This section evaluates the effects of frame length on the P_m rate of the proposed DTD algorithm. Figure 5-20 shows the plot of the P_m rate as a function of the frame length of samples used by the proposed algorithm. The same set of far-end and near-end speeches as that used in section 5.3.3.1 are used to compute the average values for the plot. The other parameters are set as follows: P_f rate is set at 0.1, the ENNR is set at 40 dB, and the NER is set in the range of [0, -10, and -20]. It can be observed from the curves that, for the frame length in the range of [64-512], the P_m rates are roughly about the same, whereas the P_m rates increase significantly for the frame lengths greater than 512 samples. This can be explained by the fact that a longer frame length carries many spectral components, and therefore may contain more unmasked spectral components. Consequently, a smaller number of spectral holes are created; and it is less likely that non-zero spectral components of the near-end speech align with the spectral holes.

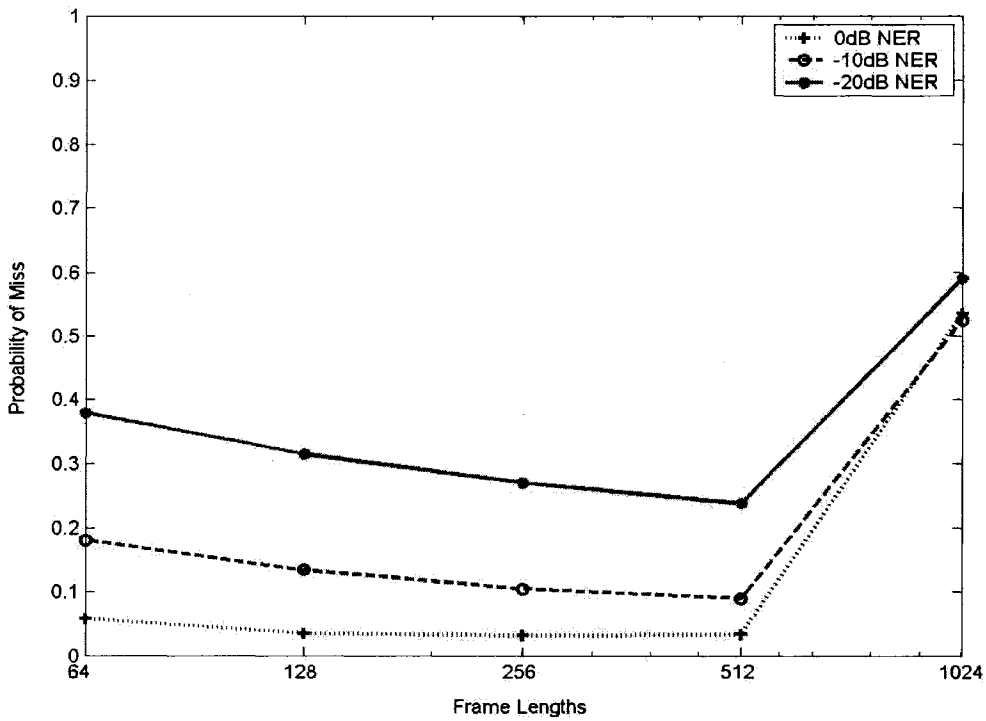

Figure 5-20: P_m versus Frame Lengths of Samples for P_f = 0.1, ENNR = 40 dB, and NER = [0, -10, and -20] dB.

5.3.3.5 Computational Complexity

In this section, the computational complexities of the considered DTD algorithms are analyzed and compared. The determining factor for computational complexities of the DTD algorithms depends on the parameters used in the algorithms such as the filter length L, the frame length N, and the FFT length M. The computational complexity can be measured in a number of different ways. One method is to count the number of multiplications and additions per time sample as a measure of the computational complexity. For simplicity, the

comparison operations and the memory requirements used in the algorithms are omitted here and are not quantified by the complexity measures, although these variables should be thoroughly considered when implementing an actual DTD system. However, a measurement of the number of multiplications and additions should be good enough to indicate the bulk part of the computational complexity of an algorithm. In addition, a division, a root, a logarithm, or an exponential operation is simply counted as a multiplication operation; and a subtraction is counted as an addition operation. It is assumed that all signals are real-valued, a complex addition requires two real additions, and a complex multiplication requires four real multiplications and two real additions. Furthermore, the authors in [SOR87] show that an M-point FFT or IFFT routine could require $\left(\frac{M}{2}\log_2 M - \frac{5M}{4}\right)$ of real multiplications and $\left(\frac{3M}{2}\log_2 M - \frac{M}{4} - 4\right)$ real additions.

The proposed algorithm processes data on a frame-by-frame basis of N samples. Since consecutive frames are 50% overlapped, there are $N/2$ new samples for each processed frame or block. To obtain the computational complexity per sample, the total number of real multiplications and additions per frame is therefore divided by $N/2$. Table 5-8 shows the estimated computational complexity per sample for the proposed DTD algorithm.

	Multiplications/Sample	**Additions/Sample**
Equation 5.6: Window Weighting	$N\left(\frac{2}{N}\right) = 2$	0
Equation 5.7: M-point FFT	$\left(\frac{M}{2}\log_2 M - \frac{5M}{4}\right)\left(\frac{2}{N}\right)$	$\left(\frac{3M}{2}\log_2 M - \frac{M}{4} - 4\right)\left(\frac{2}{N}\right)$

	$=\frac{M}{N}\log_2 M-\frac{5M}{2N}$	$=\frac{3M}{N}\log_2 M-\frac{M}{2N}-\frac{8}{N}$
Equation 5.8: Symmetrical Power Spectrum	$\left(4\frac{M}{2}\right)\left(\frac{2}{N}\right)=\frac{4M}{N}$	$\left(2\frac{M}{2}\right)\left(\frac{2}{N}\right)=\frac{2M}{N}$
Equation 5.9: Frequency to Bark (18 bands for 8 kHz)	0	$18\left(\frac{M}{18}-1\right)\left(\frac{2}{N}\right)=\frac{2M}{N}-\frac{36}{N}$
Equation 5.10: Matrix Multiplication of Spreading Function	$18^2\left(\frac{2}{N}\right)=\frac{648}{N}$	$(18)(17)\left(\frac{2}{N}\right)=\frac{612}{N}$
Equation 5.11: Spectral Flatness Measure	$\left(\frac{M}{2}-1+1+2\right)\left(\frac{2}{N}\right)=\frac{M}{N}+\frac{4}{N}$	$\left(\frac{M}{2}-1\right)\left(\frac{2}{N}\right)=\frac{M}{N}-\frac{2}{N}$
Equation 5.12: Coefficient of Tonality	$1\left(\frac{2}{N}\right)=\frac{2}{N}$	0
Equation 5.13: Threshold Offset	$(18)(2)\left(\frac{2}{N}\right)=\frac{72}{N}$	$(18)(3)\left(\frac{2}{N}\right)=\frac{108}{N}$
Equation 5.14: Spread Threshold	$(18)(3)\left(\frac{2}{N}\right)=\frac{108}{N}$	$(18)(1)\left(\frac{2}{N}\right)=\frac{36}{N}$
Equation 5.15: Normalized Threshold	$(18)(1)\left(\frac{2}{N}\right)=\frac{36}{N}$	0
Step 12: Reconstruction of Masked Signal Frame (IFFT + Overlap Add)	$\left(\frac{M}{2}\log_2 M-\frac{5M}{4}\right)\left(\frac{2}{N}\right)$ $=\frac{M}{N}\log_2 M-\frac{5M}{2N}$	$\left(\left(\frac{3M}{2}\log_2 M-\frac{M}{4}-4\right)+\frac{N}{2}\right)\left(\frac{2}{N}\right)$ $=\frac{3M}{N}\log_2 M-\frac{M}{2N}-\frac{8}{N}+1$
Equation 5.20: Spectral Hole Monitor (Window + FFT + Spectrum)	$\left(N+\left(\frac{M}{2}\log_2 M-\frac{5M}{4}\right)+4\frac{M}{2}\right)\left(\frac{2}{N}\right)$ $=2+\frac{M}{N}\log_2 M-\frac{5M}{2N}+\frac{4M}{N}$	$\left(0+\left(\frac{3M}{2}\log_2 M-\frac{M}{4}-4\right)+2\frac{M}{2}\right)\left(\frac{2}{N}\right)$ $=\frac{3M}{N}\log_2 M-\frac{M}{2N}-\frac{8}{N}+\frac{2M}{N}$
Equation 5.22: Smear Effects (Window + FFT + Spectrum)	$\left(N+\left(\frac{M}{2}\log_2 M-\frac{5M}{4}\right)+4\frac{M}{2}\right)\left(\frac{2}{N}\right)$ $=2+\frac{M}{N}\log_2 M-\frac{5M}{2N}+\frac{4M}{N}$	$\left(0+\left(\frac{3M}{2}\log_2 M-\frac{M}{4}-4\right)+2\frac{M}{2}\right)\left(\frac{2}{N}\right)$ $=\frac{3M}{N}\log_2 M-\frac{M}{2N}-\frac{8}{N}+\frac{2M}{N}$
Total:	$\frac{4M}{N}\log_2 M+\frac{3M+870}{N}+6$	$\frac{12M}{N}\log_2 M+\frac{7M+718}{N}+1$

Table 5-8: Estimated Computational Complexity of the Proposed DTD Algorithm.

Table 5-9 shows the estimated computation complexities per sample for the Geigel algorithm.

	Multiplications/Sample	Additions/Sample
Equation 5.24: Decision Variable	1	0
Total:	1	0

Table 5-9: Estimated Computational Complexity of the Geigel Algorithm.

Table 5-10 shows the estimated computation complexities per sample for the NCC algorithm.

	Multiplications/Sample	Additions/Sample
Equation 5.25: Decision Variable	$L+1+1=L+2$	$L-1$
Equation 5.26: Cross Correlation Estimation	$L+L+1=2L+1$	1
Equation 5.27: Variance Estimation	3	1
Total	$3L+6$	$L+1$

Table 5-10: Estimated Computational Complexity of the NCC Algorithm.

It looks obvious from the tables that the Geigel algorithm has the least computational complexity with its complexity order of zero, i.e. $\mathbf{O}_{\text{Geigel}}(0)$. The NCC algorithm has a linear complexity order, or $\mathbf{O}_{\text{NCC}}(L)$, which is dependent of the length of the adaptive filter. For a hands-free application, if L is as large as 1024-2048 or more, the computational complexity of the NCC algorithm can be expensive. For the proposed algorithm, its complexity order depends on the parameters M and N. Selection of a value for frame length N can be based on the perceptual quality of the output speech from the Johnston's model and on the overall

performance of the EC system. In addition, the FFT length is typically chosen as $M = 2N$. Therefore, the complexity order for the proposed algorithm can be expressed as $\mathbf{O}_{proposed}(\frac{M}{N}\log_2 M) \approx \mathbf{O}_{proposed}(\log_2 M)$. For the typical values $M = 256$, $N = 128$, and $L = 1024$ used in the simulations, the proposed algorithm would require approximately 107 multiplications and 212 additions per sample, the NCC would require 3078 multiplications and 1025 additions per sample, and the Geigel algorithm would require only one multiplication per sample. The Geigel algorithm is the computationally simplest but its performance suffers greatly as a trade-off. The proposed algorithm, despite of its complicated steps of process, has a better computational complexity than the NCC algorithm. This is because the proposed algorithm processes data in the frequency domain with parameter lengths M and N, which are much smaller than the adaptive filter length L used by the NCC algorithm. For the overall comparison of the DTD algorithms, the proposed algorithm has the best performance and a moderate computational complexity.

5.4 Summary

A novel approach to double talk detection that exploits the frequency masking properties of the human auditory system was presented. The proposed DTD algorithm was shown to have a superior overall performance in comparing with the Geigel and the NCC algorithms. Simulation results with speeches showed the best DT detection and echo return loss performance during DT conditions for the proposed algorithm. The proposed algorithm was also found to show little sensitivity to some conditions of EPCs and it outperformed the other

algorithms in overall because of its lowest false alarm rate. Analysis of the DTD algorithms showed that the proposed algorithm has a moderate computational complexity compared with the other algorithms.

Chapter 6
Conclusions and Future Work

In this thesis, a survey for solutions to the problem of DTD in the existing literature was presented. The survey provided common basics for a generic DT detector, and summarized different DTD schemes. It further classified them into two categories: energy level based and cross-correlation based, from the different signal processing techniques used. Challenges and problems in designing a simple but effective DT detector were identified. In addition, the thesis reviewed fundamental concepts in EC and psychoacoustics as foundations for the development and analysis of the proposed DTD algorithm.

The most significant contribution of the thesis was the proposal of a DTD algorithm for a single-channel EC application using a psychoacoustic auditory model. The model exploits the frequency masking properties of the human auditory system. It performs an analysis of the far-end speech signal to compute perceptual thresholds, and then removes spectral components below the perceptual thresholds to create spectral holes - without affecting the perceptual quality of the signal. DT condition is detected by monitoring the spectral levels at the created holes in the received near-end signal. Furthermore, the smearing effect resulting from the frequency masking process and the filtering of the echo path was addressed; a method to combat the effect was also proposed.

The proposed algorithm illustrated, via simulations, its effectiveness in detecting DT conditions. Simulation results with speeches demonstrated the superior performance of the

proposed algorithm in comparison with the Geigel and the NCC algorithms. The proposed algorithm also showed little or no sensitivity to typical conditions of EPCs. In addition, analysis and comparison of the DTD algorithms showed a moderate computational complexity for the proposed algorithm. In overall, the proposed DTD algorithm showed its feasibility and potential usage for single-channel EC application.

The investigation of the proposed DTD algorithm in this thesis was intended to find its fundamental limits before deciding whether it is worthwhile to pursue further work. In addition, some ideas, which appeared throughout the course of this thesis research as potential improvements to the proposed algorithm, were not followed because they would be beyond the scope of this thesis. For those reasons, the thesis is concluded with some possibilities for further future work as follow.

Since the proposed DTD algorithm was able to perform very effectively under a simulated environment, it thus seems to be interesting to validate how the algorithm would work in a practical EC system. In this aspect, one has to take into consideration implementation related issues such as system delays, non-linearity, and time varying echo path, requirements of hardware, computer power and memory etc, making it a long way from the algorithmic concept towards an actual product. As a next step, performance and assumptions about the proposed algorithm can be experimentally measured and validated. In addition, EC tests can be done to see how well the system conforms to the appropriate ITU-T recommendations. Furthermore, formal objective and subjective listening tests can be carried out to evaluate the

output speech quality of the Johnston's psychoacoustic auditory model as well as the overall speech quality of the EC system.

Performance of the proposed DTD algorithm can also be evaluated in integration with another type of EC structure such as the frequency domain adaptive filtering. Some good candidates are the multi-delay frequency domain adaptive filtering structures proposed in [PRA94], [MOU95], [SOO90], [GAY00], and [BENE01 and 03]. These structures are generally more efficient computationally than those in the time domain while allowing some degree of freedom in block delay. In the complexity aspect, the proposed algorithm would benefit further computational reductions from these structures, in which frequency-domain signals would be directly available for the processing of DTD. In the performance aspect, the delay nature of these structures would compensate for the DTD delay so the overall effect is a reduced response time of the proposed DTD algorithm.

At last, the proposed algorithm can be investigated further for stereophonic or multi-channel EC applications. Some good introductory information in this area can be found in [GAY00] and [BENE01].

Appendix A

Critical Band Numbers

The following table, extracted from [ZWI90], page 142, shows the critical band numbers and their corresponding frequency limits.

Band Number	Lower Edge F_L (Hz)	Centre F_C (Hz)	Upper Edge F_U (Hz)
0	0	50	100
1	100	150	200
2	200	250	300
3	300	350	400
4	400	450	510
5	510	570	630
6	630	700	770
7	770	840	920
8	920	1000	1080
9	1080	1170	1270
10	1270	1370	1480
11	1480	1600	1720
12	1720	1850	2000
13	2000	2150	2320
14	2320	2500	2700
15	2700	2900	3150
16	3150	3400	3700
17	3700	4000	4400
18	4400	4800	5300
19	5300	5800	6400
20	6400	7000	7700
21	7700	8500	9500
22	9500	10500	12000
23	12000	13500	15500
24	15500		

Appendix B
Echo Path and Speeches in Simulations

The acoustic echo path used in the simulations of Chapter 5 is a measured impulse response of a furnished conference room (approximately 12x6 meters). The measurement was done by applying band-limited white noise at the loudspeaker, and then recording the microphone signal to a digital audio tape. These signals were then used to train a 1200-tap NLMS adaptive filter. The impulse response was obtained by time averaging the filter weights after convergence. In Figure B - 1 plot (a) shows the impulse response in time domain and plot (b) shows the magnitude response of the echo path.

Speeches, at 8 kHz sampling rate, used in simulations were taken from the Harvard sentences speech database, which was prepared by the Acoustics and Signal Processing group at the NRC lab. The far-end speech signal is from a female and is about 10 seconds long. The near-end speech signals, which simulate DT conditions, are from different females and males; each is about 2 seconds long. The time waveforms and the power spectral densities of the speeches are shown in the following Figure B - 2, Figure B - 3, and Figure B - 4.

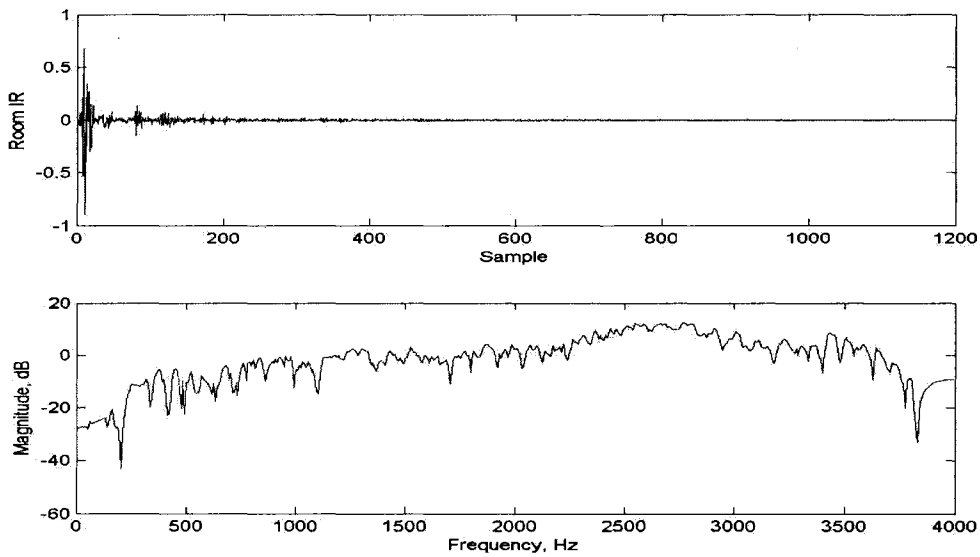

Figure B - 1: Acoustic Echo Path.

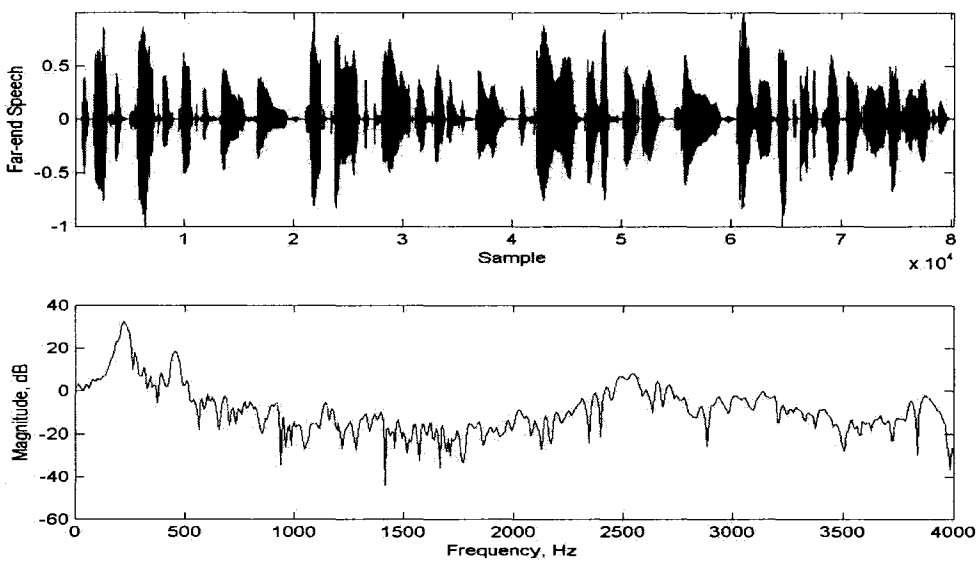

Figure B - 2: Far-End Signal

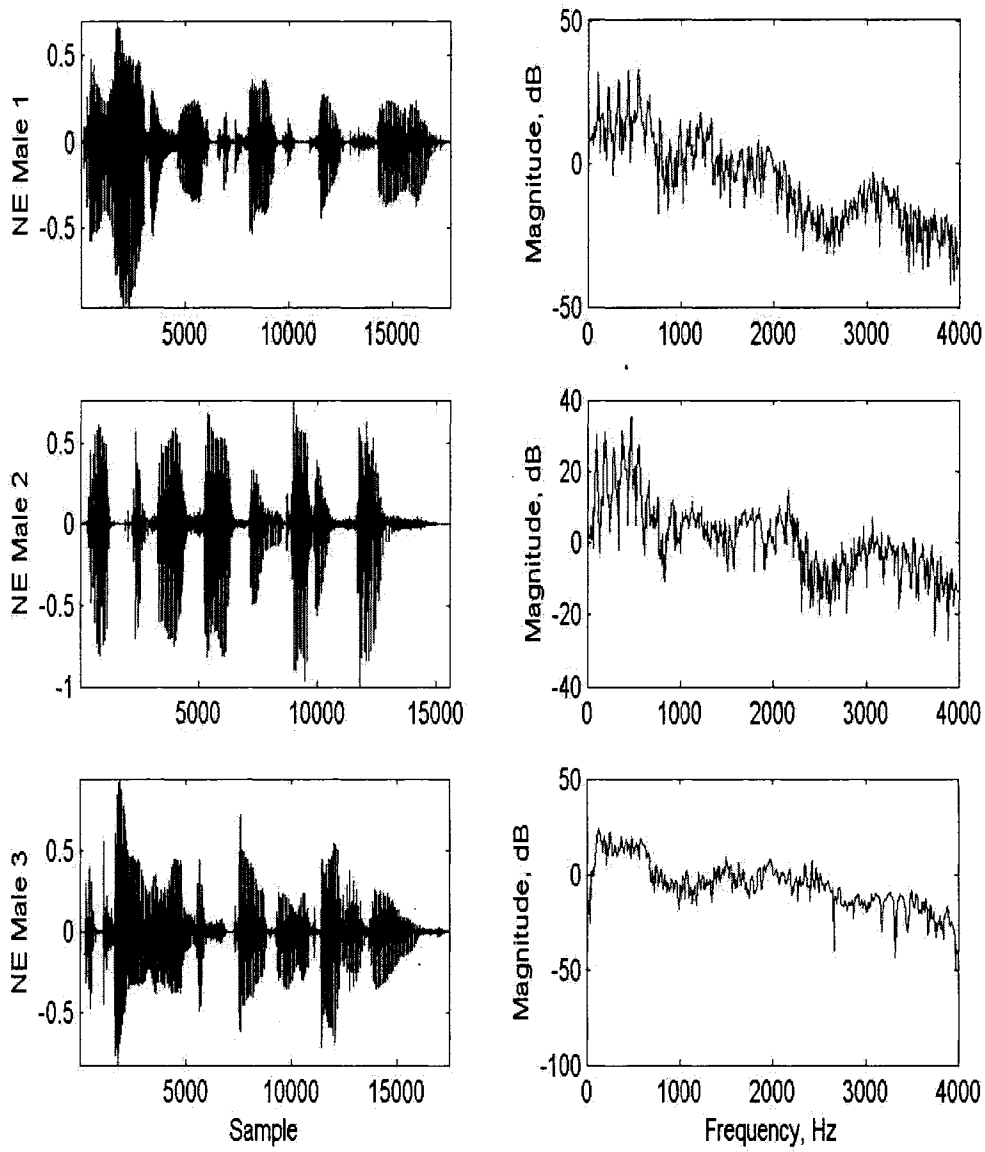

Figure B - 3: Male DT Signals

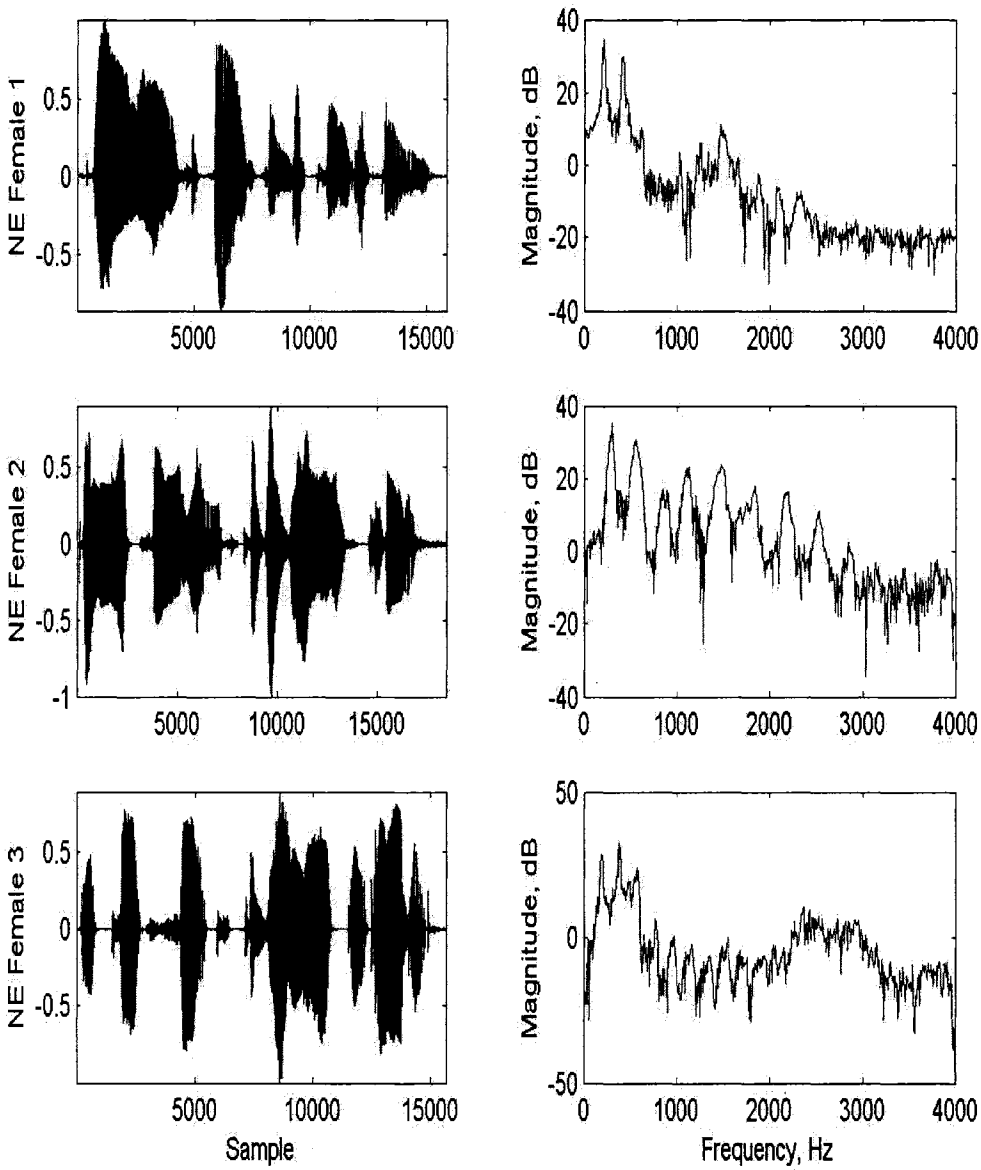

Figure B - 4: Female DT Signals.

Bibliography

[AHG04] P. Ahgren. *On System Identification and Acoustic Echo Cancellation,* Ph. D. Thesis, Uppsala University, Uppsala, Sweden, 2004.

[BEC66] F.K. Becker, and H.R. Rudin. "Application of automatic transversal filters to the problem of echo suppression," *Bell System Technical Journal,* vol. 45, pp. 1847-1850, 1966.

[BEND01] Y. Bendel et al. "Delayless frequency domain acoustic echo cancellation," *IEEE Transactions on Speech and Audio Processing,* vol. 9, no. 5, pp. 589-597, July 2001.

[BENE00] J. Benesty, D.R. Morgan, and J.H. Cho. "A new class of doubletalk detectors based on cross-correlation," *IEEE Transactions on Speech Audio Processing,* vol. 7, pp. 168-172, March 2000.

[BENE01] J. Benesty et al. *Advances in Network and Acoustic Echo Cancellation,* Springer-Verlag, 2001.

[BENE03] J. Benesty, and T. Gansler. "A multidelay double-talk detector combined with the MDF adaptive filter," *EURASIP Journal on Applied Signal Processing,* vol. 11, pp. 1056-1063, 2003.

[BRE99] C. Breining et al. "Acoustic echo control – an application of very high order adaptive filters," *IEEE Signal Processing Magazine,* pp. 42-69, July 1999.

[CAV02] C. Cave. *Perceptual Modeling for Low-Rate Audio Coding*, Master Thesis, McGill University, Montreal, Canada, 2002.

[CHO99] J.H. Cho, D.R. Morgan, and J. Benesty. "An objective technique for evaluating doubletalk detectors in acoustic echo cancellers," *IEEE Transactions on Speech and Audio Processing,* vol. 7, pp. 718-724, November 1999.

[CHU96] H. Chunfeng, Z. Zhengang, and W. Dejun. "On the algorithm and implementation of an acoustic echo canceller," *Proceedings of International Conference on Signal Processing,* vol. 1, pp. 657-660, October 1996.

[COX02] I.J. Cox, M.L. Miller, and J.A. Bloom. *Digital Watermarking,* Morgan Kaufmann Publishers, 2002.

[CRE99] T. Creasy. *Algorithms for Acoustic Echo Cancellation in the Presence of Double Talk,* Master Thesis, University of Ottawa, Ottawa, Canada, 1999.

[CRE00] T. Creasy, and T. Aboulnasr. "A projection-correlation algorithm for acoustic echo cancellation in the presence of double talk," *IEEE Conference on Acoustics, Speech, and Signal Processing,* vol. 1, pp. 436-439, 2000.

[DIN00a] H. Ding. "A stable fast affine projection adaptation algorithm suitable for low cost processors," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 360-363, 2000.

[DIN00b] H. Ding. "Circuit and method of double talk detection for use in handsfree telephony terminals," *US Patent #6049606,* April 2000.

[DIN01] H. Ding. "Method of distinguishing between echo change and double talk conditions in an echo canceller," *US Patent #6226380,* May 2001.

[DIN04] H. Ding. "Double-talk detection schemes for echo cancellation," *Proceedings of the Acoustics Week in Canada*, vol. 32, no. 3, pp. 146-147, October 2004.

[DOH97] J.F. Doherty, and R. Porayath. "A robust echo canceler for acoustic environments," *IEEE Transactions on Circuits and Systems II,* vol. 44, pp. 389-396, May 1997.

[DUT78] D.L. Duttweiller. "A twelve-channel digital echo canceller," IEEE *Transactions on Communications,* vol. 26, pp. 647-653, May 1978.

[DYB03] R.A. Dyba et al. "Motorola packet telephony echo cancellation solutions," *White Paper Series,* Freescale Semiconductor Inc., July 2003.

[DYB04] R.A. Dyba et al. "Network echo cancellers and Freescale solutions using the StarCore SC140 core," *Application Note*, Freescale Semiconductor, 2004.

[ENE98] K. Eneman, and M. Moonen. "Real implementation of an acoustic echo canceller," in *Proceedings of the COST#254 Workshop on Intelligent Communications,* Italy, 1998.

[ENE03] K. Eneman, and M. Moonen. "Iterated partitioned block frequency-domain adaptive filtering for acoustic echo cancellation," *IEEE Transactions on Speech and Audio Processing,* vol. 11, no. 2, pp. 143-158, 2003.

[FLE40] H. Fletcher. "Auditory patterns," *Revs. Modern Physics*, vol. 12, pp. 47-66, January 1940.

[GAN96] T. Gansler et al. "A double-talk detector based on coherence," *IEEE Transactions in Communications,* vol. 44, pp. 1421-1427, November 1996.

[GAN00] T. Gansler et al. "Double-talk robust fast converging algorithms for network echo cancellation," *IEEE Transactions on Signal Processing Audio and Acoustics,* vol. 8, no. 6, pp. 656-663, November 2000.

[GAN01] T. Gansler, and J. Benesty. "A frequency-domain double-talk detector based on a normalized cross-correlation vector," *Signal Processing Journal,* vol. 81, no. 8, pp. 1783-1787, August 2001.

[GAR99] R.A. Garcia. *Digital Watermarking of Audio Signal Using a Psychoacoustic Auditory Model and Spread Spectrum Theory,* Master Thesis, University of Miami, Coral Gables, Florida, 1999.

[GAY00] S.L. Gay, and J. Benesty. *Acoustic Signal Processing for Telecommunication,* Kluwer International Series in Engineering and Computer Science, 2000.

[GOR05] J.D. Gordy, and R.A. Goubran, "A low-complexity doubletalk detector for acoustic echo cancellers in packet-based telephony," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,* pp. 74-77, October 2005.

[GHO00] K. Ghose, and V. Reddy. "A double-talk detector for acoustic echo cancellation applications," *Signal Processing Journal,* vol. 80, no. 8, pp. 1459-1467, August 2000.

[GOU64] R. Gould. "Echo suppression and time delay, a subjective evaluation," *IEEE Transactions on Communications,* vol. 12, pp. 74-82, 1964.

[GUS02] S. Gustafsson et al. "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Transactions on Speech and Audio Processing,* vol. 10, no. 5, pp. 245-256, July 2002.

[HAY96] S. Haykin. *Adaptive Filter Theory,* 3rd edition, Prentice-Hall, 1996.

[HAR00] T. Harley, and S. Leese. "Acoustic echo canceller," *US Patent #6091813*, July 2000.

[HIR03] A. Hirano, K. Nakayama, and S. Ushimaru. "Double-talk resistant acoustic echo canceller with double filters," *Proceedings of 2003 IEEE International Symposium on Intelligent Signal Processing and Communication Systems,* pp. 367-370, December 2003.

[HUA00] Y. Huang. *Effects of Vocoder Distortion and Packet Loss on Network Echo Cancellation,* Master Thesis, Carleton University, Ottawa, Canada 2000.

[ITU93a] ITU-T Recommendation G.165. *Echo Cancellers*, International Telecommunication Union, 1993.

[ITU93b] ITU-T Recommendationon G.167. *Acoustic Echo Controllers,* International Telecommunication Union, 1993.

[ITU97] ITU-T Recommendation G.169. *Digital Network Echo Cancellers,* International Telecommunication Union, 1997.

[JIR00] J. Jiri. *Digital Signal Filtering, Analysis and Restoration,* Institution of Electrical Engineers, 2000.

[JIA03] T. Jia et al. "Subband doubletalk detector for acoustic echo cancellation systems," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 604-607, 2003.

[JOH88] J.D. Johnston. "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications,* vol. 6, no 2, pp 314-323, February 1988.

[JUN02] H.K. Jung, N.S. Kim, and T. Kim. "A new double-talk detector using echo path estimation," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1897-1900, 2002.

[KIM01] S.H. Kim et al. "Performance improvement of double-talk detection algorithm in the acoustic echo canceller," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 5, pp. 3249-3252, 2001.

[KOS02] B. Kosanavic. "Echo Cancellation Part 1 and 2," *www.commsdesign.com*, April 2002.

[LAH03] M. Lahdekorpi et al. "Perceptual irrelevancy removal in narrowband speech coding," *Proceedings of 2003-EuroSpeech,* pp. 1081-1084, Geneva, 2003.

[LAH03] M. Lahdekorpi. *Perceptual Irrelevancy Removal in Narrowband Speech Coding,* Master Thesis, Tampere University of Technology, Tampere, Finland, 2003.

[LIN89] C.S. Lindquist. *Adaptive and Digital Signal Processing with Digital Filtering Applications,* Steward & Sons, 1989.

[LIU01] J. Liu. "A novel adaptation scheme in the NLMS algorithm for echo cancellation," *IEEE Signal Processing Letters,* vol. 8, no. 1, pp. 20-22, January 2001.

[MA03] N. Ma, M. Bouchard, and R. Goubran. "A perceptual Kalman filtering-based approach for speech enhancement," *Proceedings of IEEE-EURASIP Seventh*

International Symposium on Signal Processing and its Applications, vol. 1, pp 373-376, France, July 2003.

[MA04] N. Ma, M. Bouchard, and R. Goubran. "Frequency and time domain auditory masking threshold constrained Kalman filter for speech enhancement," *Proceedings of IEEE Seventh International Conference on Signal Processing,* vol. 3, pp. 2659-2662, China, August 2004.

[MA05] N. Ma, M. Bouchard, and R. Goubran. "A wavelet Kalman filter with perceptual masking for speech enhancement in colored noise," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 149-152, March 2005.

[MA06] N. Ma, M. Bouchard, and R. Goubran. "Speech enhancement using a perceptually masking threshold constrained Kalman filter and its heuristic implementations," *IEEE Transactions on Audio, Speech and Language Processing,* vol. 14, no. 1, pp. 19-32, January 2006.

[MUR90] K. Murano, S. Unagami, and F. Amano. "Echo cancellation and applications," *IEEE Communications Magazine*, vol. 28, no. 1, pp. 49-55, January 1990.

[MOU95] E. Moulines, O.A. Amrane, and Y Genier. "The generalized multidelay adaptive filter: structure and convergence analysis," *IEEE Transactions on Signal Processing,* vol. 43, pp. 14-28, 1995.

[OCH77] K. Ochiai, T. Araseki, and T. Ogihara. "Echo canceler with two echo path models," *IEEE Transactions on Communications*, vol. 25, no. 6, pp. 589-595, June 1977.

[OH97] S.S. Oh. "Method and apparatus for detecting a near-end speech signal," *US Patent #5633936,* May 1997.

[OKU99] T. Okuno, M. Fukushima, and M. Tohyama. "Adaptive cross-spectral technique for acoustic echo cancellation," *IEICE Transactions on*

Fundamentals of Electronics, Communications and Computer Sciences, vol. E82-A, no. 4, pp. 634-439, April 1999.

[OPP89] A.V. Oppenheim. *Discrete-Time Signal Processing,* Prentice-Hall, 1989.

[PAI00] T. Painter, and A. Spanias. "Perceptual Coding of Digital Audio," *Proceedings of the IEEE,* vol. 88, no. 4, pp. 451-513, April 2000.

[PAR90] S. Park, and G. Hillman. "A novel implementation of acoustic and electrical echo cancellers for full-duplex teleconference phone," *IEEE Conference on Signals, Systems and Computers,* vol. 1, pp. 47-50, November 1990.

[PAR02] S.J. Park et al. "Integrated echo and noise canceler for hands-free applications," *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing,* vol. 49, no. 3, pp. 188-195, March 2002.

[PRA94] J. Prado, and E. Moulines. "Frequency domain adaptive filtering with applications to acoustic echo cancellation," *Annales des Telecommunications,* vol. 49, no. 7-8, pp. 414-428, 1994.

[PRO88] J.G. Proakis, and D.G. Manolakis. *Introduction to Digital Signal Processing,* Macmillan, 1988.

[PRO01] J.G. Proakis. *Digital Communications,* 4th edition, McGraw Hill, 2001.

[RAB78] L.R. Rabiner, and R.W. Schafer. *Digital Processing of Speech Signals,* Prentice-Hall, 1978.

[SCH79] M.R. Schroeder, B.S. Atal, and J.L. Hall. "Optimizing digital speech coders by exploiting masking properties of the human ear," *Journal of the Acoustical Society of America,* vol. 66, no. 6, pp. 1647-1652, 1979.

[SCH00] D. Schlichtharle. *Digital Filters: Basics and Design,* Springer-Verlag, 2000.

[SHA00] D. O'Shaughnessy. *Speech Communications: Human and Machine,* 2nd edition, IEEE Press, 2000.

[SHA01] R.V. Shannon, J.J. Galvin III, D. and Baskent. "Holes in hearing," *Journal of the Association for Research in Otolaryngology,* vol. 3, no. 2, pp. 185-199, June 2001.

[SMI98] S.W. Smith. *The Scientist and Engineer's Guide to Digital Signal Processing,* 2nd edition, California Technical Publishing, 1998.

[SON66] M.M. Sondhi, and A.J. Presti. "A self-adaptive echo canceler," *Bell System Technical Journal,* vol. 45, pp.1851-1854, 1966.

[SON67] M.M. Sondhi. "An adaptive echo canceler," *Bell System Technical Journal,* vol. 46, pp. 497-511, 1967.

[SOO90] J.S. Soo, and K.K. Pang. "Multidelay block frequency domain adaptive filter," *IEEE Transaction on Acoustics, Speech, and Signal Processing,* vol. 38, no. 2, pp. 373-376, 1990.

[SOR87] H.V. Sorensen et al. "Real-valued fast Fourier transform algorithms," *IEEE Transaction on Acoustics, Speech, and Signal Processing*, vol. 35, no. 6, pp. 849-863, 1987.

[STE03] S.D. Stearns. *Digital Signal Processing with Examples in Matlab,* CRC Press, 2003.

[STE90] S.D. Stearns. *Digital Signal Analysis,* Prentice-Hall, 1990.

[SUG05] A. Sugiyama, J. Berclz, and M. Sato. "Noise-robust double-talk detection based on normalized cross correlation and a noise offset," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 3, pp. 153-156, March 2005.

[SWA98] M. Swanson et al. "Robust audio watermarking using perceptual masking," *Signal Processing Journal,* vol. 66, no. 3, pp. 337-355, May 1998.

[TER79] E. Terhardt. "Calculating virtual pitch," *Hearing Research*, vol. 1. pp 155-182, 1979.

[TER82] E. Terhardt et al. "Algorithm for extraction of pitch and pitch salience form complex tonal signals," *Journal of Acoustical Society of America*, vol. 71, pp. 679-688, 1982.

[THI01] J. Thiemann. *Acoustic Noise Suppression for Speech Signals Using Auditory Masking Effects*, Master Thesis, McGill University, Montreal, Canada, 2001.

[TOB70] J. Tobias. *Foundation of Modern Auditory Theory,* Academic Press, 1970.

[TRE01] J.R. Treichler, C.R. Johnson, and M.G. Larimore. *Theory and Design of Adaptive Filters,* Prentice Hall, New Jersey, 2001.

[TSO97] D.E. Tsoukalas, J. Mourjopoulos, and G. Kokkinakis. "Perceptual filters for audio signal enhancement," *Journal of the Audio Engineering Society,* vol. 45, no. 1-2, pp. 22-36, 1997.

[VIR99] N. Virag. "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech and Audio Processing,* vol. 7, no. 2, pp. 126-137, March 1999.

[VU04] T.A. Vu, H. Ding, and M. Bouchard. "A survey of double talk detection schemes for echo cancellation applications," *Proceedings of the Acoustics Week in Canada*, vol. 32, no. 3, pp. 144-145, Canada, October 2004.

[VU06] T.A. Vu, H. Ding, and M. Bouchard. "A double talk detection algorithm using a psychoacoustic auditory model," *Proceedings of 14th European Signal Processing Conference (EUSIPCO)*, Florence, Italy, September 2006.

[WAL03] F. Wallin. *Combining Acoustic Echo Cancellation and Suppression,* Master Thesis, Linkoping University, Linkoping, Sweden, 2003.

[WID99] B. Widrow, and S. D. Stearns. *Adaptive Signal Processing,* Prentice Hall, Upper Saddle River, NJ, 1999.

[XU00] X. Xu. *Effects of Delay Jitter and Clock Slippage on Network Echo Cancellation,* Master Thesis, Carleton University, Ottawa, Canada, 2000.

[YAS93] H. Yasukawa, and S. Shimada. "An acoustic echo canceller using subband sampling and decorrelation methods," *IEEE Transaction on Signal Processing,* vol. 41, pp. 926-930, 1993.

[YE91] H. Ye, and B.X. Wu. "A new double-talk detection algorithm based on orthogonality theorem," *IEEE Transaction on Communications,* vol. 39, pp. 1542-1545, November 1991.

[ZWI90] E. Zwicker, and H. Fastl. *Psychoacoustics: Facts and Models,* Springer-Verlag, 1990.

[ZWI99] E. Zwicker. *Psychoacoustics: Facts and Models,* 2nd edition, Springer, New York, 1999.