Robust Echo Cancellation in Harsh Environments

submitted by

James D. Gordy, B.Sc., M.Sc.

A thesis submitted to

the Faculty of Graduate Studies and Research

in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

Ottawa-Carleton Institute for Electrical and Computer Engineering

Faculty of Engineering and Design

Department of Systems and Computer Engineering

Carleton University

Ottawa, Ontario, Canada K1S 5B6

February 28, 2007

Your file Votre référence
ISBN: 978-0-494-27096-7
Our file Notre référence
ISBN: 978-0-494-27096-7

Abstract

This thesis deals with the development of adaptive filter structures, adaptation algorithms, and control algorithms capable of operating in harsh environments, and applied to the problem of echo cancellation in telecommunications systems. The specific environments considered are moderate-to-high levels of background noise, doubletalk conditions, and echo cancellation in Voice-over-IP (VoIP) networks. A secondary focus is maintaining low complexity in the resulting structures and algorithms.

The problem of doubletalk detector calibration is addressed by statistical modeling and applied to the normalized cross-correlation-based doubletalk detector. Detection probability is a function of the parameter estimation window, echo to noise ratio (SNR) and near-end speech to echo ratio (NER), and methods for compensating background noise bias do not eliminate the detection statistic's SNR dependency. Signal-adaptive algorithms are presented for constructing thresholds based on statistical criteria, which are shown to be capable of increasing detection rates.

Psychoacoustic limits of echo canceller performance in the presence of noise are quantified using a perceptual model of hearing, and it is shown that average-power-based performance measures may under- or over-estimate the amount of audible echo removed by an echo canceller. Two algorithms are proposed for estimating the audible echo signal reduction provided by an echo canceller, and verified using informal listening tests.

Affine Projection (AP) and normalized cross-correlation-based doubletalk detection algorithms are derived for echo cancellers employing critically sampled subband adaptive filters. Subband AP, even with only 2 – 4 subbands, can improve the rate of convergence

over fullband AP employing the same projection order. Background noise is not spectrally flat, and so per-subband adaptive detection thresholds can be constructed which provide an improvement in detection rates over fullband doubletalk detectors.

Adaptation and control algorithms utilizing linear-prediction-based speech parameters are proposed for echo cancellers deployed in VoIP networks. Decorrelated adaptation and doubletalk detection algorithms are presented that avoid the cost of constructing decorrelation filter coefficients. A power spectrum estimation algorithm is proposed for residual echo from nonlinear vocoder distortion. When incorporated into a frequency-domain post-filter, near-end speech spectral distortion is improved by 0.98 dB, with 0.4 increase in estimated mean opinion score.

Acknowledgements

First of all, I would like to express my gratitude to my supervisor, Dr. Rafik Goubran, for providing consistent support, patience, and motivation that has kept me focused on my work. It is his enthusiasm and interest in research that has made my time at Carleton University an enjoyable and rewarding experience.

I would also like to gratefully acknowledge financial support from the Faculty of Graduate Studies and Research and the Department of Systems and Computer Engineering at Carleton University, the Natural Sciences and Engineering Research Council of Canada (NSERC), Ontario Graduate Scholarships in Science and Technology (OGS-ST), Ontario Centres of Excellence (OCE), and Mitel Networks. My research and this thesis would not have been possible without their assistance.

Finally, I wish to thank Danny Lemay and the SCE technical staff for keeping the DSP lab and our networks running smoothly. To the staff of Mitel Networks, special thanks go to Dr. Franck Beaucoup for providing valuable feedback and technical discussions. My fellow graduate students in the DSP lab, particularly Trevor Burton, Brady Laska, Anita McKee and Joseph Gammal, receive many thanks for their friendship and helpful discussions (technical and otherwise).

Table of Contents

List of Figures

List of Tables

List of Symbols

An underlined uppercase letter generally refers to a matrix of numbers or variables, and an underlined lowercase letter generally refers to a vector of numbers or variables. A non-underlined uppercase or lowercase letter is a scalar quantity.

Arabic Symbols

Symbol	Description
$\underline{C}(n)$	Input signal correlation matrix estimate at time *n*.
$\underline{C}_i(m)$	Input signal correlation matrix estimate for subband *i* at time *n*.
D	(i) Downsampling factor; (ii) end-to-end transmission delay.
d(n)	Reference signal at time *n*.
$d_f(n)$	Filtered reference signal at time *n*.
$d_i(m)$	Reference signal for subband *i* at time *m*.
e(n)	Error signal at time *n*.
$e_f(n)$	Filtered error signal at time *n*.
$e_f(n, l)$	Filtered error signal for the l^{th} coefficient at time *n*.
$e_i(m)$	Error signal for subband *i* at time *m*.
F	Decorrelation filter order.
$F_i(z)$	Synthesis filter for subband *i* in the z domain.
f(n)	Decorrelation filter at time *n*.
$f(n, i)$	i^{th} coefficient for time-varying decorrelation filter at time *n*.
f_s	Sampling frequency.
$f_Z(z)$	Probability distribution function of random variable *Z*.
$F_Z(z)$	Cumulative distribution function of random variable *Z*.
$H(\omega)$	Weighting filter transfer function in the frequency domain.
$H_i(z)$	Analysis filter for subband *i* in the z domain.
J(n)	Adaptive filter cost function at time *n*.

Symbol	Description
$J_i(n)$	Adaptive filter cost function for subband i at time n.
$J(\omega)$	Frequency-domain cost function.
K	Parameter estimation window length.
L	Linear echo path system length.
M	Number of subbands in analysis/synthesis filter bank.
N	Adaptive filter length.
N_D	Subband adaptive filter length.
N_F	Synthesis filter length.
N_H	Analysis filter length.
$N(\omega)$	Noise signal in the frequency domain.
$N(\mu_x, \sigma_x^2)$	Gaussian random variable X with mean μ_x and variance σ_x^2.
P	Projection order.
P_F	Probability of false alarm.
P_M	Probability of miss.
$r(n)$	Short-term excitation signal at time n.
$\underline{r}_{i,j}$	Cross-correlation vector of input and reference signals for subbands i and j.
$\underline{\hat{r}}_{i,j}(m)$	Estimated cross-correlation vector of input and reference signals for subbands i and j at time m.
$\underline{r}_{xd}$	Cross-correlation vector of input and reference signals.
$\underline{\hat{r}}_{xd}(n)$	Estimated cross-correlation vector of input and reference signals at time n.
$\underline{r}_{xe}(n)$	Cross-correlation vector of input and error signals.
$\underline{\hat{r}}_{xe}(n)$	Estimated cross-correlation vector of input and error signals.
$\underline{R}_{i,j}$	Input signal cross-covariance matrix for subbands i and j.
$\underline{\hat{R}}_{i,j}(m)$	Estimated input signal cross-covariance matrix for subbands i and j at time m.
$\underline{R}_{xx}$	Autocorrelation matrix of input signal vector $x(n)$.

Symbol	Description
$S_{XX}(\omega)$, $S_{XX}(k)$	Power spectrum of signal $x(n)$ in continuous- and discrete-frequency domains.
SNR	Echo to background noise power ratio.
SNR_D	Signal to quantization noise ratio for reconstructed reference signal.
SNR_X	Signal to quantization noise ratio for reconstructed input signal.
$SNR_D(\omega)$	Frequency-weighted signal to quantization noise ratio for reconstructed reference signal.
$SNR_X(\omega)$	Frequency-weighted signal to quantization noise ratio for reconstructed input signal.
T	Doubletalk detection threshold.
$T_A(f)$, $T_A(k)$	Absolute threshold of hearing in continuous- and discrete-frequency domains.
$T_M(\omega)$, $T_M(k)$	Masking threshold in continuous- and discrete-frequency domains.
$T_{PF}(n)$	Doubletalk detection threshold for given probability of false alarm at time n.
$T_{PF,i}(m)$	Subband doubletalk detection threshold for given probability of false alarm at time m.
$T_{PM}(n)$	Doubletalk detection threshold for given probability of miss at time n.
$T_{PM,i}(m)$	Subband doubletalk detection threshold for given probability of miss at time m.
$v(n)$	Near-end speech signal at time n.
$v_i(m)$	Near-end speech signal for subband i at time m.
$\underline{w}(n)$	Echo path impulse response vector at time n.
$w_i(n)$	i^{th} coefficient of echo path impulse response vector at time n.
$\underline{w}_i(m)$	Echo path impulse response vector for subband i at time m.
$\underline{w}_{opt}$	Optimal Wiener filter coefficient vector.
$\underline{\hat{w}}(n)$	Adaptive filter coefficient vector at time n.
$\hat{w}_i(n)$	i^{th} coefficient of adaptive filter vector at time n.

Symbol	Description
$\underline{\hat{w}}_i(m)$	Adaptive filter coefficient vector for subband i at time m.
$\underline{X}(n)$	Input signal matrix at time n.
$\underline{X}_{i,j}(m)$	Input signal matrix for subbands i and j at time m.
$x(n)$	Input signal at time n.
$x_f(n)$	Filtered input signal at time n.
$x_{i,j}(m)$	Input signal for subbands i and j at time m.
$y(n)$	Echo signal at time n.
$y_i(m)$	Echo signal for subband i at time m.
$\hat{y}(n)$	Estimated echo signal at time n.
$\hat{y}_i(m)$	Estimated echo signal for subband i at time m.

Greek Symbols

Symbol	Description
$\alpha(n)$	Doubletalk detector scale factor affecting adaptation step size parameter.
$\delta(n)$	Residual echo signal at time n.
$\Delta\underline{w}(n)$	Adaptive filter error vector at time n.
$\Delta w_i(n)$	i^{th} coefficient of adaptive filter error vector at time n.
$\xi(n)$	Doubletalk detection statistic at time n.
$\xi_i(m)$	Doubletalk detection statistic for subband i at time m.
$\eta(n)$	Background noise signal at time n.
λ	Smoothing factor for parameter estimation.
λ_i	i^{th} eigenvalue for the input autocorrelation matrix.
μ	Adaptation step size parameter.
$\mu(n)$	Adaptation step size parameter at time n.
$\mu_i(m)$	Adaptation step size parameter for subband i at time m.
$\sigma_x^2(n)$	Average power of signal $x(n)$ at time n.

Symbol	Description
$\hat{\sigma}_x^2(n)$	Estimated average power of signal $x(n)$ at time n.
$\sigma_{x,i}^2(m)$	Average power of signal $x(n)$ for subband i at time m.
$\hat{\sigma}_{x,i}^2(m)$	Estimated average power of signal $x(n)$ for subband i at time m.
$\sigma_{y,audible}^2$	Audible echo signal power.
$\sigma_{\delta,audible}^2$	Audible residual echo signal power.
$\sigma_{\Delta D}^2$	Excitation signal quantization noise for reconstructed reference signal.
$\sigma_{\Delta X}^2$	Excitation signal quantization noise for reconstructed input signal.

Chapter 1 Introduction

1.1 Problem Statement

Traditional wireline telephony over the public switched telephone network (PSTN) remains the standard method of day-to-day voice communications. However, increasing computing power and more sophisticated digital signal processing algorithms are causing an evolution in communications. Videoconferencing systems, hands-free terminals and wireless networks continue to become popular modes of communication [1], [2], [3]. A more recent trend is the migration of voice services onto packet-switched Internet Protocol (IP) networks in homes and businesses, often referred to as Voice-over-IP (VoIP). In addition, telecom providers are transparently merging their circuit-switched voice networks into integrated IP-based voice-and-data core networks [4], [5], [6]. A fundamental requirement for user acceptance of these new technologies is speech quality as good as, or better than, traditional wireline telephony. Of particular importance is the performance of digital echo cancellers, which are typically deployed in terminals and throughout the PSTN for removing acoustic and network echo. The problem of echo cancellation is a popular application of adaptive filter theory, and there is an established body of literature devoted to the design, implementation, and analysis of echo cancellers [2], [7].

Typical requirements of echo cancellers are stability, fast convergence and tracking capabilities, low computational complexity. Often these requirements are in contention, leading to trade-offs in the choice of structures and adaptation algorithms. There are also

many harsh physical and environmental aspects that limit an echo canceller's performance in practice. In particular, the presences of background noise, doubletalk conditions, and nonlinearity in the echo path greatly inhibit echo canceller performance [8], [9], [10]. End-user perception and tolerability of echo is a function of the round-trip time, and increased delay reduces the perceived effectiveness of existing echo cancellers [11]. Therefore, in addition to an adaptive filter and adaptation algorithm, secondary structures such as a doubletalk detector and a nonlinear processor (NLP) or post-filter are required in practice [12]. The performance and interaction of each of these components has a direct impact on the stability and performance of the echo canceller as a whole.

The presence of background noise in the environment is a common harsh condition in which an echo canceller must operate. In mobile environments, for example, the background noise is typically of a moderate-to-high level, time-varying, and not spectrally flat. Echo canceller performance is typically expressed using mean squared error (MSE) and echo return loss enhancement (ERLE) provided by the echo canceller [12]. These two measures are simple to calculate in terms of average signal powers, but they assume the presence of low background noise. However, higher noise will tend to mask the presence of residual (un-cancelled) echo, at some point forming an upper bound on perceived echo canceller performance. This effect has not been thoroughly studied in the literature, and there is a need for echo canceller performance measures that can take into account the presence of masking effects of background noise.

The presence of doubletalk is another harsh environment encountered in practice, and it is usually mitigated by incorporating a doubletalk detector into an echo canceller

implementation. A number of doubletalk detection algorithms have been proposed in the literature, with varying degrees of computational complexity and performance [9], [13], [14], [15], [16]. However, an open problem is that of calibration, or selecting operating parameters – such as a detection threshold – that are "optimal" in some sense. Of equal importance is being able to calibrate a doubletalk detector using the same approach for arbitrary echo path environments. It would also be helpful to be able to determine a given doubletalk detector's expected performance and response time under various conditions. An intuitive approach to solving these problems is by applying statistical analysis to construct a model of doubletalk detector behavior. However, the literature lacks such an approach to doubletalk detector modeling and calibration.

A third set of harsh conditions arises from the nature of VoIP technologies, and interconnections of IP networks with the legacy PSTN. Differences between the networks have resulted in new echo cancellation problems and exacerbated existing ones. First of all, voice is transmitted through the PSTN as a continuous stream of pulse code modulated samples (ITU-T G.711), whereas in VoIP networks speech frames are often compressed using low-bit-rate parametric coders such as the ITU-T G.729 and G.723.1 speech codecs [4], [17], [18], [19]. A second major difference is that the PSTN provides guaranteed bandwidth and relatively constant delay for each voice channel, whereas transmission over packet-switched networks is subject to variable end-to-end delay and losses due to network congestion [5]. The combination of compression, packetization, transport, and playback delays results in longer overall round-trip delay, which is known to increase end-user perception of echo [20]. As a result, there is a still a need for

adaptation algorithms that can offer increased convergence and tracking rates. Another possible problem arises when the signal paths to or from the echo path employ low-bit-rate speech compression, such as an encoder / decoder pair. The resulting distortion is highly nonlinear and will inhibit echo canceller performance [21].

Although the processing power of digital signal processors has been increasing in recent years, not all of this power can be devoted to echo cancellation. In a modern digital cellular telephone, for example, an echo canceller must share computational resources with other firmware providing noise suppression, speech enhancement and compression, network interfaces, channel coding, and radio frequency (RF) processing [23]. Therefore, new adaptation and control algorithms for dealing with harsh environments must still be tempered with computational complexity constraints. Subband adaptive filter structures are one approach to implementing echo cancellers while reducing the computational complexity [24]. Computational savings are obtained by segmenting the input and reference signals into frequency bands and decimating the signals in each subband, which also leads to a shorter-length adaptive filter in each subband. Maximal savings are obtained by using filter banks employing critical sampling, but these structures require unique algorithms to handle aliasing in each subband [25]. Only recently have structures been proposed offering more robust handling of aliasing. In subband adaptive filters, each adaptive filter is adapted independently of the others, and so a natural idea is to attempt to combine other components such as doubletalk detection and more sophisticated adaptation algorithms

into each subband. However, it is not obvious how to incorporate such algorithms into echo cancellers employing critically sampled filter banks.

1.2 Objectives and Motivation

This thesis is primarily about the development and analysis of echo canceller structures and algorithms capable of operating in harsh environments, with specific application to those outlined in the previous section: doubletalk conditions, background noise, and the effects of speech coding in VoIP networks. A secondary goal is maintaining a low computational complexity in the resulting algorithms. In particular, there are five basic research questions addressed by this thesis:

1. What is the behavior of doubletalk detectors in the presence of background noise and time-varying signal statistics, and how can an appropriate detection threshold be constructed given these influences?
2. What are the perceived limits of echo canceller performance given psychoacoustic aspects of human hearing, and how does the presence of background noise affect perceived echo canceller performance?
3. Given that additional structures are required to reduce aliasing in critically sampled subband filters, is it possible to incorporate more sophisticated adaptation and doubletalk detection algorithms into echo cancellers based on these structures?

4. Is it possible to incorporate parametric representations of speech, utilized by low-bit-rate speech encoders in VoIP networks, in the design of echo canceller structures for VoIP gateways and IP-based terminals?
5. What is the effect of vocoder distortion on echo canceller performance, in particular due to the presence of encoder / decoder pairs on the send and / or receive paths? Are there ways to mitigate the presence of this distortion?

1.3 Contributions of the Thesis

The contributions of this thesis are the development of structures and adaptation algorithms for achieving echo cancellation in harsh environments such as background noise and doubletalk conditions. In particular, the main contributions of this thesis to the field of adaptive echo cancellation are as follows:

1. **Statistical modeling of doubletalk detectors** – An approach to doubletalk detector calibration is proposed based on the statistical performance measures of probability of false alarm (Type I error) and probability of miss (Type II error). This approach is applied to an existing doubletalk detection algorithm, resulting in the development of models of the detection statistic's probability density function in the absence and presence of near-end speech. Two algorithms are proposed for constructing detection thresholds adaptive to changes in the environment, such as time-varying input signal and noise statistics. These algorithms allow the practitioner to construct statistically optimal doubletalk detection thresholds instead of relying on sub-optimal empirical methods. These contributions are

described in Chapter 4 of this thesis, and the results of this work are published in [26] and [27].

2. **Perceptual performance limitations of echo cancellers** – An investigation is conducted into limits of perceived echo canceller performance based on psychoacoustic limitations of human hearing, and in particular the masking effects of moderate-to-high levels of background noise. It is found that existing echo canceller performance measures based on average power estimates, such as echo return loss enhancement (ERLE), cannot incorporate frequency-dependent psychoacoustic limitations of human hearing. As a result, ERLE may over- or under-estimate the audible echo power reduction achieved by an echo canceller. Two echo canceller performance measures are proposed that estimate the amount of audible echo power reduction by incorporating psychoacoustic limitations. These contributions are described in Chapter 5, and the results of this work are published in [28] and [29].
3. **Affine Projection (AP) and doubletalk detection in critically sampled subband adaptive filters (CS-SBAF)** – In this work the use of AP in a recently proposed SBAF employing critical sampling is investigated, and a convergence analysis of the algorithm is proposed. In addition, a derivation is shown for a cross-correlation-based doubletalk detector for use in the CS-SBAF structure. One result of this work is a convergence analysis of AP in the critically sampled SBAF, which offers insight into the factors affecting the rate of convergence of such structures. One outcome of deriving the doubletalk detector is that a

doubletalk decision can be made for each subband adaptive filter based on a measure of near-end speech power in each subband. This contribution is described in Chapter 6, and the results appear in [30] and [31].

4. **Low-complexity decorrelated NLMS adaptation and cross-correlation-based doubletalk detection algorithms for VoIP** – Speech decoders are often co-located with network or acoustic echo cancellers at VoIP gateways and IP-based phones. In this work, signal processing algorithms are proposed that employ information from compressed speech frames available at LPC-based speech decoders. The results are a decorrelated NLMS algorithm for echo canceller adaptation, and a cross-correlation-based doubletalk detection algorithm. The availability of speech parameters allows the algorithms to be implemented with lower complexity than a straightforward implementation. This contribution is described in Chapter 7, and the results appear in [32] and [33].
5. **Post-filtering algorithms for suppression of residual echo due to vocoder distortion in VoIP** – This work addresses the problem of residual echo resulting from the presence of nonlinear distortion from LPC-based speech encoders in the echo path. An investigation into the power spectrum properties of the residual echo resulting from ITU-T G.729A encoder / decoder pairs in the echo path is presented. A power spectrum estimation algorithm for modeling the residual echo is presented for use as part of a frequency-domain post-filter for enhancing near-end speech. This contribution is described in Chapter 7, and the results are published in [34].

1.4 Organization and Scope

This thesis has four central chapters. Chapters 4 and 5 address the issues of statistical modeling of doubletalk detectors for calibration and performance evaluation, and the perceptual performance limitations of echo cancellers in the presence of background noise. Chapter 6 investigates new adaptation and control algorithms for subband adaptive filters. Chapter 7 presents structures and algorithms for echo cancellation, doubletalk detection, and post-filtering in VoIP networks. Supporting Chapters 2, 3, and 8 review the echo cancellation problem and provide necessary background material on adaptive filter theory, doubletalk detection, post-filtering and VoIP structures, and summarize and draw conclusions from this work.

Chapter 2 presents an overview of the echo cancellation problem, and provides a review of echo canceller structures, adaptation algorithms, and supporting components such as doubletalk detection and post-filtering algorithms. In addition, it provides a review of low-bit-rate speech compression technologies commonly employed in VoIP systems.

Chapter 3 provides a description of the experimental and simulation setups used to obtain results for the algorithms and structures described in this thesis. It provides a description of the room dimensions and contents, the playback and recording equipment employed, and sources of noise and speech signals used in experiments.

Chapter 4 addresses the problem of doubletalk detector calibration in noisy environments. This is done by applying statistical modeling to derive probability

density functions characterizing the detection statistic's behavior in the absence and presence of doubletalk. These results are used to develop two calibration algorithms based on statistically optimal criteria, and simulation results are provided confirming the validity of the models.

Chapter 5 investigates performance limitations of echo cancellers by considering the perceptual limitations of the human auditory system. In particular, the masking effects of background noise on residual echo are analyzed using the MPEG Psychoacoustic Model. The results are used to construct an echo canceller performance measure that takes into account these masking effects. The results of confirmatory informal listening tests are presented.

Chapter 6 investigates the incorporation of Affine Projection and cross-correlation-based doubletalk detection algorithms into a subband adaptive echo cancellers employing critically sampled filter banks. A mathematical derivation of the two algorithms is provided to show how they can be derived for each subband. A convergence analysis of the subband AP algorithm is provided. The behavior of the algorithms is investigated through computer simulations.

Chapter 7 analyzes the statistical properties of parametric speech representations and the structures of low-bit-rate speech coders. The results are used to derive three low-complexity algorithms for performing echo cancellation, cross-correlation-based doubletalk detection, and audio mixing in VoIP systems. The algorithms achieve complexity reduction by taking advantage of compressed speech parameters. In addition, the problem of suppressing residual echo caused by nonlinear vocoder

distortion in the echo path is studied. A power spectrum estimation algorithm is described for modeling the residual echo signal, and simulation results are presented verifying the improvement in near-end speech quality.

Chapter 8 highlights the important contributions arising from the thesis, and suggests future areas of research to build upon the contributions of this thesis.

Chapter 2 Background

2.1 Overview

This chapter presents a brief literature review of echo canceller structures and algorithms, as well as additional supporting structures required for typical echo canceller implementations. The sections of this chapter are organized as follows. Section 2.2 provides an overview of the echo cancellation problem in the context of network and acoustic echo, as well as a description of typical echo canceller implementations and supporting components. Section 2.3 reviews commonly used echo canceller structures and adaptation algorithms, with a focus on low-complexity algorithms. Section 2.4 describes the problem of doubletalk detection, and reviews several commonly used algorithms for detecting doubletalk conditions. Section 2.5 provides a review of post-filtering structures for suppression of residual echo after linear echo cancellation. Finally, a review of Voice-over-IP (VoIP) structures and related algorithms is provided in Section 2.6.

2.2 The Echo Cancellation Problem

Network echo arises in the legacy public switched telephone network (PSTN) at the analog interface between the central office (CO) and the local loop leading to customer premises, as shown in Figure 2.1 [13]. It is at this point that four-wire unidirectional digital connections from the central office are converted to and from a bidirectional analog two-wire twisted pair leading to the customer premises. This is done using digital-to-analog (DAC) and analog-to-digital (ADC) converters on the input and output

ports, and a transformer-like coupling device called a hybrid to interface with the two-wire port. In an ideal configuration, the hybrid transfers the incoming discrete-time signal, $x(n)$, from the input port to an analog signal $x(t)$ on the two-wire port with no signal leakage onto the output port. Similarly, the outgoing speech signal $v(t)$ from the customer premises is ideally transferred from the two-wire port to the central office's output port $d(n)$ with no leakage back into the local loop. However, impedance mismatch between the two sides of the hybrid leads to imperfect separation of incoming and outgoing signals. The result is a distorted version of the input signal, $y(t)$, returned to the far end in the form of echo.

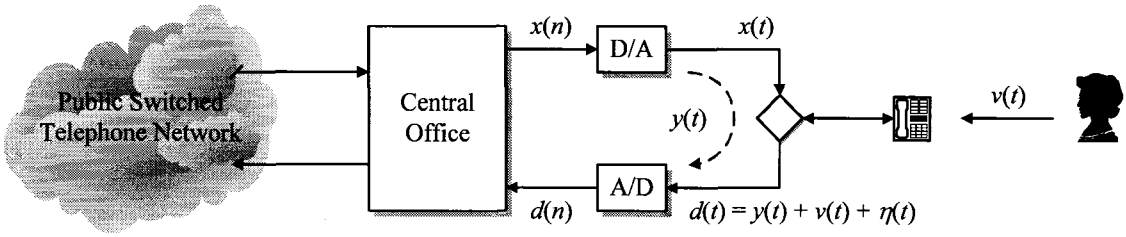


Figure 2.1 – Network echo in the public switched telephone network. In addition to an echo signal $y(t)$, the reference signal may contain noise $\eta(t)$ and near-end speech $v(t)$ signals.

Acoustic echo is predominantly a problem of hands-free telephones for conferencing or videoconferencing systems [7]. However, more recently the problem has arisen in cellular telephones offering hands-free operating modes. These systems are typically designed to provide full-duplex communications allowing both far-end and near-end talkers to be active at the same time, offering partial or full overlap in their speech bursts. A block diagram of a typical configuration is shown in Figure 2.2. In full-duplex systems such as these, far-end speech $x(n)$ is played over a loudspeaker into the near-end room after conversion to a continuous-time waveform $x(t)$ through a DAC. An echo signal $y(t)$ is recorded by the microphone, consisting of an attenuated version of the far-end speech

signal via a direct path from the loudspeaker (acoustic coupling), and reflected versions of the signal caused by walls objects within the room (reverberation). In addition to echo, the microphone records near-end speech $v(t)$ and background noise $\eta(t)$ to form the reference signal $d(n)$ after conversion using an ADC.

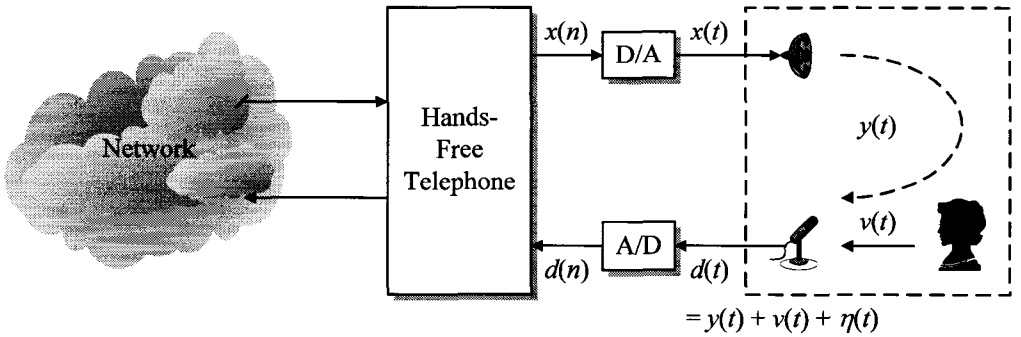


Figure 2.2 - Acoustic echo in a hands-free telephone. In addition to an echo signal $y(t)$, the reference signal may contain noise $\eta(t)$ and near-end speech $v(t)$ signals.

The echo paths and operating environments of network and acoustic echo have markedly different characteristics. Network echo paths are typically of a relatively short duration of 5 – 10 ms, whereas acoustic environments can induce a much longer echo path of up to 250 ms in duration [13], [7]. Although the network echo path is typically different for each telephone call, it is usually stationary once the call is established. In acoustic environments the echo path may change continuously during communications due to the movement of people and objects within the room. The actual echo path itself in both cases can be well-modeled as a linear system, but both hybrid transformers and hands-free telephones often contain nonlinear elements which introduce some nonlinearity into the echo path [10], [36]. Background noise is usually more of an issue in hands-free environments, and in conferencing systems the use of stereo or multiple loudspeakers introduces additional sources of echo [37], [38]. In both cases, however,

the presence of echo is disruptive for two reasons. During single-talk periods it is annoying to hear a delayed version of one's own voice. Additionally, during double-talk periods the quality of near-end speech is reduced when it contains echo. The result of this is often an induced "half duplex" conversation between participants, and a perceived degradation in the quality of the communications [38]. Finally, the relative signal powers of echo and near-end speech are often quite different between network and acoustic echo environments. A typical hybrid transformer itself introduces at least 5 – 15 dB of attenuation in coupled signals, resulting in near-end speech power at least 5 – 15 dB higher than that of the echo [13]. In a hands-free telephone, however, the loudspeaker and microphone are often closely situated in the enclosure compared to the relative location of the near-end talker. As a result, it is common for near-end speech power to be at least 10 – 20 dB lower than that of the echo [2].

2.2.1 Echo Canceller Structure and Conventions

The presence of network and acoustic echoes has traditionally been mitigated through the use of digital echo cancellers [13]. Figure 2.3 shows a block diagram of a typical acoustic echo canceller implementation and its components in the context of a full-duplex hands-free telephone. The echo path is modeled as a system to which a far-end speech signal $x(n)$ is applied, resulting in a near-end reference signal $d(n)$ consisting of three uncorrelated components: the echo signal $y(n)$, near-end speech $v(n)$, and an aggregate noise signal $\eta(n)$ containing background and measurement noise. The echo path system parameters are estimated and used to construct an estimate of the echo signal, $\hat{y}(n)$, which is subtracted from the reference signal. In the ideal case of perfect cancellation,

the echo canceller error signal $e(n)$ consists of only the near-end speech and noise signal. In practice, the error signal also contains a residual echo signal $\delta(n)$ resulting from error in the estimated echo signal:

$$d(n) = y(n) + v(n) + \eta(n) \tag{2.1}$$

$$e(n) = d(n) - \hat{y}(n) = \delta(n) + v(n) + \eta(n) \tag{2.2}$$

The echo path is often assumed to be capable of continuous changes, so the linear system is usually adaptive to track changes in the system parameters. As shown in Figure 2.3, a practical echo canceller implementation consists of the following components, which are described in more detail in subsequent sections:

- An echo canceller to estimate and cancel the echo signal from the reference signal, and a corresponding adaptation algorithm to model and track the echo path parameters.
- A doubletalk detector to detect the presence of near-end speech occurring at the same time as far-end speech, and to slow or halt adaptation of the echo canceller for the duration of the near-end speech.
- A nonlinear processor or post-filter to provide additional suppression of residual echo remaining at the echo canceller output.

Note that in practice there are other structures typically found in an echo canceller implementation that, although important, fall outside the scope of this thesis. For example, the echo canceller's adaptation algorithm can only operate effectively in the presence of far-end speech. Therefore, in practice a voice activity detector (VAD) is

typically employed to differentiate between speech bursts and silence periods in the far-end signal $x(n)$ [40].

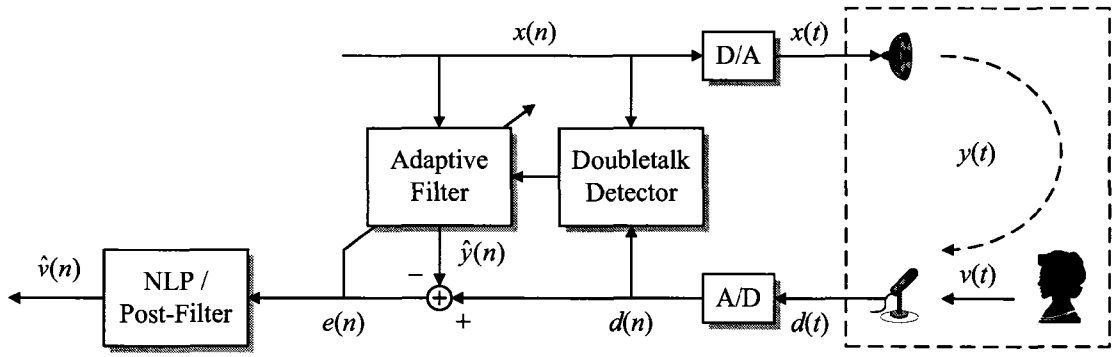


Figure 2.3 - Block diagram of typical practical echo canceller components.

2.2.2 Performance Measures and Operating Requirements

Objective echo canceller performance measures are the steady-state mean squared error (MSE), echo return loss enhancement (ERLE), and system distance (DIST). MSE and ERLE are defined, respectively, as the average error signal power and the ratio of average reference to error signal powers, both expressed in decibels as follows [12]:

$$MSE(dB) = 10\log_{10}\{E[e^2(n)]\} = 10\log_{10}[\sigma_e^2(n)] \approx 10\log_{10}\frac{1}{K}\sum_{k=0}^{K-1} e^2(n-k) \qquad (2.3)$$

$$ERLE(dB) = 10\log_{10}\left\{\frac{E[d^2(n)]}{E[e^2(n)]}\right\} = 10\log_{10}\left[\frac{\sigma_d^2(n)}{\sigma_e^2(n)}\right] \approx 10\log_{10}\left[\frac{\sum_{k=0}^{K-1} d^2(n-k)}{\sum_{k=0}^{K-1} e^2(n-k)}\right] \qquad (2.4)$$

where $\sigma_d^2(n)$ and $\sigma_e^2(n)$ are the steady-state variances of the reference and error signals at time n, respectively, and $E[\cdot]$ is the statistical expectation operator. MSE and ERLE can be estimated as a function of time n over a window of samples using the latter expressions in (2.3) and (2.4) above. MSE provides a measure of the residual echo power after the echo canceller, while ERLE provides a measure of how much the echo is

attenuated by the echo canceller. Both measures assume low levels of background noise, and no near-end speech in the reference and error signals. Theoretically it is possible to have an infinitely small or infinitely large MSE and ERLE, respectively, but in practice the values are limited by the presence of background noise and finite word-length effects in the system implementation.

Operating requirements have been proposed by the International Telecommunication Union (ITU) governing the performance of echo cancellers with respect to the objective measures described above, such as ITU-T G.168 [12]. In addition to objective measures, subjective performance guidelines and measurement techniques have been proposed. In particular, the required amount of echo cancellation to prevent objectionable echo, expressed in terms of ERLE, is a function of the round-trip delay in ITU-T G.131 [11]. Longer round-trip delays, such as those inherent in long-distance telephone calls, increases echo perception and annoyance of echo. In addition, the Mean Opinion Score (MOS) is a subjective measure of speech quality useful, for example, in determining the quality of near-end speech containing residual echo [41], [42].

2.3 Filter Structures and Adaptation Algorithms

Figure 2.3 shows an illustration of the basic adaptive filter structure in the context of echo cancellation. The unknown system is modeled as a potentially time-varying linear system consisting of a finite impulse response (FIR) of length L samples, whose coefficients are represented as an $L \times 1$ vector $\underline{w}(n)$. The resulting echo signal $y(n)$ is represented as the convolution of the input signal and the impulse response coefficients:

$$y(n) = \underline{x}^T(n)\underline{w}(n) \tag{2.5}$$

$$\underline{w}(n) = [w_0(n) \quad w_1(n) \quad \cdots \quad w_{L-1}(n)]^T \tag{2.6}$$

$$\underline{x}(n) = [x(n) \quad x(n-1) \quad \cdots \quad x(n-L+1)]^T \tag{2.7}$$

The most straightforward and commonly used approach is to model the echo path impulse response as a linear filter $\underline{\hat{w}}(n)$ with a finite impulse response (FIR) of length $N \leq L$ samples. The output $\hat{y}(n)$ of the adaptive filter is an estimate of the echo signal $y(n)$ resulting from applying the input signal $x(n)$ to the unknown system:

$$\hat{y}(n) = \underline{x}^T(n)\underline{\hat{w}}(n) \tag{2.8}$$

$$\underline{\hat{w}}(n) = [\hat{w}_0(n) \quad \hat{w}_1(n) \quad \cdots \quad \hat{w}_{N-1}(n)]^T \tag{2.9}$$

$$\underline{x}(n) = [x(n) \quad x(n-1) \quad \cdots \quad x(n-N+1)]^T \tag{2.10}$$

This approach is often chosen because it is easily implemented in hardware as a tapped delay line with $N - 1$ delay elements and N multipliers, and as such it is stable and, in practice, relatively robust to finite word length effects of numerical representations [43]. The weights of the linear filter are adapted using an algorithm that minimizes some function of the error signal $e(n)$ between the desired signal $d(n)$ and the adaptive filter output $\hat{y}(n)$. The most commonly used methods of filter adaptation are based on variations of the method of steepest decent [44]. Steepest decent iteratively converges upon the target impulse response vector using the gradient of a cost function $J(n)$ that it is desirable to minimize:

$$\underline{\hat{w}}(n+1) = \underline{\hat{w}}(n) - \mu(n)\underline{P}(n)\underline{\nabla}J(n) \tag{2.11}$$

where $\underline{\hat{w}}(n)$ and $\underline{\hat{w}}(n+1)$ are the adaptive filter coefficients at times n and $n + 1$, $\mu(n)$ is a step size parameter, and $P(n)$ is a matrix constructed to accelerate the rate of convergence

of the adaptation. $\underline{\nabla}J(n)$ is the gradient of the cost function taken with respect to the adaptive filter coefficients:

$$\underline{\nabla}J(n) = \frac{\partial J(n)}{\partial \underline{\hat{w}}(n)} \tag{2.12}$$

Different choices for the cost function $J(n)$ and acceleration matrix $\underline{P}(n)$ lead to different classes of iterative adaptation algorithms, and as a result there have been many adaptation algorithms proposed in the literature, with varying degrees of complexity, stability, and robustness. Surveys of various classes of adaptation algorithms for echo cancellation can be found in [7] and [44]. In Section 2.3.1 several gradient-based adaptation algorithms relevant to this thesis are reviewed in more detail.

Modeling the echo path as a linear system also results in a performance measure that complements the MSE and ERLE defined in (2.3) and (2.4). The system error norm, or system distance (DIST), is expressed in decibels as follows:

$$DIST(dB) = 10\log_{10}\left\{\frac{[\underline{w}(n) - \underline{\hat{w}}(n)]^T[\underline{w}(n) - \underline{\hat{w}}(n)]}{\underline{w}^T(n)\underline{w}(n)}\right\} \tag{2.13}$$

where $\underline{w}(n)$ is the true echo path impulse response of (2.6), and it is assumed that $N = L$. One advantage to employing the system distance is that it is not dependent upon the input or reference signals. In practice, however, the echo path impulse response coefficients are not known, restricting the use of system distance to simulations.

It is important to note that although the finite impulse response model is popular in practice, it is not the only implementation structure available. Adaptive echo cancellers have been constructed using recursive linear filters consisting of an infinite impulse response (IIR) with varying degrees of success [48]. Both FIR and IIR structures can be

implemented using feed-forward and recursive lattice structures to further improve immunity to finite word length effects, but at some expense of computational efficiency [43], [44]. In addition, echo cancellers implemented using frequency domain adaptive filters (FDAF) have been extensively studied [24], [44], [45], [46], [47]. Finally, adaptive echo cancellers constructed using a subband decomposition of the input and reference signals have been reported in the literature [49], [50], [51], [52]. Section 2.3.2 reviews in more detail several subband adaptive filter structures relevant to this thesis.

2.3.1 Gradient-Based Adaptation Algorithms

In gradient adaptation algorithms the cost function $J(n)$ to be minimized is the expected squared value of the echo canceller error signal given by (2.2):

$$J(n) = E[e^2(n)] \tag{2.14}$$

It can be shown in this case that the optimum weight vector $\underline{w}_{opt}$ minimizing $J(n)$ can be obtained by solving the Wiener-Hopf equations [44]:

$$E[\underline{x}(n)\underline{x}^T(n)]\underline{w}_{opt} = E[d(n)\underline{x}(n)] \tag{2.15}$$

One practical approach to implementing the steepest descent algorithm is to replace the statistical expectation operator in (2.15) with an estimate formed from the instantaneous error [44]. Substituting the instantaneous error into the gradient yields the following:

$$\underline{\nabla} J(n) = \frac{\partial[e^2(n)]}{\partial \underline{\hat{w}}(n)} = 2e(n)\frac{\partial[d(n) - \hat{y}(n)]}{\partial \underline{\hat{w}}(n)} = -2e(n)\frac{\partial[\underline{x}^T(n)\underline{\hat{w}}(n)]}{\partial \underline{\hat{w}}(n)} = -2e(n)\underline{x}(n) \tag{2.16}$$

If the input signal is stationary and ergodic, then the ensemble-averaged instantaneous gradient estimate is equal to the true gradient vector. Substituting (2.16) into (2.11) yields the general form of the steepest descent algorithm:

$$\underline{\hat{w}}(n+1) = \underline{\hat{w}}(n) + \mu(n)\underline{P}(n)\underline{x}(n)e(n) \tag{2.17}$$

where the factor of 2 in (2.16) is included in the step size $\mu(n)$ and, as before, the choice of matrix $\underline{P}(n)$ is chosen to improve the rate of convergence of the algorithm. In the following subsections two particular variations of gradient adaptation algorithms are presented, the Normalized LMS and Affine Projection algorithms.

2.3.1.1 Normalized Least Mean Square (NLMS) Algorithm

The Least Mean Square (LMS) algorithm is obtained by setting $\underline{P}(n)$ equal to the identity matrix and by fixing the step-size parameter $\mu(n)$ to a constant value [44]. The resulting adaptation algorithm is given by the following equation, along with the resulting stability bounds for the step-size parameter μ:

$$\underline{\hat{w}}(n+1) = \underline{\hat{w}}(n) + \mu\underline{x}(n)e(n) \tag{2.18}$$

$$0 < \mu < \frac{2}{\sum_{i=0}^{N-1}\lambda_i} \tag{2.19}$$

where λ_i, $0 \le i \le N-1$, are the eigenvalues of the $N \times N$ input signal correlation matrix.

In practice the input signal correlation matrix is not known, and so a more commonly employed version of LMS is obtained by normalizing the step size by the l^2 (or Euclidean) norm of the input signal vector $\underline{x}(n)$ [44]. The resulting algorithm and step-size parameter bounds for stability are given as follows:

$$\underline{\hat{w}}(n+1) = \underline{\hat{w}}(n) + \mu \frac{\underline{x}(n)e(n)}{\underline{x}^T(n)\underline{x}(n) + \delta} \tag{2.20}$$

$$0 < \mu < 2 \tag{2.21}$$

The LMS and NLMS algorithms are popular in practice because of their low complexity, approximately $2N$ multiplications per sample, and because of their stability and general robustness to finite word length effects on fixed-point processors [53]. However, the algorithms generally suffer from a slow rate of convergence in the presence of correlated input signals such as speech [54]. This is because the rate of convergence is dependent on the eigenvalue spread of the input signal correlation matrix, and faster performance can be obtained when the input signal is white. One simple approach to improving the rate of convergence of LMS / NLMS is to employ a fixed or adaptive decorrelation filter to whiten the input signal [7].

2.3.1.2 Affine Projection (AP) Algorithm

The Affine Projection (AP) algorithm employs an adaptive filter update vector constructed from a P-dimensional projection of the current and $P - 1$ previous input signal vectors onto the *a posteriori* error signal vector [58]. The parameter P is commonly referred to as the projection order. The *a posteriori* error signal vector is first constructed by re-calculating the set of P error signals using the current adaptive filter coefficients:

$$\underline{e}(n) = \underline{d}(n) - \underline{\hat{y}}(n) \tag{2.22}$$

$$\underline{d}(n) = \left[d(n) \quad d(n-1) \quad \cdots \quad d(n-P+1)\right]^T \tag{2.23}$$

$$\underline{\hat{y}}(n) = \underline{X}^T(n)\underline{\hat{w}}(n) \tag{2.24}$$

$$\underline{X}(n) = [\underline{x}(n) \quad \underline{x}(n-1) \quad \cdots \quad \underline{x}(n-P+1)]^T \tag{2.25}$$

An estimate of the gradient vector is then constructed such that it forces the a posteriori error signal vector to zero. The resulting adaptive filter update equation is given by the following equations:

$$\underline{\hat{w}}(n+1) = \underline{\hat{w}}(n) + \mu \underline{X}(n)\underline{C}^{-1}(n)\underline{e}(n) \tag{2.26}$$

$$\underline{C}(n) = \underline{X}^T(n)\underline{X}(n) + \delta \underline{I} \tag{2.27}$$

where $\underline{C}(n)$ is may be viewed as an estimate of the input signal autocorrelation matrix, μ is the step size parameter, and δ is a small scalar regularization parameter inserted to avoid potential stability problems while calculating the inverse of $\underline{C}(n)$. Although δ is often chosen to be an arbitrary small value, recently methods have been proposed for choosing optimal regularization parameters [56], [57].

It has been shown that Affine Projection with projection order P can decorrelate an autoregressive (AR) input signal of the same order [59]. As a result, the algorithm is popular for adaptive systems with speech input signals, because the AR model has successfully been used to model speech signals [60]. In addition, it is important to note from (2.22) to (2.27) that for a projection order of P = 1, AP reduces to the NLMS algorithm of (2.20) as a special case. As a result, AP is commonly viewed as a generalization of NLMS. The computational complexity of Affine Projection is higher than that of LMS and NLMS, requiring approximately $2NP + P_{INV}P^2$ multiplications per sample, where P_{INV} is a constant associated with calculating the inverse of the $P \times P$ matrix $\underline{C}(n)$. However, the algorithm has also become popular recently because of the

introduction of lower-complexity versions such as the Fast Affine Projection (FAP) algorithm and its variants [61], [62], [63].

2.3.2 Subband Filters and Adaptation Algorithms

Another low-complexity approach to adaptive filter design is to employ filter banks and multirate signal processing techniques [49]. Figure 2.4 shows a block diagram of a typical echo canceller employing subband adaptive filters. Subband filters employ an M-channel analysis filter bank on the input and reference signals, which decomposes the signals into a set of M subband signals each containing the content of a different band of frequencies. This frequency division allows each subband signal to be downsampled by a factor of $D \leq M$. Each of the M subband signals is processed by a corresponding adaptive filter, and the subband error signals are upsampled and reconstructed by a synthesis filter bank. There are numerous filter bank design criteria, tradeoffs, and implementation considerations, of which there is a vast literature. A survey can be found in [50], but several aspects are highlighted below.

The benefits of using subband over fullband adaptive filters are threefold. Computational savings are obtained by downsampling each of the subband signals by a factor of D, and the required length of each subband adaptive filter is typically reduced by a factor of D. In addition, as the number of subbands increases, the analysis filter bank has a decorrelating effect on each of the subband signals, potentially leading to an increased rate of convergence of each adaptive filter. Finally, employing a set of independent adaptive filters allows additional mechanisms, such as doubletalk detection, step-size control, and post-filtering, to be incorporated into each subband based on signal

characteristics localized in frequency [64], [65]. In the context of echo cancellation, one drawback of filter banks is the delay introduced into the signal path between the reference and error signals [66], [67].

Subband adaptive filters are classified as oversampled if the downsampling factor is less than the number of subbands, or $D < M$, and critically sampled if $D = M$. A maximal reduction in complexity is obtained with the latter, but it is not possible to physically realize filter banks with perfect stopband attenuation at the cutoff frequencies [43]. As a result, maximal decimation will result in aliasing in each of the subband signals. Early critically sampled subband adaptive filters were proposed with additional "cross-adaptive" filters to alleviate aliasing, but they generally suffered from poor convergence performance [25]. Recently, a critically sampled subband adaptive filter was proposed that handles aliasing while maintaining a performance increase over fullband echo cancellers [68]. This structure is of interest in this thesis and is reviewed in more detail in the following subsections. Most of the literature on subband adaptive filters has focused on the use of LMS or NLMS algorithms to adapt the adaptive filter in each subband. However, recently Affine Projection was employed to adapt each subband in an oversampled subband adaptive filter with positive results [69].

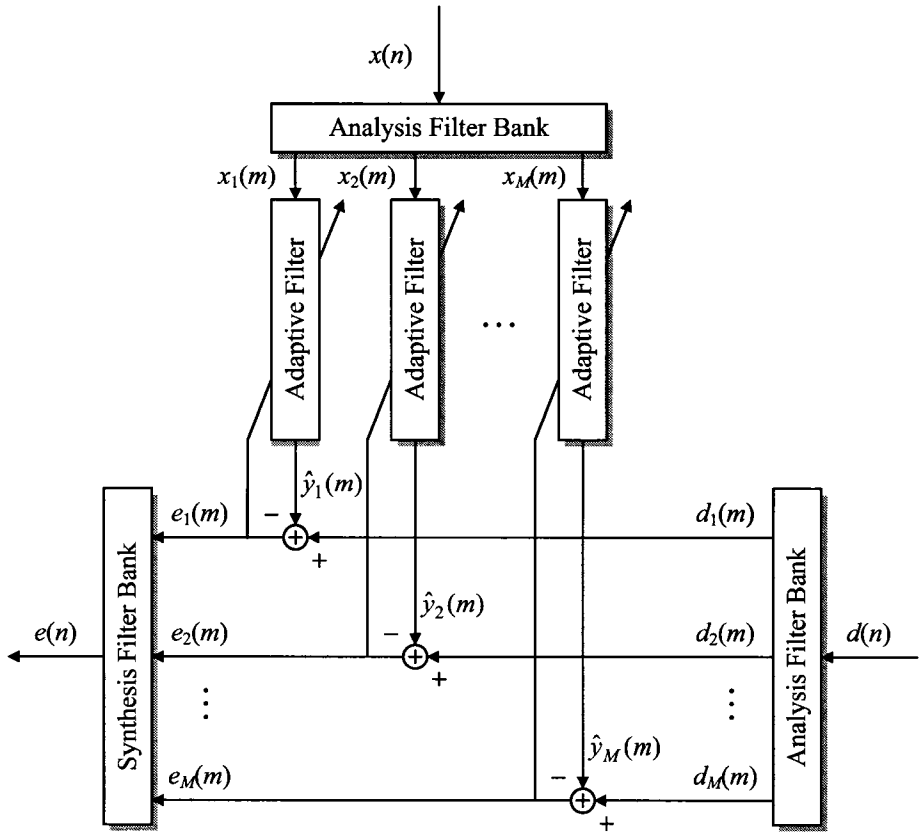


Figure 2.4 - Block diagram of a typical subband adaptive filter.

2.3.2.1 Critically Sampled Subband Adaptive Filters (CS-SBAF)

A block diagram of the critically sampled subband adaptive filter structure of [68] is shown in Figure 2.5 for the case of $M = 3$ subbands. Assume the availability of an M-channel perfect reconstruction analysis and synthesis filter bank. The input signal $x(n)$ is first passed through the analysis filter bank consisting of $2M - 1$ filters formed by convolving each of the M analysis filters with those of adjacent subbands. The reference signal $y(n)$, also containing background noise $v(n)$, is passed through the standard M-channel analysis filter bank. Each of the signals is critically downsampled by factor of $D = M$, which introduces aliasing into the subband signals. Aliasing is cancelled by

incorporating adjacent subbands as part of the M subband adaptive filter update equations.

Let $x(n)$ and $x_{i,j}(m)$ be the input and i,j^{th} subband input signals, where n and m represent normal and downsampled time indices. Similarly, let $d(n)$ and $d_i(m)$ be the reference and i^{th} subband reference signals, and let $X(z)$, $X_{i,j}(z)$, $D(z)$ and $D_i(z)$ represent the z-transforms of these signals, respectively. Let $H_i(z)$ and $F_i(z)$ be the analysis and synthesis filters of length N_H and N_F samples, respectively, for the i^{th} subband of the M-channel perfect-reconstruction filter bank. In addition, let $\underline{w}_i(m)$ be the $N_D \times 1$ adaptive filter coefficient vector for the i^{th} subband at time m, where $N_D = (N + N_F) / M + 1$, and N is the length of the fullband echo path to be modeled by the system. In all cases $1 \leq i, j \leq M$. The $2M - 1$ subband input signals $x_{i,j}(m)$ are formed by filtering the input signal with the set of analysis filters and, after filtering with the analysis filters of adjacent subbands, downsampling by a factor of M. In this structure it is assumed that the analysis filters have sufficient stopband attenuation that aliasing between nonadjacent subbands can be neglected. The subband reference signals $d_i(m)$, $1 \leq i \leq M$, are formed using only the analysis filters. In particular:

$$X_{i,j}(z^M) = X(z)H_i(z)H_j(z) \tag{2.28}$$

$$D_i(z^M) = D(z)H_i(z) \tag{2.29}$$

Finally, as shown in Figure 2.5, the fullband error signal $e(n)$ is reconstructed from the subband error signals at the outputs of the subband adaptive filters by upsampling and passing the signals through the synthesis filter bank.

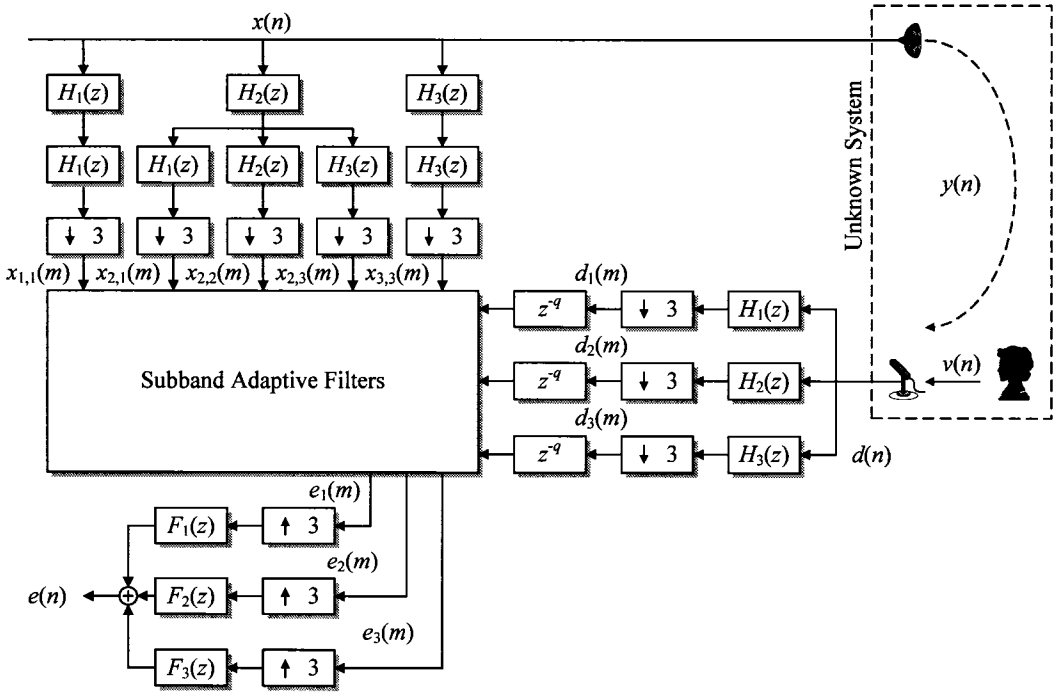


Figure 2.5 - Block diagram of the critically sampled subband adaptive filter structure for $M = 3$ subbands.

2.3.2.2 NLMS Adaptation Algorithm for CS-SBAF

In the critically sampled subband adaptive filter of [68], the analysis and synthesis filters are assumed to have sufficient stopband attenuation so that the aliasing introduced from non-adjacent subbands can be neglected. As a result, the echo signal estimate and corresponding error signal for each subband is formed from the contribution of adjacent subband input signals as well as the current one:

$$e_i(m) = d_i(m-q) - \hat{y}_i(m) \tag{2.30}$$

$$\hat{y}_i(m) = \underline{x}_{i,i-1}^T(m)\underline{\hat{w}}_{i-1}(m) + \underline{x}_{i,i}^T(m)\underline{\hat{w}}_i(m) + \underline{x}_{i,i+1}^T(m)\underline{\hat{w}}_{i+1}(m) \tag{2.31}$$

where $\underline{x}_{i,j}(m) = [x_{i,j}(m) \quad x_{i,j}(m-1) \quad \cdots \quad x_{i,j}(m-N_D+1)]^T$ is the $N_D \times 1$ vector of subband input signal samples at time m, and $q = (N_H + N_F) / 2M$ compensates for the delay introduced by the filter bank.

The adaptive filter update equation minimizes an error function comprised of the sum of squared errors across all M subbands:

$$J(m) = \sum_{i=1}^{M} e_i^2(m) \tag{2.32}$$

The resulting NLMS filter update equation for the i^{th} subband is given by an instantaneous estimate of the gradient using the current and adjacent subband input signals and the corresponding subband error signals:

$$\underline{\hat{w}}_i(m+1) = \underline{\hat{w}}_i(m) + \mu_i(m)\underline{\nabla J}_i(m) \tag{2.33}$$

$$\underline{\nabla J}_i(m) = \underline{x}_{i,i-1}(m)e_{i-1}(m) + \underline{x}_{i,i}(m)e_i(m) + \underline{x}_{i,i+1}(m)e_{i+1}(m) \tag{2.34}$$

$$\mu_i(m) = \frac{\mu}{P_{i,i-1}(m) + P_{i,i}(m) + P_{i,i+1}(m)} \tag{2.35}$$

where the normalization parameter $P_{i,j}(m)$ is the estimated power in the i,j^{th} subband input signal obtained by calculating the dot product of the vector $\underline{x}_{i,j}(m)$ with itself.

For subbands $i = 1$ and $i = M$, (2.31), (2.34), and (2.35) are adjusted to remove terms involving subbands $i - 1$ and $i + 1$, respectively. Subband adaptive filters generally introduce a fixed delay into the signal path between the fullband reference and reconstructed error signal, which may be undesirable in the context of echo cancellation. A delayless echo canceller may be implemented by reconstructing the fullband adaptive filter vector using the analysis filter bank polyphase matrix, at a cost of an increase in complexity [86].

2.4 Doubletalk Detection Algorithms

Most of the adaptation algorithms outlined in Section 2.3 assume only the presence of a far-end input signal and low levels of background noise [44]. However, full-duplex communications systems must handle doubletalk, which is the presence of near-end speech in the reference signal occurring at the same time as the echo signal [13]. The presence of near-end speech is equivalent to injecting a high noise signal into the adaptation algorithm, which in most cases causes the adaptive filter coefficients to diverge after only a few samples [70]. The result is higher residual echo for the duration of time required for the adaptive filter to re-converge. Therefore, practical echo canceller implementations employ a doubletalk detector capable of sensing the presence of near-end speech, and a control mechanism implemented as a scale factor $0 \leq \alpha(n) \leq 1$ applied to the adaptation step size to slow or halt adaptation accordingly.

Doubletalk detectors typically calculate a detection statistic $\xi(n)$ and a statistical test with the hypotheses H_0 and H_1 that doubletalk is and is not present, respectively. Many doubletalk detection algorithms have been proposed in the literature, ranging from simple energy-based approaches such as the Geigel algorithm, to more sophisticated algorithms employing cross-correlation, frequency-domain coherence, and robust statistics [13], [14], [15], [16]. Surveys and comparisons of different algorithms appear in [16] and [70]. A cross-correlation-based doubletalk detector relevant to this thesis is reviewed in the following subsection.

Two simultaneous and conflicting goals of doubletalk detection are fast and accurate identification of doubletalk conditions. Accuracy is required to avoid missing doubletalk

and to avoid unnecessary delay in adaptive filter convergence and tracking. Doubletalk detector performance is somewhat subjective because of the possibility of adaptive filter divergence and the effect of residual echo on near-end speech quality. However, an objective method was proposed for evaluating the performance of doubletalk detection algorithms using the probability of detection as a function of the power of near-end speech in relation to the echo signal [70].

2.4.1 Cross-Correlation and Power-Based Algorithms

Improved detection performance and noise immunity can be obtained by employing doubletalk detectors based on measures of cross-correlation between far- and near-end signals involved in echo canceller operation. An early example is the algorithm proposed in [14] based on the cross-correlation vector between the far-end input signal $x(n)$ and the echo canceller error signal $e(n)$. A more robust doubletalk detector was proposed in [16] based on the normalized cross-correlation vector between the input signal $x(n)$ and the reference signal $d(n)$, and briefly reviewed as follows. First of all, assume stationarity of the echo path coefficient vector $\underline{w}(n)$ and the far-end input signal $x(n)$. Under these conditions, the expected variance of the reference signal can be written in terms of the $N \times 1$ cross-correlation vector between the input and reference signal, and the $N \times N$ autocorrelation matrix of the input signal. The doubletalk detection statistic is obtained by normalizing the expected echo signal variance by the measured reference signal variance, and then taking the square root of the result:

$$\sigma_y^2 = \underline{w}^T \underline{R}_{xx} \underline{w} = \underline{r}_{xd}^T \underline{R}_{xx}^{-1} \underline{r}_{xd} \tag{2.36}$$

$$\xi = \sqrt{\frac{\underline{r}_{xd}^T \underline{R}_{xx}^{-1} \underline{r}_{xd}}{\sigma_d^2}} \tag{2.37}$$

Equation (2.37) is simply the ratio of the expected and actual reference signal variances. The former is constructed from the input signal and the cross-correlation vector, while the latter is obtained from the reference signal itself. In the absence of doubletalk the numerator and denominator terms are equal and $\xi = 1$. When doubletalk is present, the actual reference signal variance is larger and $\xi < 1$. Taking the expected value of (2.37) reveals that the detection statistic's average value is a function of the near-end speech to echo signal power ratio (NER), which is plotted in Figure 2.6:

$$E[\xi] = \sqrt{\frac{\sigma_y^2}{\sigma_y^2 + \sigma_v^2}} = \sqrt{\frac{1}{1 + \sigma_v^2/\sigma_y^2}} = \sqrt{\frac{1}{1 + \mathrm{NER}}} \tag{2.38}$$

In practice the input and near-end speech signals are time-varying, and the parameters of (2.37) must be estimated and tracked. A direct implementation suffers from the drawback of having to construct estimates of the autocorrelation matrix and calculate its inverse. This is expensive for the long impulse responses typical of acoustic environments (up to 250 ms). Therefore, a simplification from [16] is to assume that the adaptive filter coefficients have converged, at which time the numerator can be simplified with the following:

$$\underline{R}_{xx}^{-1} \underline{r}_{xd} = \underline{w} \approx \underline{\hat{w}}(n) \tag{2.39}$$

A straightforward approach to estimating the cross-correlation vector of (2.37) is by averaging over a window of K previous samples. Similarly, an unbiased estimate of the reference signal variance at time n can be obtained as follows [101]:

$$\underline{\hat{r}}_{xd}(n) = \frac{1}{K}\sum_{k=0}^{K-1}\underline{x}(n-k)d(n-k) \tag{2.40}$$

$$\hat{\sigma}_d^2(n) = \frac{1}{K-1}\sum_{k=0}^{K-1}\left[d(n-k) - \frac{1}{K}\sum_{j=0}^{K-1}d(n-j)\right]^2 \tag{2.41}$$

Substituting (2.39) – (2.41) into (2.37) results in an estimated doubletalk detection statistic at time n:

$$\xi(n) = \sqrt{\frac{\underline{\hat{r}}_{xd}^T(n)\underline{\hat{w}}(n)}{\hat{\sigma}_d^2(n)}} \tag{2.42}$$

For a stationary input signal and echo path, the accuracy of the parameter estimates of (2.40) and (2.41) increase with the estimation window size. However, the per-sample computational complexity also increases with K and, as will be shown in Chapter 4, there is a tradeoff between the accuracy of detection and the speed of detection after the onset of doubletalk conditions.

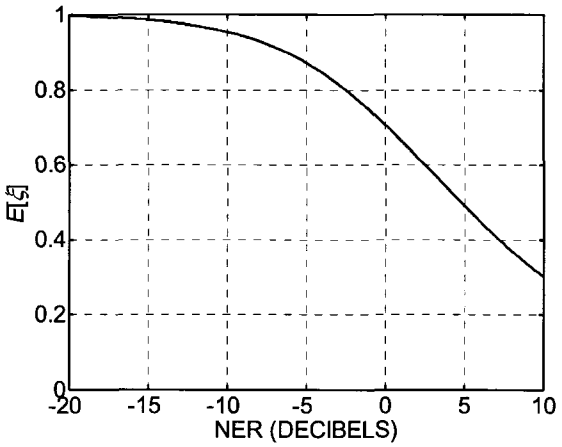


Figure 2.6 - Expected value of the doubletalk detection statistic $\xi(n)$ as a function of near-end speech to echo signal power ratio (NER).

2.5 Post-Filtering Algorithms

The presence of an echo canceller may not be enough to sufficiently cancel the echo signal, resulting in a residual echo component $\delta(n)$ in the error signal. Residual echo occurs during initial convergence and tracking of the echo path, when error in the adaptive filter coefficients induces error in the estimated echo signal. After the filter has converged, under-modeling of the echo path ($N < L$) results in residual echo if the true echo path contains significant components past the adaptive filter order. Another cause is the presence of nonlinearity in the echo path, particularly from analog hybrid transformers and from loudspeakers in acoustic environments, which cannot be modeled by a linear echo canceller [10]. Therefore, as shown in Figure 2.3, a practical echo canceller implementation typically contains a nonlinear processor (NLP) or post-filter at the echo canceller output to provide additional echo suppression.

During singletalk periods, the goal of the NLP is to suppress residual echo so that it is at least tolerable to the far-end user, and preferably no longer perceivable. Classical suppression algorithms are center-clipping and adaptive whitening [71], [72]. During doubletalk conditions, the presence of residual echo will impair the perceived quality of the near-end speech, so in these periods the goal is to suppress residual echo while minimizing distortion of near-end speech. Typical approaches are implemented in the frequency domain and drawn from the speech enhancement literature. Key problems, outlined below, is that there is no easy way to "switch" between NLP and frequency-domain post-filtering algorithms.

2.5.1 Center-Clipping

Residual echo is typically a low-amplitude signal, and so a classical NLP, the center-clipper, passes the echo canceller error signal through a nonlinear function that attenuates samples with amplitudes smaller than some threshold γ[71]:

$$e'(n) = \begin{cases} 0, & |e(n)| \leq \gamma \\ e(n), & |e(n)| > \gamma \end{cases} \tag{2.43}$$

Center-clipping is effective at suppressing residual echo during singletalk periods, and introduces no delay into the error signal. During doubletalk conditions or periods of no far-end speech, the NLP must be disabled to avoid attenuating or distorting near-end speech signals. Therefore, the performance of NLP depends on the accuracy of the doubletalk detector [73]. In environments with moderate levels of background noise, selectively enabling and disabling the NLP may result in "noise modulation", where background noise is audible during doubletalk or near-end-speech-only periods, and suppressed during singletalk periods. Another consequence is that disabling the post-filter during doubletalk may result in audible residual echo during those periods.

2.5.2 Frequency-Domain and Psychoacoustic Post-Filtering

The goal of frequency-domain post-filtering algorithms is to suppress residual echo during doubletalk conditions by employing algorithms from the speech enhancement literature [74]. In particular, the echo canceller error signal is modeled as a (desirable) near-end speech signal corrupted by residual echo and background noise. A time-varying linear filter is constructed and applied to the error signal to enhance the near-end speech while suppressing the residual echo and noise. As shown in Figure 2.7, the error signal

$e(n)$ is transformed by an analysis stage into a frequency domain representation $E(\omega)$ consisting of the spectrums of near-end speech $V(\omega)$, residual echo $\Delta(\omega)$, and background noise $N(\omega)$. A weighting function $H(\omega)$ is applied to enhance the near-end speech before reconstruction by a synthesis stage:

$$e(n) = v(n) + \delta(n) + \eta(n) \leftrightarrow E(\omega) = V(\omega) + \Delta(\omega) + N(\omega) \tag{2.44}$$

$$\hat{V}(\omega) = H(\omega)E(\omega) = H(\omega)[V(\omega) + \Delta(\omega) + N(\omega)] \tag{2.45}$$

A survey of frequency-domain post-filtering algorithms appears in [75]. A generic approach to designing the weighting function is to minimize the distortion of near-end speech while suppressing the residual echo below some specified level [76]. Define a cost function $J(\omega)$ as the squared error between the true and estimated near-end speech frequency representations. Assuming statistical independence between the near-end speech, residual echo, and background noise, $J(\omega)$ can be written as the sum of three cost functions $J_V(\omega)$, $J_\Delta(\omega)$, and $J_N(\omega)$:

$$J(\omega) = [V(\omega) - \hat{V}(\omega)]^2 = J_V(\omega) + J_\Delta(\omega) + J_N(\omega) \tag{2.46}$$

$$J_V(\omega) = [1 - H(\omega)]^2 S_{VV}(\omega) \tag{2.47}$$

$$J_\Delta(\omega) = H^2(\omega) S_{\Delta\Delta}(\omega) \tag{2.48}$$

$$J_N(\omega) = H^2(\omega) S_{NN}(\omega) \tag{2.49}$$

where $S_{VV}(\omega)$, $S_{\Delta\Delta}(\omega)$, and $S_{NN}(\omega)$ are the power spectral density (PSD) functions of the near-end speech, residual echo and background noise, respectively. $J_V(\omega)$ represents the distortion of near-end speech, while $J_\Delta(\omega)$ and $J_N(\omega)$ represent the residual echo and noise

PSD after post-filtering. A Wiener weighting function can be constructed to minimize $J_V(\omega)$ [76]:

$$H(\omega) = \frac{S_{VV}(\omega)}{S_{VV}(\omega) + S_{\Delta\Delta}(\omega) + S_{NN}(\omega)} \tag{2.50}$$

To employ (2.50), it is necessary to have estimates of the near-end speech and residual echo PSD functions. In [104] it is assumed that the background noise is negligible, and independence of the near-end speech and residual echo signals is used to obtain an estimate of $S_{VV}(\omega)$ from the cross-PSD between the error and reference signal, or via spectral subtraction:

$$S_{VV}(\omega) \approx S_{EE}(\omega) - S_{\hat{Y}\hat{Y}}(\omega) \approx S_{ED}(\omega) \tag{2.51}$$

Frequency-domain post-filtering structures for echo cancellation and speech enhancement have also been proposed incorporating psychoacoustic models of human hearing [76]. In these approaches, the post-filter suppresses the residual echo to a level at or below the masking threshold induced by the near-end speech signal. As shown in Figure 2.8, in this approach a preliminary estimate of the near-end speech power spectrum is constructed from the error signal using an estimate of the residual echo power spectrum. The masking threshold $T_M(\omega)$ of the near-end speech is then obtained using a psychoacoustic model such as [106]. Finally, $H(\omega)$ is constructed under the assumption that residual echo below the masking threshold will be inaudible to the listener [107].

Recall the frequency-domain cost functions of (2.46) – (2.49). Minimizing $J_V(\omega)$ such that $J_\Delta(\omega)$ is at the masking threshold $T_M(\omega)$ of the near-end speech results in a real-valued transfer function given by [76]:

$$H(\omega) = \max\left\{\sqrt{\frac{T_M(\omega)}{S_{\Delta\Delta}(\omega)}}, 1\right\} \tag{2.52}$$

In order to accurately construct the masking threshold of the near-end speech, it is important to have a "good" estimate of the residual echo power spectrum. From (2.52) it is clear that inaccuracies in $T_M(\omega)$ or $S_{\Delta\Delta}(\omega)$ may lead to audible distortion of the near-end speech and / or insufficient residual echo suppression. In addition, the computational complexity of post-filtering is typically much higher than center-clipping and adaptive whitening due to the time- or frequency-domain post-filtering operation, as well as the cost of constructing and adapting the weighting function $H(\omega)$. Finally, the use of frequency-domain algorithms in general implies that there is delay introduced into the error signal due to block processing.

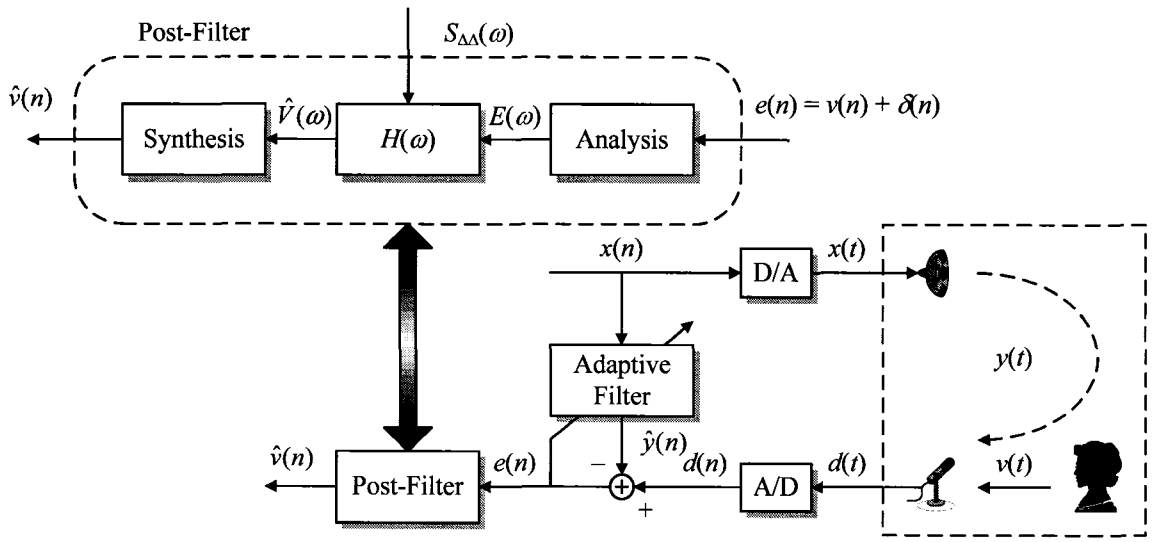


Figure 2.7 - Block diagram of a typical frequency-domain post-filter for residual echo suppression and near-end speech enhancement.

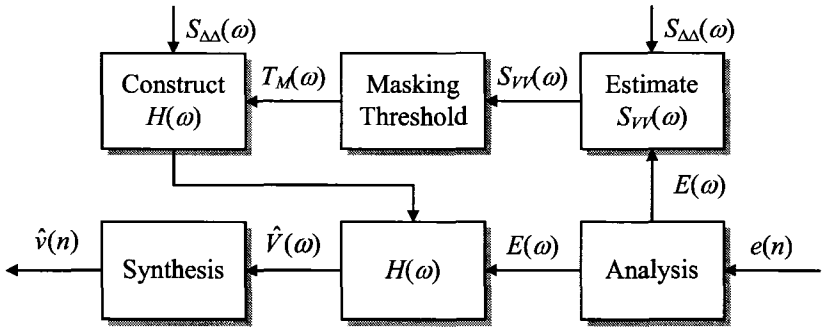


Figure 2.8 - Block diagram of a psychoacoustic post-filter for suppressing residual echo from near-end speech.

2.6 Echo Cancellation in VoIP

As discussed in Chapter 1, a current trend is the migration of voice services from the legacy public switched telephone network (PSTN) to integrated voice-and-data services delivered over a common packet-switched network, or VoIP [4], [5], [6], [77]. Along with that trend is a host of supporting technologies required for high-quality voice transmission over unreliable packet-switched networks such as the Internet. Figure 2.9 shows a block diagram of a typical VoIP configuration where the PSTN is interconnected with an IP-based network through an IP gateway [5]. As shown in the figure, a typical implementation employs LPC-based low-bit-rate compression algorithms, such as ITU-T G.729A or G.723.1, to compress toll-quality speech from 64 kbit/sec down to 5.3 – 8 kbit/sec [18], [19]. A playback (or "jitter") buffer is also commonly employed at the receiving end to collect frames of compressed speech for decoding and playback [80]. The buffer is necessary to compensate for the transmission delay of data through packet-switched networks, and is typically adaptive to track the network's variable transmission delay. The combination of network transport, playback, and codec delays has the effect of increasing the end-user perception of echo at the far end [11].

In addition to speech compression and reconstruction components, new signaling and transport protocols have been developed for call setup and transmission of voice packets across an inherently unreliable network. Although outside the scope of this thesis, popular standards are the Session Initiation Protocol (SIP) and Real-Time Protocol (RTP) [78], [79].

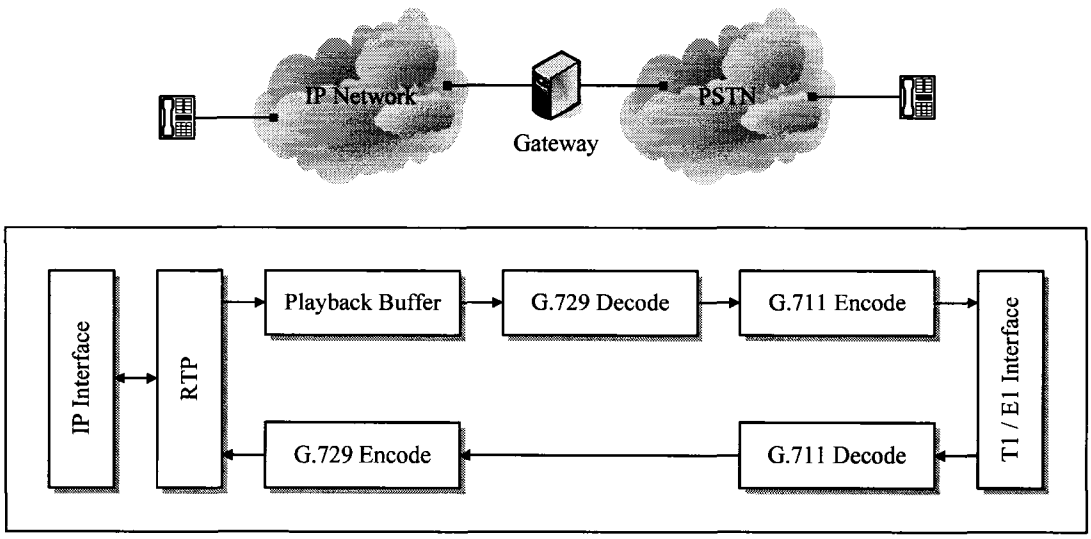


Figure 2.9 - Block diagram of components at an IP / PSTN gateway. In this example configuration, the ITU-T G.729 codec is employed on the IP side, and G.711 (mu-law) encoding is employed on the PSTN side.

2.6.1 Analysis-By-Synthesis Speech Compression

Most low-bit-rate speech coders are based on linear predictive coding (LPC) techniques and an autoregressive model of speech [81], [82]. A typical encoder segment signals into frames of 10 – 30 ms in duration, and for each frame determines a parametric representation of the signal using analysis-by-synthesis techniques. The result is an optimal set of parameters which are quantized and transmitted to the receiver. Figure 2.10 shows a typical structure employed by an LPC-based analysis-by-synthesis encoder, and its parameters and operation are reviewed as follows. First of all, an excitation signal

$c(n)$ is constructed from a codebook and scaled by a time-varying codebook gain parameter $g_c(n)$. A long-term synthesis filter models the pitch period of voiced speech periods, and is represented by a pitch period $p(n)$ and a pitch gain $g_p(n)$ (or adaptive codebook gain). The excitation signal is filtered by the long-term synthesis filter to form the short-term excitation signal $r(n)$:

$$r(n) = g_p(n)r(n - p(n)) + g_c(n)c(n) \tag{2.53}$$

A time-varying short-term synthesis filter is used to recreate the broad spectrum of the speech signal, and is represented by a m^{th}-order set of LPC coefficients $a_m(n, i)$, for $0 \leq i \leq m$. Similarly, the short-term excitation signal is filtered by the short-term synthesis filter to form the reconstructed speech signal for the frame:

$$\tilde{x}(n) = \sum_{i=1}^{m} a_m(n,i)\tilde{x}(n-i) + r(n) \tag{2.54}$$

The steps in the encoding process can be summarized as follows. First the LPC coefficients are obtained using the Levinson-Durbin algorithm and an estimate of the autocorrelation function calculated from the input signal [44]. The LPC coefficients are then used to construct a weighting filter $w_p(n)$ that emphasizes perceptually significant frequencies over those with a higher noise masking threshold. A closed-loop search is then performed to find the remaining parameters that minimize the weighted mean squared error between the original and reconstructed speech signals. In particular:

$$e(n) = x(n) - \tilde{x}(n) \approx x(n) - r(n) \otimes h(n) \tag{2.55}$$

$$e_w(n) = e(n) \otimes w_p(n) \tag{2.56}$$

where $h(n)$ is the impulse response of the short-term synthesis filter, and $\otimes$ denotes the convolution operator. The optimal parameters minimizing the mean squared error of (2.56) are compressed and transmitted to the decoder.

It is expensive to find an optimal encoding for the speech signal by performing a brute-force search over the entire space of the parameters $c(n)$, $g_c(n)$, $p(n)$ and $g_p(n)$. As a result, many methods exist for reducing the complexity of the search process, such as calculating a per-frame open-loop pitch estimate p_{OL}, using structured codebooks, and segmenting frames into smaller subframes of 5 – 7.5 ms, over which the pitch period and gain parameters are held constant [81]. By inspection of Figure 2.10, it can be seen that the perceptual weighting filter $w_p(n)$ may be applied to $x(n)$ and $h(n)$ before the closed-loop search is performed:

$$x_w(n) = x(n) \otimes w_p(n) \tag{2.57}$$

$$h_w(n) = h(n) \otimes w_p(n) \tag{2.58}$$

$$e_w(n) = x_w(n) - r(n) \otimes h_w(n) \tag{2.59}$$

A coarse estimate of the pitch period $p(n)$ is usually obtained by first locating an open-loop pitch period p_{OL} from on the perceptually weighted input signal $x_w(n)$. The closed-loop search for $p(n)$ is then limited to a more narrow subset of delays around the open-loop estimate. Finally, most vocoders also employ highly structured codebooks to compactly represent the excitation signal $c(n)$ and to facilitate faster closed-loop searches minimizing (2.56).

A block diagram of a typical LPC-based speech decoder is shown in Figure 2.11. It mainly consists of the encoder's synthesis structure, although in this case the parameters

are only used to reconstruct the speech signal using (2.53) and (2.54). As a result the decoder is much less computationally expensive. In addition to the long- and short-term synthesis filters, a harmonic post-filter is often employed after the short-term synthesis filter [81]. The purpose of the post-filter is to increase the quality of the reconstructed speech signal by masking noise introduced by parameter quantization at the encoder.

It should be noted that many variations of the LPC-based speech production model described above exist in practice [81]. A key difference between them is the method used to model and encode the excitation signal $r(n)$. For example, hybrid speech coders such as multi-pulse excitation (MPE) and regular-pulse excitation (RPE) techniques model the excitation signal as a series of pulses which may explicitly include the pitch period [81]. In this case the excitation signal $c(n)$ is equivalent to the short-term excitation signal $r(n)$, and the pitch synthesis filter of (2.54) is not required.

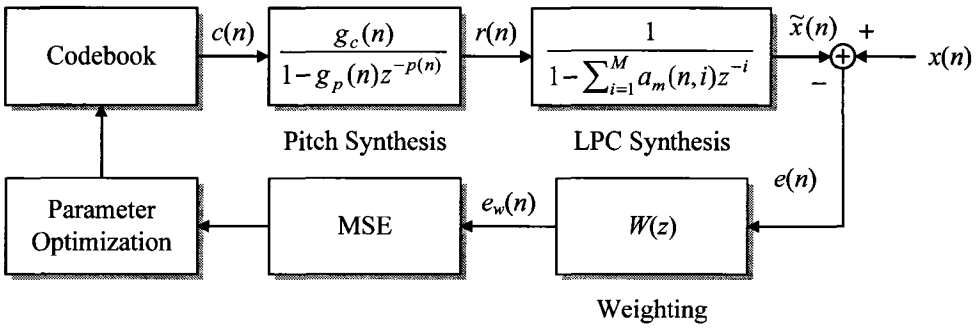


Figure 2.10 - Block diagram of a typical analysis-by-synthesis LPC-based speech encoder.

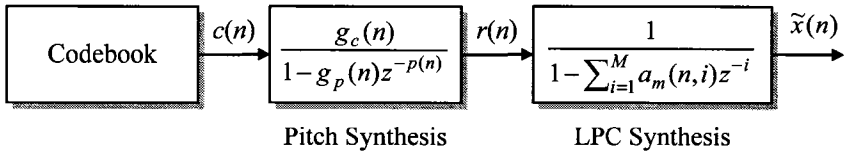


Figure 2.11 - Block diagram of a typical LPC-based speech decoder.

2.6.2 Centralized Echo Cancellation in VoIP

Digital echo cancellers are ideally deployed as close as possible to echo sources, as shown in Figures 2.1 and 2.2. However, existing echo cancellers may not provide a sufficient level of cancellation given the increased round-trip delays introduced by VoIP and mobile networks [11]. In addition, interconnections between IP-based networks and the PSTN complicate decisions of where to deploy echo cancellers. One solution is to employ *centralized* echo cancellers at IP gateways and mobile switching centers (MSC) [104], [105]. However, a unique problem is that send and receive paths in the PSTN may be transparently replaced with an all-IP core network, with speech coding and decoding introduced at the network edge [103]. If an echo canceller is deployed facing into the network, as shown in Figure 2.12, the presence of vocoders introduces distortion into the echo path that manifests itself as residual echo [21]. Another unique problem introduced by vocoders arises when signal processing operations must be performed on compressed speech, such as audio mixing [35]. Linear mixing for conferencing systems is often performed at gateways, and the operations of decoding, mixing, and re-encoding speech signals introduces further delay and distortion.

As noted in Section 2.5, frequency-domain post-filtering approaches to residual echo suppression require an estimate of the residual echo power spectrum. One linear estimation technique for undermodeled acoustic echo cancellers was proposed in [109], but is ineffective for the nonlinear distortion introduced by vocoders. More recent approaches estimate the residual echo power spectrum as a linearly-weighted estimate of the estimated echo power spectrum, but address only the simplified case of speech encoders / decoders along the receive path [104], [105].

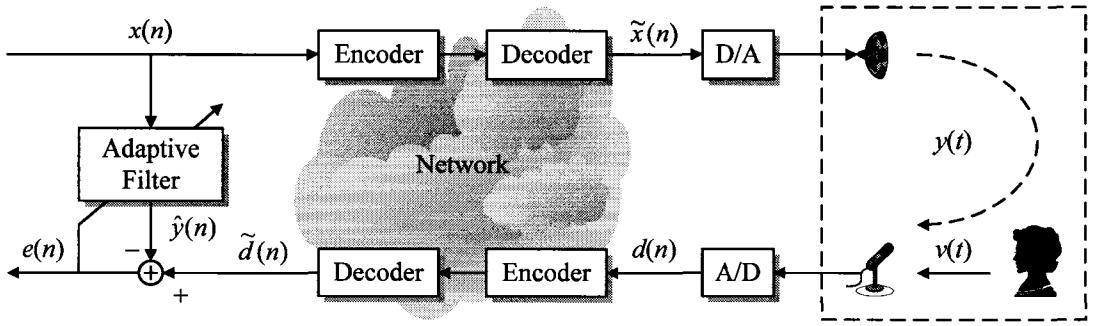


Figure 2.12 - Echo canceller in a network with vocoder distortion introduced into the input (send path) and reference (receive path) signals.

Chapter 3 Experimental Setup

3.1 Physical and Simulation Environments

Experiments were performed in a small conference room (MC 3033) measuring 12 feet, 6 inches (3.8 meters) wide by 17 feet, 6 inches (5.3 meters) long. Flooring consisted of commercial-grade tiles, and three of the four walls were standard plaster with one wall of bare cement. The room contained one conference table, one desk, and eight chairs. The layout of the room is shown in Figure 3.1, along with the approximate locations of the loudspeaker and microphone. Far-end signals were played through a Tannoy Reveal loudspeaker, and the reference signal was simultaneously recorded using an Audio Technica microphone (ATM-803b) and a Pioneer SX-201 receiver. Audio samples were digitized with an M-Audio Delta 1010 sound card providing a 24-bit analog-to-digital converter (ADC). Adobe Audition 1.5 was used to control playback and recording sessions. Simulations were performed on the data off-line using a Dell Dimension 4550 computer consisting of a 2.5 GHz Intel Pentium 4 processor with 512 MB of RAM running Microsoft Windows XP. Simulations were performed using MATLAB 7.0.1, and compressed speech samples for Chapter 7 were obtained using the reference fixed-point "C" implementation of G.729A provided by the ITU [18].

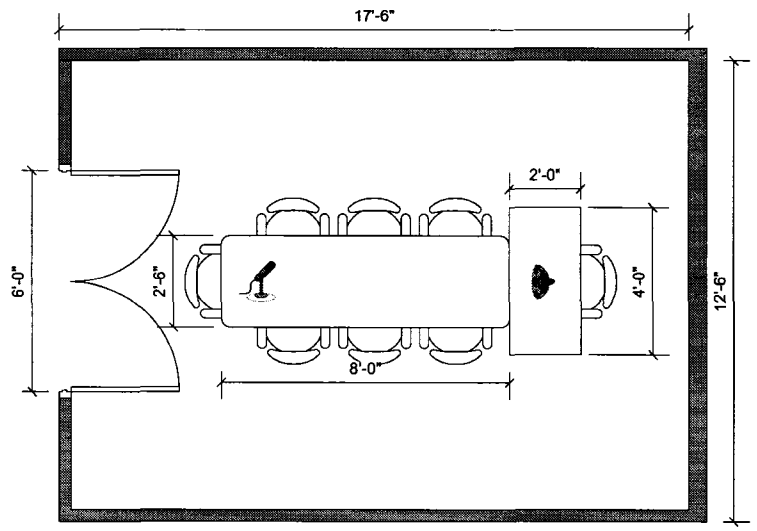


Figure 3.1 - Dimensions and layout of objects within conference room (MC 3033).

3.2 Echo Path Impulse Responses

An estimate of the impulse response of the conference room of Figure 3.1 was obtained by playing Gaussian white noise from the loudspeaker and simultaneously recording the microphone signal for five minutes. Normalized LMS was used offline to construct an estimate of the room impulse response using the playback and recorded data signals with a step size of $\mu = 0.1$, which was then truncated to $L = 500$ samples. A plot of the resulting room impulse response is shown in Figure 3.2. Several synthetic impulse responses were also constructed by modulating a Gaussian white noise process of unity variance with an exponentially decaying envelope in accordance with:

$$w_n = \eta(n)\exp(-\alpha n) \tag{3.1}$$

where $\eta(n)$ is the white noise process at time n, and $\alpha > 0$ is a small constant controlling the rate of decay. Two sets of synthetic impulse responses were constructed, one with a decay parameter of $\alpha = 0.01$ for $N = 500$ samples, and a second with $\alpha = 0.005$ for $N = 2000$ samples. For the latter case, a secondary response was generated to represent an

additional echo signal occurring at N = 1000 samples. Plots of the respective impulse responses are shown in Figure 3.3 and 3.4.

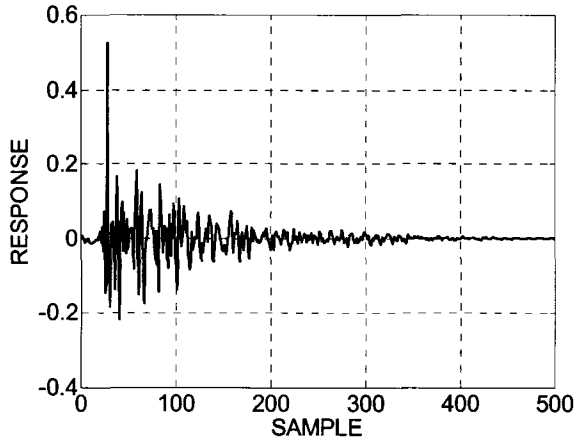


Figure 3.2 - Plot of room impulse response captured from MC 3033 (N = 500 samples).

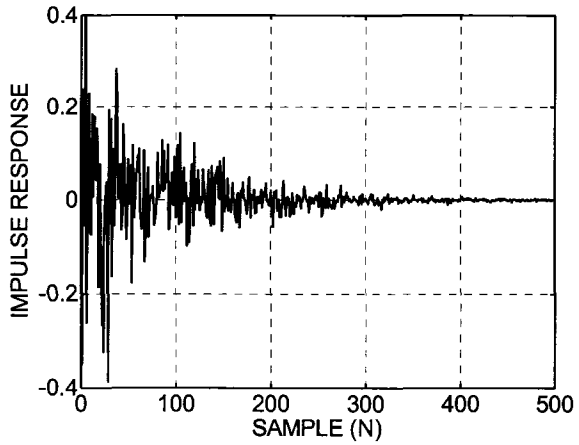


Figure 3.3 - Plot of synthetic room impulse response (N = 500 samples).

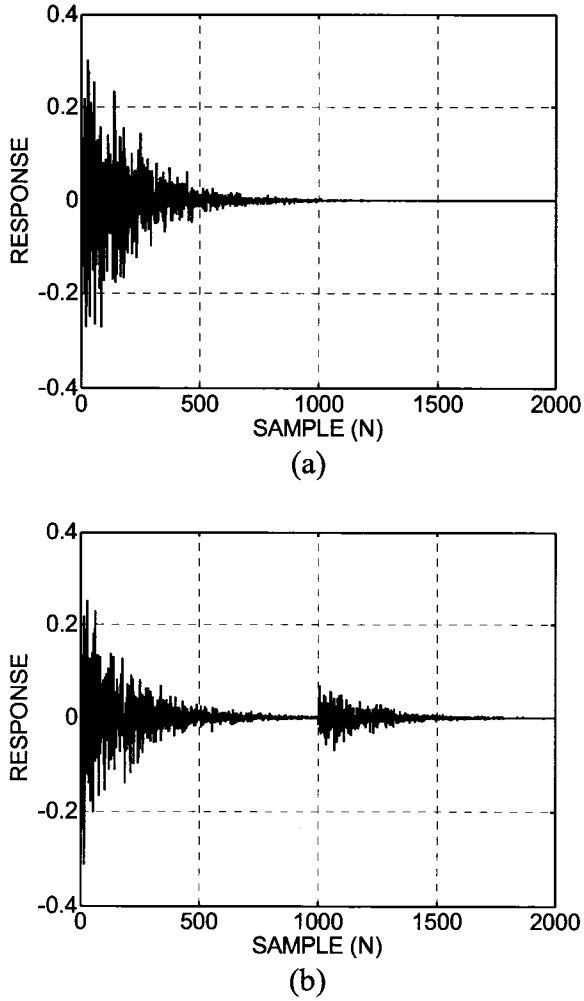


Figure 3.4 - Plot of synthetic room impulse responses (N = 2000 samples).

Chapter 4 Robust Calibration of Doubletalk Detectors

4.1 Overview

As described in Chapter 2, in practical echo canceller implementations a doubletalk detector is required to sense the presence of near-end speech in the reference signal. Most doubletalk detectors calculate a detection statistic and compare it to some pre-defined threshold [70]. However, one problem is choosing a threshold that is "optimal" in some sense, and also appropriate for different echo path environments. In this chapter the problem of doubletalk detector calibration is approached by applying statistical analysis, and applied to the normalized cross-correlation-based doubletalk detector of [16]. It is shown that this model is useful for constructing statistically optimal detection thresholds, and to determine expected performance in different environments.

The sections of this chapter are organized as follows. Section 4.2 discusses implementation issues of the cross-correlation-based doubletalk detection algorithm. Section 4.3 presents a statistical analysis of the algorithm's behavior in the absence and presence of doubletalk conditions. Section 4.4 proposes calibration algorithms based on this statistical model, and presents simulation results. Finally, in Section 4.5 the primary results and conclusions of this chapter are summarized.

4.2 Doubletalk Detector Implementation and Calibration Issues

Recall the cross-correlation-based doubletalk detector of [16] reviewed in Section 2.4.1 and, in particular, the parameter estimates described in (2.39) – (2.41). Using these

equations to calculate a time-varying doubletalk detection statistic introduces a number of issues in practice. First of all, assuming convergence of the adaptive filter coefficients in (2.39) leaves $\xi(n)$ vulnerable to abrupt changes in the true echo path impulse response $\underline{w}(n)$, which may be reported erroneously as doubletalk conditions [94]. The problem of differentiating between doubletalk and echo path changes has been studied before [95], [96], [97]. However, even if (2.39) holds, in practice the echo path impulse response is typically at least slowly time-varying, resulting in variability in the adaptive filter coefficients. In addition, the fact that (2.40) and (2.41) are estimated over a window of K samples implies that the doubletalk detector's accuracy is dependent on the window size. Finally, the presence of background noise $\eta(n)$ in the environment will increase the reference signal variance estimated using (2.41), which will in turn bias the detection statistic calculated using (2.42).

Since the parameter estimates of (2.39) – (2.41) inherently contain error and possibly bias due to noise, $\xi(n)$ is a random process with a probability density function (PDF) centered (ideally) at 1 in the absence of doubletalk, and at some value less than one in the presence of doubletalk. Therefore, a decision can be made by comparing $\xi(n)$ to a threshold T, below which doubletalk is declared to be present:

$$\xi(n) < T \Rightarrow H_0 \quad \xi(n) > T \Rightarrow H_1 \tag{4.1}$$

where H_0 and H_1 are the hypotheses that doubletalk is and is not present, respectively.

An important implementation problem is how to select the detection threshold T in practice. Figure 4.1 illustrates the statistical importance of the threshold graphically, showing the conditional PDFs of a typical double detection statistic given no doubletalk

and doubletalk. Choosing too low a threshold will increase the probability of miss P_M (Type I error), which is critical as it may cause the adaptive filter to diverge. The false alarm probability P_F (Type II error) increases with a higher threshold and is less critical, but a high false alarm rate will slow the echo canceller's convergence and tracking capability [70]. If the near-end speech characteristics are not known (which is generally the case), one approach is to choose T for a prescribed maximum false alarm probability P_F conditional on the absence of doubletalk. A second approach is to make some assumptions of the near-end speech and select the threshold for a maximum probability of miss P_M conditional on the presence of doubletalk:

$$P[\xi(n) < T \mid H_1] \leq P_F \tag{4.2}$$

$$P[\xi(n) > T \mid H_0] \leq P_M \tag{4.3}$$

Equations (4.2) and (4.3) assume *a priori* knowledge of the doubletalk detection statistic's conditional PDFs, which in general is not known. In [70] an approach was presented for choosing T by using training speech signals to gather empirical data on the detection statistic. An alternative approach is proposed in the following section by deriving PDFs the doubletalk detection statistic of (2.42) in the absence and presence of doubletalk, allowing (4.2) or (4.3) to be used for calibration and evaluation purposes.

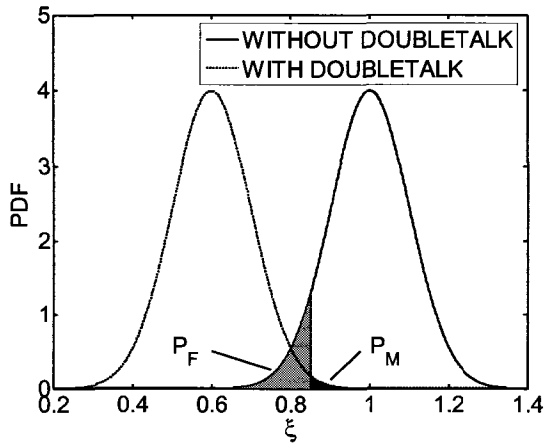


Figure 4.1 - Typical doubletalk detection statistic PDFs conditional on the absence and presence of doubletalk, with probability of miss (P_M) and false alarm (P_F) for threshold T.

4.3 Statistical Analysis of Doubletalk Detection

4.3.1 Assumptions

The following assumptions are employed to ease the task of constructing a statistical model. First of all, it is assumed that the approximation of (2.39) holds in general, which implies that the adaptive filter coefficients have converged and any variations in the time-varying true echo path impulse response $\underline{w}(n)$ are sufficiently slow so that (2.39) continues to hold. The actual difference between the two vectors is represented by $\Delta\underline{w}(n)$, the adaptive filter error vector at time n:

$$\hat{\underline{w}}(n) = \underline{w}(n) - \Delta\underline{w}(n) \tag{4.4}$$

It is also assumed that the input signal $x(n)$ and near-end speech signal $v(n)$ are zero-mean uncorrelated random processes that are stationary within the estimation windows of (2.40) and (2.41). Background noise $\eta(n)$ is assumed to be stationary, white and Gaussian with zero mean and variance σ_η^2, and uncorrelated with the input and near-end speech signals. It has been proposed to model speech signals with Gaussian, Laplacian,

or more complicated probability density functions [99]. However, in this chapter no particular distributions for $x(n)$ and $v(n)$ are assumed.

4.3.2 Probability Distribution in the Absence of Doubletalk

Note from (2.40) that the estimated cross-correlation vector is formed by averaging the scalar product of the input signal vector $\underline{x}(n)$ and reference signal $d(n)$. As a result, the numerator of (2.42) can be viewed as a sum of random variables weighted by the adaptive filter coefficients. From statistics, under general conditions the PDF of the sum of a large number of random variables approaches a Gaussian distribution even when the random variables themselves are not Gaussian [100]. Using (4.4), the numerator of (2.42) can be written in terms of the true echo path impulse response and bias from the adaptive filter error at time n:

$$\underline{\hat{r}}_{xd}^{T}(n)\underline{\hat{w}}(n) = \underline{\hat{r}}_{xd}^{T}(n)\underline{w}(n) - \underline{\hat{r}}_{xd}^{T}(n)\Delta\underline{w}(n) \tag{4.5}$$

From the reasoning above, (4.5) can be modeled as the difference of two Gaussian random variables. If the input signal autocorrelation matrix $\underline{R}_{xx}$ is Toeplitz, then the adaptive filter error bias may be written in terms of the cross-correlation vector between the input and error signals, and the adaptive filter coefficient vector at time n:

$$\begin{aligned} \underline{r}_{xd}^{T}(n)\Delta\underline{w}(n) &= \underline{w}^{T}(n)\underline{R}_{xx}\Delta\underline{w}(n) = \Delta\underline{w}^{T}(n)\underline{R}_{xx}\underline{w}(n) \\ &= \underline{r}_{xe}^{T}(n)[\underline{\hat{w}}(n) + \Delta\underline{w}(n)] \\ &= \underline{r}_{xe}^{T}(n)\underline{\hat{w}}(n) + \sigma_{\delta}^{2}(n) \end{aligned} \tag{4.6}$$

where $\sigma_{\delta}^{2}(n)$ is the variance of the residual echo signal of the echo canceller output at time n. In general the bias is nonzero and time-varying due to changes in the echo path impulse response and adaptive filter error. One approach to correct this bias is to

compute estimates of the parameters of (4.6) over a window of K samples similar to (2.40) and (2.41):

$$\hat{\underline{r}}_{xe}(n) = \frac{1}{K}\sum_{k=0}^{K-1}\underline{x}(n-k)e(n-k) \tag{4.7}$$

$$\hat{\sigma}_{\delta}^{2}(n) \approx \hat{\sigma}_{e}^{2}(n) = \frac{1}{K-1}\sum_{k=0}^{K-1}\left[e(n-k) - \frac{1}{K}\sum_{j=0}^{K-1}e(n-j)\right]^{2} \tag{4.8}$$

Using the error signal to estimate the residual echo variance will include the variance of background noise. For a converged echo canceller, the background noise will be at or higher than the bias introduced by the residual echo. Since the background noise is assumed to be stationary, it can be represented as a constant term estimated during quiet periods, so $\sigma_{\delta}^{2}(n) \approx \sigma_{\eta}^{2}$. When the latter is added to the numerator of (2.41) to compensate for bias, this will also compensate for noise in the reference signal variance:

$$\xi(n) = \sqrt{\frac{\hat{\underline{r}}_{xd}^{T}(n)\hat{\underline{w}}(n) + \sigma_{\eta}^{2}}{\hat{\sigma}_{d}^{2}(n)}} \approx 1 \tag{4.9}$$

It is known from statistics that Equation (2.41) produces an unbiased estimate of the reference signal variance, with mean and variance given as follows [101]:

$$E[\hat{\sigma}_{d}^{2}(n)] = \mu_{2} \tag{4.10}$$

$$VAR[\hat{\sigma}_{d}^{2}(n)] = \frac{(K-1)\mu_{4} - (K-3)\mu_{2}^{2}}{K(K-1)} \tag{4.11}$$

where μ_2 and μ_4 are the second and fourth central moments of the underlying signal, respectively. Note that $d(n)$ contains echo $y(n)$ plus background noise $\eta(n)$, the former of which is formed by the convolution of the input signal with the echo path impulse response, or a weighted sum of samples from $x(n)$. Therefore, it is assumed that the

reference signal is Gaussian, in which case (2.41) has a Pearson type III distribution [101]. For a large estimation window ($K >> 1$) the latter approaches a Gaussian distribution completely specified by the reference signal variance and estimation window length:

$$E[\hat{\sigma}_d^2(n)] = \sigma_y^2(n) + \sigma_\eta^2 \tag{4.12}$$

$$VAR[\hat{\sigma}_d^2(n)] = \frac{2[\sigma_y^2(n) + \sigma_\eta^2]^2}{K-1} \tag{4.13}$$

In practice, the numerator and denominator terms calculated with (2.40) and (2.41) are estimated using the same reference signal $d(n)$ and window length K. As a result, with no near-end speech present, the bias-compensated numerator of (4.9) is expected to be highly correlated with the denominator and also Gaussian. The background noise variance, estimated during quiet periods, is a constant and does not contribute to the numerator variance:

$$E[\underline{\hat{r}}_{xd}^T(n)\underline{\hat{w}}(n) + \sigma_\eta^2] = E[\underline{\hat{r}}_{xd}^T(n)\underline{\hat{w}}(n)] + E[\sigma_\eta^2] = \sigma_y^2(n) + \sigma_\eta^2 \tag{4.14}$$

$$VAR[\underline{\hat{r}}_{xd}^T(n)\underline{\hat{w}}(n) + \sigma_\eta^2] = \frac{2\sigma_y^4(n)}{K-1} \tag{4.15}$$

Since the echo and background noise signals are uncorrelated, the denominator of (4.9) can be written as consisting of two separate Gaussian random processes. This leads to a representation of the numerator and denominator terms of (4.9) as a set of Gaussian random processes represented as follows:

$$\underline{\hat{r}}_{xd}^T(n)\underline{\hat{w}}(n) + \sigma_\eta^2 \sim N\left\{\sigma_y^2(n) + \sigma_\eta^2, \frac{2\sigma_y^4(n)}{K-1}\right\} \tag{4.16}$$

$$\begin{aligned}\hat{\sigma}_d^2(n) &\sim N\left\{\sigma_y^2(n)+\sigma_\eta^2, \frac{2[\sigma_y^2(n)+\sigma_\eta^2]^2}{K-1}\right\} \\ &= N\left\{\sigma_y^2(n)+\sigma_\eta^2, \frac{2\sigma_y^4(n)}{K-1}\right\} + N\left\{0, \frac{2[2\sigma_y^2(n)\sigma_\eta^2+\sigma_\eta^4]}{K-1}\right\}\end{aligned} \tag{4.17}$$

Let $B(n)$ represent the random process of (4.16), and let $A(n)$ represent the second term of (4.17). Substituting (4.16) and (4.17) into (4.9) gives an approximation of the doubletalk detection statistic as a function of $Z(n)$, the ratio of $A(n)$ to $B(n)$ as follows:

$$\xi(n) \approx \sqrt{\frac{B(n)}{B(n)+A(n)}} = \sqrt{\frac{1}{1+A(n)/B(n)}} = \sqrt{\frac{1}{1+Z(n)}} \tag{4.18}$$

Let the means and variances of $A(n)$ and $B(n)$ be represented by μ_A, μ_B, σ_A^2 and σ_B^2, respectively. The PDF of $Z(n)$, $f_Z(z)$, has a generalized Gaussian ratio distribution given by the following equation [101]:

$$\begin{aligned}f_Z(z) = &\frac{e^{-C/2}}{2\pi\sigma_A\sigma_B A(z)} \\ &+ \frac{B(z)}{\sqrt{2\pi}\sigma_A\sigma_B A(z)^{3/2}} \exp\left\{-\frac{1}{2}\left[C-\frac{B^2(z)}{A(z)}\right]\right\} Erf\left[\frac{B(z)}{\sqrt{2A(z)}}\right]\end{aligned} \tag{4.19}$$

where

$$A(z) = \frac{1}{\sigma_A^2}z^2 + \frac{1}{\sigma_B^2} \tag{4.20}$$

$$B(z) = \frac{\mu_A}{\sigma_A^2}z + \frac{\mu_B}{\sigma_B^2} \tag{4.21}$$

$$C = \frac{\mu_A^2}{\sigma_A^2} + \frac{\mu_B^2}{\sigma_B^2} \tag{4.22}$$

and $Erf(\cdot)$ is the error function. Note that for the more specific case of $\mu_A = \mu_B = 0$, (4.19) reduces to the well-known Cauchy distribution. Given that $A(n)$ and $B(n)$ are assumed to

be Gaussian with means μ_A, $\mu_B > 0$, and assuming a large estimation window K, both $A(n)$ and $B(n)$ have nearly vanishing density at zero. In this case, the PDF of $Z(n)$ can be represented by a simplified expression as follows [101]:

$$f_Z(z) = \frac{\sigma_A^2 \mu_B + \sigma_B^2 \mu_A z}{\sqrt{2\pi}(\sigma_A^2 + \sigma_B^2 z^2)^{3/2}} \exp\left\{ -\frac{(\mu_A - \mu_B z)^2}{2(\sigma_A^2 + \sigma_B^2 z^2)} \right\} \tag{4.23}$$

Regardless of the underlying distribution of $Z(n)$, the resulting conditional PDF of the doubletalk detection statistic of (2.42) in the absence of doubletalk can be obtained from (4.18) by solving for $Z(n)$ in (4.18) and taking the derivative of the resulting function with respect to $\xi(n)$:

$$f_{\Xi}[\xi(n) \mid H_1] = \left| \frac{2}{\xi^3(n)} \right| f_Z\left[\frac{1 - \xi^2(n)}{\xi^2(n)} \right] \tag{4.24}$$

for $\xi > 0$. The PDF is characterized by substituting into (4.19) or (4.23) the statistics of $A(n)$ and $B(n)$ given in (4.16) – (4.17):

$$\mu_A = 0 \tag{4.25}$$

$$\sigma_A^2 = \frac{2[2\sigma_y^2(n)\sigma_\eta^2 + \sigma_\eta^4]}{K-1} \tag{4.26}$$

$$\mu_B = \sigma_y^2(n) + \sigma_\eta^2 \tag{4.27}$$

$$\sigma_B^2 = \frac{2\sigma_y^4(n)}{K-1} \tag{4.28}$$

Figure 4.2(a) shows the conditional PDF of the doubletalk detection statistic $\xi(n)$ calculated using (4.24) for two different input signals: white Gaussian noise with unity variance, and a 10^{th}-order autoregressive process driven by white Gaussian noise. For both cases an echo path impulse response of $N = 500$ samples and an estimation window

length of $K = 200$ were employed, with white noise added to the reference signal with an SNR of 30 dB with respect to the echo signal power. Also shown are the corresponding detection thresholds calculated using (4.2) with $P_F \leq 0.1$. For these test signals the respective detection thresholds are $T = 0.9902$ and $T = 0.9606$, which is a considerable difference. Figure 4.2(b) shows the detection threshold calculated using (4.2) as a function of SNR for $P_F \leq 0.1$ and $P_F \leq 0.2$. This figure reveals that the conditional PDF of $\xi(n)$ is dependent upon the statistics of the input signal $x(n)$ and, as a result, the SNR of the reference signal. One consequence is that using (4.2) to select a doubletalk detection threshold based on a desired P_F implies that the threshold must be adaptive to changes in signal characteristics. Figure 4.2(c) shows the detection threshold as a function of SNR for $P_F \leq 0.1$ and estimation window lengths of $K = 50$, 100, and 200 samples. For stationary input signals, increasing K reduces the variance of the detection statistic, which leads to higher detection thresholds for a given P_F.

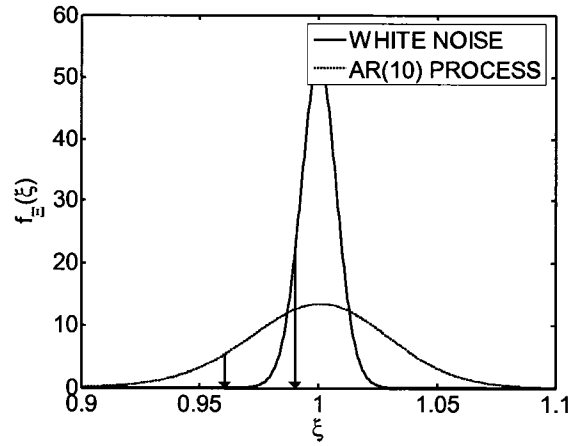


Figure 4.2 - (a) PDF of $\xi(n)$ in the absence of doubletalk for two input signals, along with their corresponding detection thresholds for $P_F \leq 0.1$; (b) detection threshold as a function of SNR for $P_F \leq 0.1$ and $P_F \leq 0.2$, and (c) as a function of estimation window K for $P_F \leq 0.1$.

4.3.3 Probability Distribution in the Presence of Doubletalk

In the presence of doubletalk, the reference signal $d(n)$ contains the echo signal $y(n)$, background noise $\eta(n)$, and near-end speech $v(n)$ as in (2.1). Since it is assumed that the input signal $x(n)$ and near-end speech are uncorrelated, the bias-compensated numerator of (4.9) can again be written in terms of a Gaussian random variable as in (4.16). The constituent signals of the reference signal are also assumed to be uncorrelated. Therefore, in the presence of doubletalk the estimated reference signal variance of (2.41) can again be approximated as a Gaussian random variable for a large estimation window K, with distribution as follows:

$$\underline{\hat{r}}_{xd}^{T}(n)\underline{\hat{w}}(n)+\sigma_{\eta}^{2} \sim N\left\{\sigma_{y}^{2}(n)+\sigma_{\eta}^{2}, \frac{2\sigma_{y}^{4}(n)}{K-1}\right\} \tag{4.29}$$

$$\hat{\sigma}_{d}^{2}(n) \sim N\left\{\sigma_{y}^{2}(n)+\sigma_{\eta}^{2}+\sigma_{v}^{2}(n), \frac{2[\sigma_{y}^{2}(n)+\sigma_{\eta}^{2}+\sigma_{v}^{2}(n)]^{2}}{K-1}\right\} \tag{4.30}$$

As before, (4.30) can be written as two separable contributions to the denominator of (4.9):

$$\begin{aligned}\hat{\sigma}_{d}^{2}(n) &\sim N\left\{\sigma_{y}^{2}(n)+\sigma_{\eta}^{2}+\sigma_{v}^{2}(n), \frac{2[\sigma_{y}^{2}(n)+\sigma_{\eta}^{2}+\sigma_{v}^{2}(n)]^{2}}{K-1}\right\} \\ &= N\left\{\sigma_{y}^{2}(n)+\sigma_{\eta}^{2}, \frac{2\sigma_{y}^{4}(n)}{K-1}\right\} \\ &+ N\left\{\sigma_{v}^{2}(n), \frac{2[2\sigma_{y}^{2}(n)\sigma_{\eta}^{2}+2\sigma_{y}^{2}(n)\sigma_{v}^{2}(n)+2\sigma_{\eta}^{2}(n)\sigma_{v}^{2}(n)+\sigma_{\eta}^{4}+\sigma_{v}^{4}(n)]}{K-1}\right\}\end{aligned} \tag{4.31}$$

Let $B(n)$ again represent the random process of (4.29), and let $A(n)$ represent the second term of (4.31). Therefore, one can again approximate $\xi(n)$ in the presence of doubletalk as a function of $Z(n)$, the ratio of $A(n)$ to $B(n)$ as follows:

$$\xi(n) \approx \sqrt{\frac{B(n)}{B(n)+A(n)}} = \sqrt{\frac{1}{1+A(n)/B(n)}} = \sqrt{\frac{1}{1+Z(n)}} \tag{4.32}$$

Note that (4.32) is identical to the expression derived for $\xi(n)$ in the absence of doubletalk in Section 4.3.2. Therefore, the resulting conditional PDF of the doubletalk detection statistic of (2.42) in the presence of doubletalk is given by the following expression:

$$f_{\Xi}[\xi(n) \mid H_0] = \left|\frac{2}{\xi^3(n)}\right| f_Z\left[\frac{1-\xi^2(n)}{\xi^2(n)}\right] \tag{4.33}$$

for $\xi > 0$, where $f_Z(z)$ is given by (4.19) or (4.23). The PDF is characterized by substituting the contributions of the echo, background noise, and near-end speech variances from (4.29) – (4.30):

$$\mu_A = \sigma_v^2(n) \tag{4.34}$$

$$\sigma_A^2 = \frac{2[2\sigma_y^2(n)\sigma_\eta^2 + 2\sigma_y^2(n)\sigma_v^2(n) + 2\sigma_\eta^2\sigma_v^2(n) + \sigma_\eta^4 + \sigma_v^4(n)]}{K-1} \tag{4.35}$$

$$\mu_B = \sigma_y^2(n) + \sigma_\eta^2 \tag{4.36}$$

$$\sigma_B^2 = \frac{2\sigma_y^4(n)}{K-1} \tag{4.37}$$

Figure 4.3(a) shows the PDF of the doubletalk detection statistic for a 10^{th}-order autoregressive input signal process without doubletalk, and compared to the PDF in the presence of near-end speech with a near-end to echo signal power ratio (NER) of -15 dB, calculated using (4.33) – (4.37). This example employed the same echo path impulse response, SNR = 30 dB, and estimation window of K = 200 samples as Figure 4.2. In this case, using the original detection threshold of T = 0.9606 results in an expected miss

probability obtained with (4.3) of $P_M = 0.0085$. Using the equations above, the doubletalk detector's expected performance can be characterized over a range of conditions. For example, Figure 4.3(b) shows the expected P_M as a function of NER for the same environment, with detection thresholds calculated using (4.2) for $P_F \leq 0.1$ at SNR of 30, 20, and 10 dB. As these figures illustrate, the doubletalk detector's performance is dependent upon the SNR of the reference signal and the power of the near-end speech. Figure 4.3(c) shows the expected P_M for estimation windows of $K = 50$, 100, and 200 samples, revealing that increasing K reduces the expected P_M. In addition to using (4.33) to obtain an expected P_M, it will be shown in Section 4.4 that one can also use (4.33) to construct a detection threshold T for a specified maximum P_M if the NER has a known minimum value.

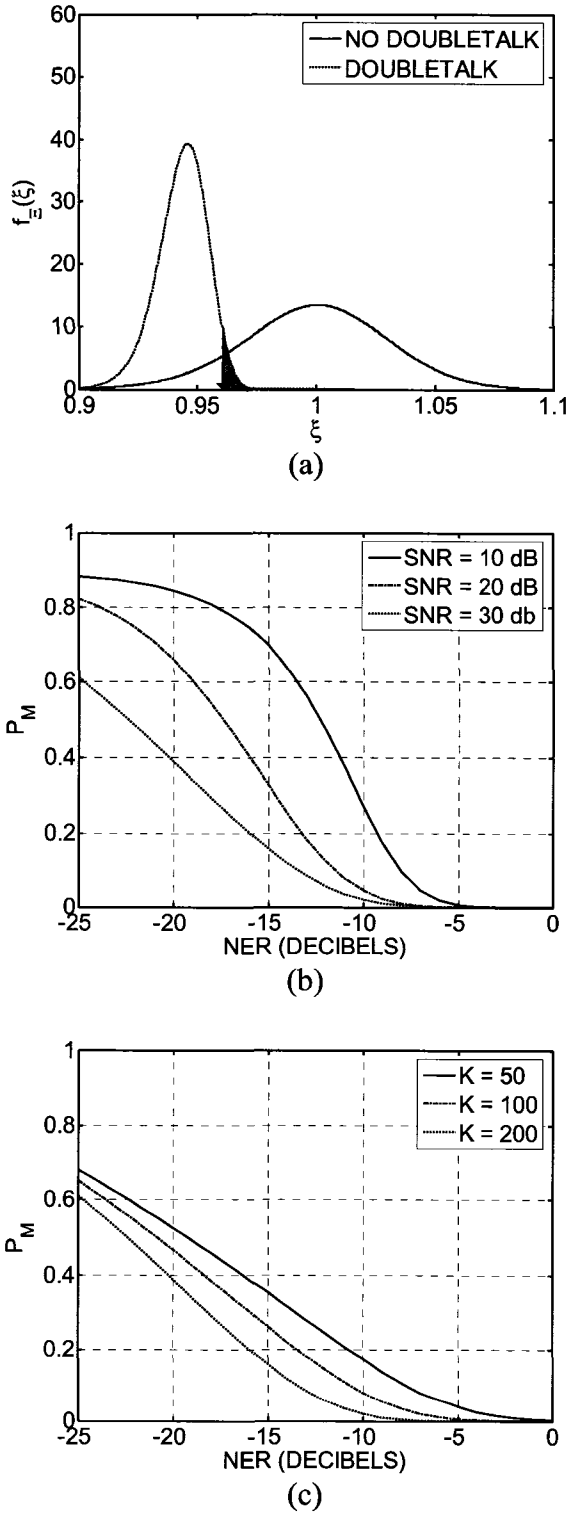


Figure 4.3 - (a) PDF of $\xi(n)$ in the presence of doubletalk, with expected $P_M = 0.0085$ at NER = -15 dB for detection threshold $T = 0.9606$; (b) P_M as a function of NER under various SNR for detection thresholds constructed for $P_F \leq 0.1$, and (c) as a function of estimation window K for SNR = 30 dB.

4.3.4 Expected Doubletalk Detector Response Time

The probability density function derived in Section 4.3.3 assumes that the near-end speech signal is stationary within the estimation window. In practice, the reference signal variance of (2.41) will not reach a steady-state value until a number of samples after the onset of near-end speech. During this time the doubletalk detector may miss doubletalk conditions, thus increasing the overall miss probability. Knowing the probability distribution in the presence of doubletalk, (4.33), allows one to obtain an expected response time for given near-end speech power and detection threshold T. Assume that the near-end speech is stationary as before, but applied "abruptly" to the reference signal as follows. For an estimation window of K samples, let M represent the number of reference signal samples containing both echo and near-end speech, for $1 \le M \le K$. For a near-end speech onset at sample n, the signal used to calculate the reference signal variance estimate of (2.41) can be written in terms of M, the echo signal, background noise, and near-end speech as follows:

$$d(n-k) = \begin{cases} y(n-k)+\eta(n-k)+v(n-k), & 0 \le k \le M-1 \\ y(n-k)+\eta(n-k), & M \le k \le K-1 \end{cases} \tag{4.38}$$

Also assume that the detection statistic values after the onset of near-end speech are independent from each other. In practice this is not the case, although it simplifies the analysis considerably. As in (4.30), the estimated reference signal variance of (2.41) can be approximated as the sum of two terms, the estimated echo signal variance and a term containing the estimated near-end speech variance. However, the contribution of the near-end speech variance will be proportional to the number of reference signal samples containing near-end speech (M out of K).

$$\hat{\sigma}_d^2(n) \sim N\left\{\sigma_y^2(n)+\sigma_\eta^2+\tfrac{M}{K}\sigma_v^2(n), \frac{2[\sigma_y^2(n)+\sigma_\eta^2+\tfrac{M}{K}\sigma_v^2(n)]^2}{K-1}\right\} \tag{4.39}$$

The resulting conditional PDF of $\xi(n)$ in the presence of doubletalk can again be represented using (4.33), where $Z(n) = A(n) / B(n)$ is the Gaussian ratio distribution. In addition, from (4.39), the parameters characterizing $f_Z(z)$ become functions of M and K:

$$\mu_A = \tfrac{M}{K}\sigma_v^2(n) \tag{4.40}$$

$$\sigma_A^2 = \frac{2[2\sigma_y^2(n)\sigma_\eta^2+2\sigma_y^2(n)\tfrac{M}{K}\sigma_v^2(n)+2\sigma_\eta^2\tfrac{M}{K}\sigma_v^2(n)+\sigma_\eta^4+\tfrac{M^2}{K^2}\sigma_v^4(n)]}{K-1} \tag{4.41}$$

$$\mu_B = \sigma_y^2(n)+\sigma_\eta^2 \tag{4.42}$$

$$\sigma_B^2 = \frac{2\sigma_y^4(n)}{K-1} \tag{4.43}$$

Now let $P(M)$, $1 \le M \le K$, be the probability that the doubletalk detection statistic takes M samples to reach the detection threshold T. For $M = 1$, $P(1)$ can be obtained by finding the probability that (4.33) is less than the threshold, or equivalently, evaluating the CDF of (4.33) at the threshold T:

$$P(1) = P[\xi(n) < T \mid H_0, M=1] = \int_0^T f_\Xi[\xi(n) \mid H_0, M=1]d\xi = F_\Xi[T \mid H_0, M=1] \tag{4.44}$$

Assuming still that doubletalk detection statistics are not correlated, $P(M)$ for $M > 1$ can be obtained as a geometric series by multiplying the probability of detection for M samples by the probability that the detection threshold was not reached in any of the previous $M-1$ samples:

$$P(M) = P[\xi(n) < T \mid H_0, M]\{1 - P[\xi(n) < T \mid H_0, M-1]\}\cdots\{1 - P[\xi(n) < T \mid H_0, 1]\}$$
$$= P[\xi(n) < T \mid H_0, M]\prod_{m=1}^{M-1}\{1 - P[\xi(n) < T \mid H_0, m\}$$
$$= F_{\Xi}[T \mid H_0, M]\prod_{m=1}^{M-1}\{1 - F_{\Xi}[T \mid H_0, m]\}$$

(4.45)

Therefore, the expected doubletalk detector response time can be calculated by substituting the variances of the echo, background noise, and near-end speech variances into (4.33). By evaluating this PDF, the expected response time can be obtained simply by finding the expected value of $P(M)$ over all possible delays from $1 \leq M \leq K$:

$$\text{Expected Response Time} = \sum_{m=1}^{K} mP(m) \tag{4.46}$$

Figure 4.4 shows a plot of expected response time as a function of the near-end speech to echo power ratio (NER) for a Gaussian white noise input signal with SNR = -30 dB for K = 50, 100, and 200 samples. It is clear from this plot that the expected response time increases proportionally with the length estimation window K. As a result, there is a tradeoff between the doubletalk detector's response time and the accuracy of parameter estimates of (2.40) – (2.41) obtained by increasing the estimation window length. For low-to-moderate levels of near-end speech (NER in the range of -25 to -10 dB), the expected response time is in the range of 10 – 40 samples.

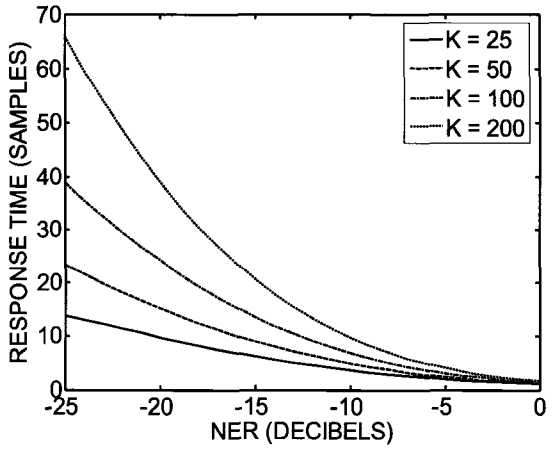


Figure 4.4 - Expected doubletalk detector response time calculated using (4.45) – (4.46) as a function of NER and estimation window size K, for $P_F \leq 0.1$ and SNR = 30 dB.

4.4 Optimal Doubletalk Detector Calibration Algorithms

An important result of Section 4.3 was the characterization of the probability distribution of $\xi(n)$ as functions of statistics of the input, background noise and, in the presence of doubletalk, near-end speech signals. In this section, methods are described for constructing statistically optimal detection thresholds $T(n)$ that are adaptive to changes in signal statistics. Figure 4.5 shows a block diagram of an echo canceller and doubletalk detector employing this proposed additional component for calculating the detection threshold. A parameter estimation block provides estimates of the parameters required to construct the doubletalk detection statistic's PDF in the absence or presence of doubletalk. Finally, either (4.2) or (4.3) are used to construct the threshold based on a desired probability of false alarm P_F or probability of miss P_M.

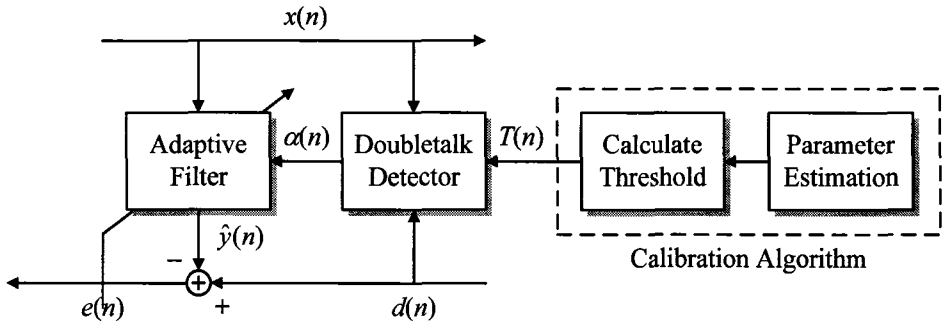


Figure 4.5 - Block diagram of the proposed signal-adaptive doubletalk detection threshold calculation.

4.4.1 Algorithm Description

Figure 4.6 shows a block diagram of the steps required to construct a detection threshold using prescribed P_F or P_M in accordance with (4.2) or (4.3). Each of these approaches requires evaluating the probability density functions of (4.24) or (4.33), respectively, with the variances of the background noise σ_η^2, the echo signal $\sigma_y^2(n)$ and, for (4.33), the maximum near-end speech variance $\sigma_{v,MAX}^2$. The background noise variance, σ_η^2, can be estimated from the reference signal $d(n)$ during quiet periods containing neither echo nor near-end speech. In this algorithm, the per-sample estimates are smoothed with a first-order lowpass filter characterized by a smoothing parameter $0 < \lambda < 1$ applied to individual background signal variance estimates:

$$\hat{\sigma}_d^2(n) = \frac{1}{K-1}\sum_{k=0}^{K-1}\left[d(n-k) - \frac{1}{K}\sum_{j=0}^{K-1} d(n-j)\right]^2 \tag{4.47}$$

$$\hat{\sigma}_\eta^2(n) = \lambda\hat{\sigma}_\eta^2(n-1) + (1-\lambda)\hat{\sigma}_d^2(n) \tag{4.48}$$

It is straightforward to identify periods of no echo by applying a voice activity detector (VAD) to the input signal $x(n)$. However, since the near-end speech is unknown, it is important to only estimate the background noise variance when there is assurance

that no near-end speech is present. One solution to this problem was presented in [90], where periods of no near-end speech were determined by evaluating the correlation between samples in the reference signal when no echo is present.

If it is assumed that the adaptive filter has converged, then the variance of the estimated echo signal, $\hat{y}(n)$, will be approximately equal to that of the true echo. Therefore, the time-varying echo signal variance is obtained by estimating the variance of the estimated echo, and again applying a smoothing factor to the individual estimates:

$$\hat{\sigma}_{\hat{y}}^2(n) = \frac{1}{K-1}\sum_{k=0}^{K-1}\left[\hat{y}(n-k) - \frac{1}{K}\sum_{j=0}^{K-1}\hat{y}(n-j)\right]^2 \tag{4.49}$$

$$\hat{\sigma}_{y}^2(n) = \lambda\hat{\sigma}_{y}^2(n-1) + (1-\lambda)\hat{\sigma}_{\hat{y}}^2(n) \tag{4.50}$$

Finally, once the parameters are estimated, (4.23) or (4.32) can be used to construct the time-varying detection threshold $T_{PF}(n)$ or $T_{PM}(n)$, respectively, by integrating over the PDF until the desired false alarm or miss probability is reached:

$$T_{PF}(n) = T(n) \text{ such that } P[\xi(n) < T(n) \mid H_1] = P_F \rightarrow \int_{\xi(n)=0}^{T(n)} f_{\Xi}[\xi(n) \mid H_1]d\xi = P_F \tag{4.51}$$

$$T_{PM}(n) = T(n) \text{ such that } P[\xi(n) > T(n) \mid H_0] = P_M \rightarrow \int_{\xi(n)=T(n)}^{\infty} f_{\Xi}[\xi(n) \mid H_0]d\xi = P_M \tag{4.52}$$

The most computationally expensive part of the procedure outlined above is performing integration of a PDF once its constituent parameters are estimated. However, note from (4.24) – (4.28) that the PDF is a function of only the echo and noise signal variances. If the SNR is known, based on the fact that the noise variance is assumed known and the echo signal variance is estimated by (4.49) – (4.50), then one can pre-

compute a look-up table providing detection threshold values for a given desirable P_F. An example of this was given in Figure 4.2(b), which shows a plot of $T_{PF}(n)$ as a function of SNR for $P_F \leq 0.1$ and $P_F \leq 0.2$.

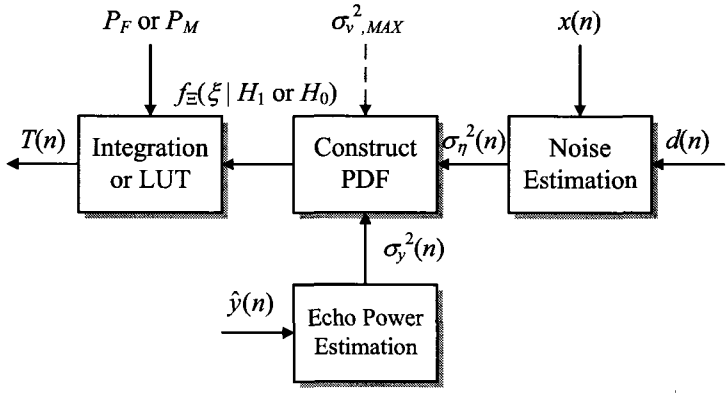


Figure 4.6 - Block diagram of steps for constructing a time-varying doubletalk detection threshold $T(n)$. Background noise variance is estimated from the reference signal during quiet periods, and the time-varying near-end speech variance is estimated from the estimated echo.

4.4.2 Simulation Results

4.4.2.1 Simulation Setup

The echo path impulse response from Figure 3.2 was employed, corresponding to the small conference room described in Chapter 3. Without loss of generality, the impulse response vector was normalized such that $\underline{w}^T\underline{w} = 1$ to equate the echo signal power to the far-end input signal power. Ten input and ten near-end speech sequences were obtained from the TIMIT database and downsampled to a sampling rate of f_s = 8 kHz [92]. The average power of the near-end speech was adjusted relative to that of the input signal to obtain a desired near-end to echo signal power ratio (NER). A simple voice activity detector (VAD) was employed on each of the input and near-end speech signals to identify periods of far-end speech and doubletalk conditions [70]. Background noise was added with a power of -30 dB relative to the average echo signal, and assumed to be of

known variance σ_η^2. The doubletalk detector was implemented using (2.42) with K = 200 samples for the estimation window. It was assumed that the adaptive filter had converged, and variability in the echo path was modeled by applying white noise modulation with variance 0.01. Normalized LMS was used to track the adaptive filter coefficients with step size of $\mu = 0.25$. For constructing adaptive detection thresholds, near-end speech power was estimated using (4.49) and (4.50) with a smoothing factor of $\lambda = 0.98$. A procedure was described in [70] for selecting a fixed detection threshold T_{FIXED} using empirical techniques that was implemented for comparison. In this approach, the doubletalk detector of (2.42) was applied to the training signals with no doubletalk present. The fixed detection threshold was selected by simply evaluating (4.2) for all detection statistic outputs corresponding to active voice periods.

4.4.2.2 Comparison of Adaptive and Fixed Detection Thresholds

The goal of this experiment was to compare the adaptive detection threshold $T_{PF}(n)$ to the fixed threshold T_{FIXED} with respect to the accuracy of calibration. Figure 4.7(a) shows a plot of one of the test far-end speech input signals, and Figure 4.7(b) shows the corresponding fixed and adaptive detection thresholds, both constructed for a false alarm probability of $P_F \leq 0.1$. Figure 4.7(c) shows the corresponding expected probability of miss calculated using (4.33) for NER $\geq$ -15 dB for each of the two thresholds. It is clear that there is a significant difference between T_{FIXED} and $T_{PF}(n)$ over the entire course of the signal, which implies that the actual probability of false alarm using T_{FIXED} will differ from the expected value of 0.1. In addition, a lower detection threshold increases the likelihood that doubletalk conditions will be missed. This effect is evident in Figure

4.7(c), where P_M is higher for the fixed threshold in regions where the fixed threshold is lower than the adaptive one. This implies that whether doubletalk is detected is dependent upon the instantaneous echo signal power during those periods. These effects are due to the fact that the fixed threshold is calculated as an average over all training signals, and is not responsive to the time-varying statistics of an individual input signal.

It is clear from Figure 4.7 that there is an advantage to using adaptive detection thresholds. However, there is a trade-off between the two algorithms: a lower desired miss rate will result in a higher false alarm rate, and vice-versa. To illustrate this effect, Figure 4.8 shows a plot of P_F as a function of the desired maximum P_M, where the latter is calculated for near-end speech signals at NER $\geq$ -25, -20, and -15 dB. Note that for all three of these near-end speech powers, the resulting false alarm rate is greater than 0.5 even for a relatively high P_M of 0.1. This implies that convergence and tracking will occur at least twice as slowly due to the higher P_F, suggesting that using $T_{PM}(n)$ would require, in practice, a relatively high minimum NER.

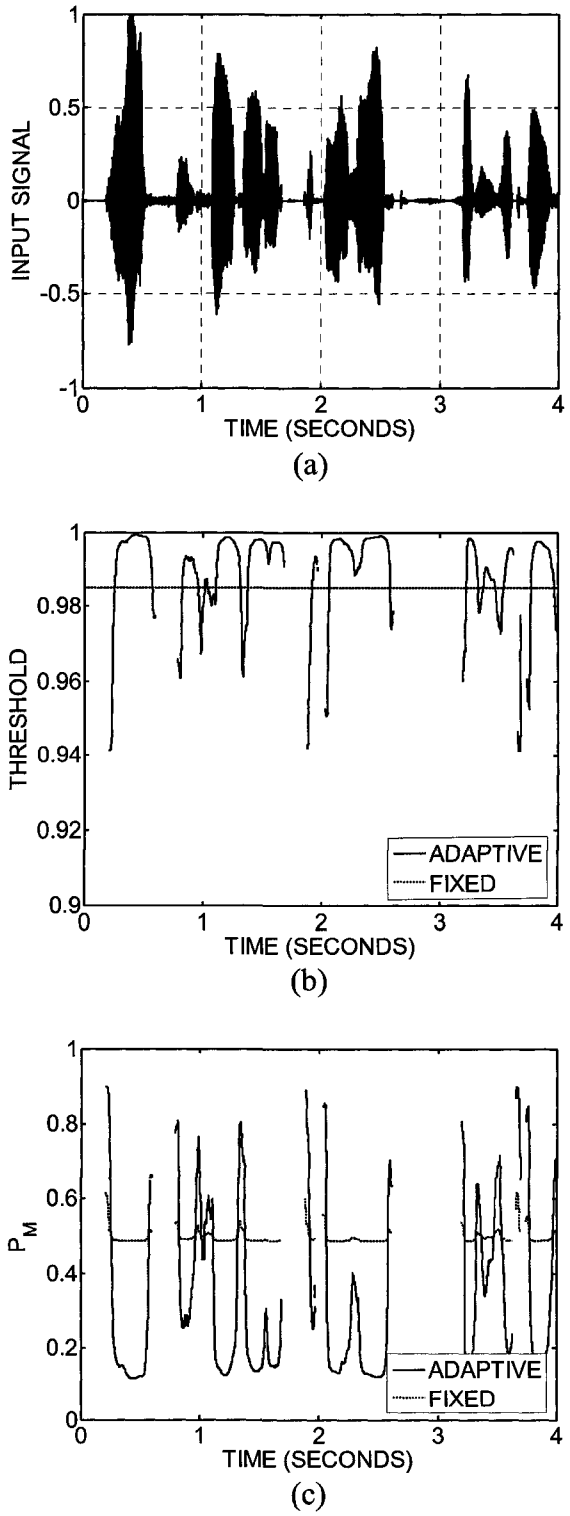


Figure 4.7 – (a) Plot of a test far-end speech input signal; (b) fixed and adaptive detection thresholds constructed for $P_F \leq 0.1$, and (c) corresponding expected P_M for NER $\geq$ -15 dB.

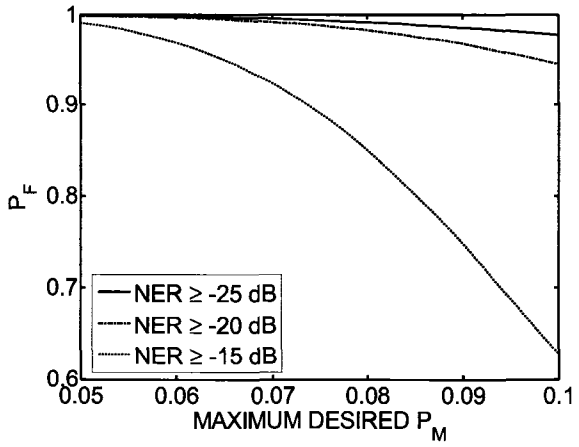


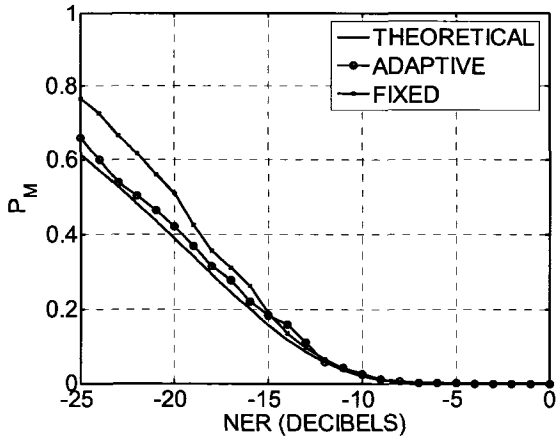
Figure 4.8 - Expected P_F using $T_{PM}(n)$ constructed for a desired maximum P_M and minimum NER.

4.4.2.3 Effect of Adaptive Detection Thresholds on Detection Probability

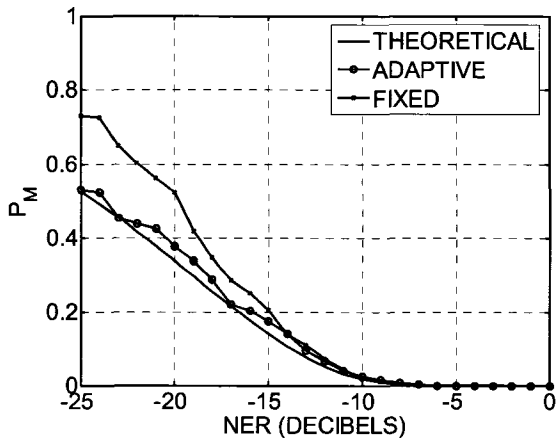
The goal of this experiment was to evaluate the fixed and signal-adaptive detection thresholds with respect to the average probability of miss (P_M) under doubletalk conditions. P_M was obtained by evaluating (4.3) only at samples for which both far- and near-end speech were present over the entire estimation window of K samples. This was done to meet the assumptions of the statistical model in Section 4.3.1 by removing the effects of doubletalk detector response time. The results were averaged over the sets of test input and near-end speech signals. Figure 4.9(a) and Figure 4.9(b) show plots of P_M resulting from using the fixed and adaptive detection thresholds as a function of NER for false alarm probabilities of $P_F \leq 0.1$ and $P_F \leq 0.2$, respectively. Also shown in these figures is the expected P_M for the given NER and adaptive detection threshold, which was calculated using (4.3) and (4.32). It is clear from these figures that using an adaptive detection threshold provides a marked reduction in the miss probability compared to a fixed threshold. This is likely a result of the time-varying difference in P_M shown in Figure 4.7(c). Another observation is that the probability of miss obtained from speech

input and doubletalk signals corresponds closely with the expected value from (4.3) and (4.32).

Recall that Figure 4.7(c) showed that a time-varying detection threshold may produce a lower P_M than a fixed threshold. To illustrate this effect, Figure 4.10(a) shows a test near-end speech signal consisting of two short speech bursts at approximately 1.5 and 2.25 seconds applied to the echo signal of Figure 4.7(a) at NER = -15 dB. Figure 4.10(b) and Figure 4.10(c) show the corresponding ERLE and system distance as functions of time using adaptation controlled with T_{FIXED} and $T_{PF}(n)$. Note that in this case the fixed threshold allows some adaptation and partial divergence during the doubletalk periods, leading to a reduction in ERLE of up to 7.5 dB.



(a)



(b)

Figure 4.9 - Probability of miss (P_M) as a function of near-end to echo signal power ratio (NER) for adaptive and fixed thresholds with (a) $P_F \leq 0.1$ and (b) $P_F \leq 0.2$.

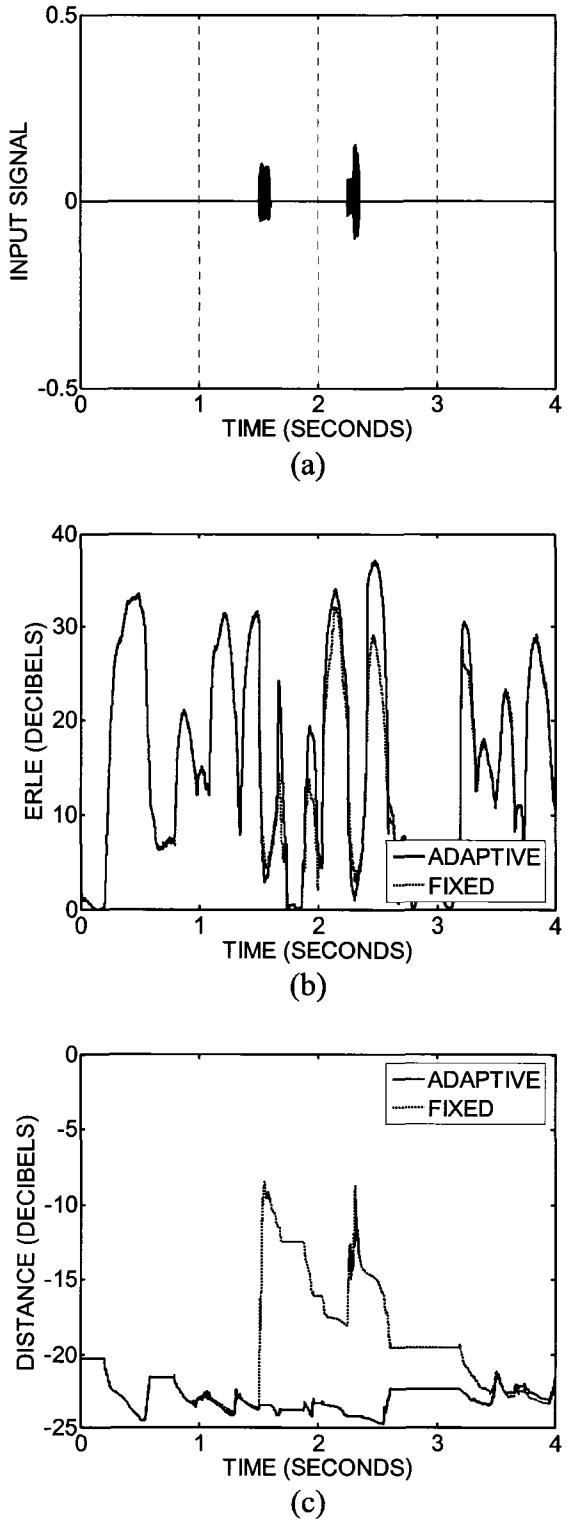


Figure 4.10 – (a) Plot of test near-end speech signal with NER = -15 dB; effect of employing fixed and adaptive detection thresholds on performance in terms of (b) ERLE and (c) system distance.

4.4.2.4 Expected Versus Actual Doubletalk Detector Response Time

The expected response time model derived in Section 4.3.4 was validated experimentally by measuring the actual response time of the doubletalk detector at the onset of near-end speech for 50 pairs of input and doubletalk speech sequences from the TIMIT database. For simplicity, a fixed detection threshold T_{FIXED} was constructed corresponding to $P_F \leq 0.1$. Figure 4.11(a) and 4.11(b) shows the expected and measured doubletalk detector response times as a function of NER for estimation windows of K = 100 and K = 200 samples, respectively, along with the 95% confidence intervals of the measured response times. The plot reveals a fairly close correspondence between the expected and observed response times for NER $\geq$ -15 dB, which supports the response time model constructed in Section 4.3.4.

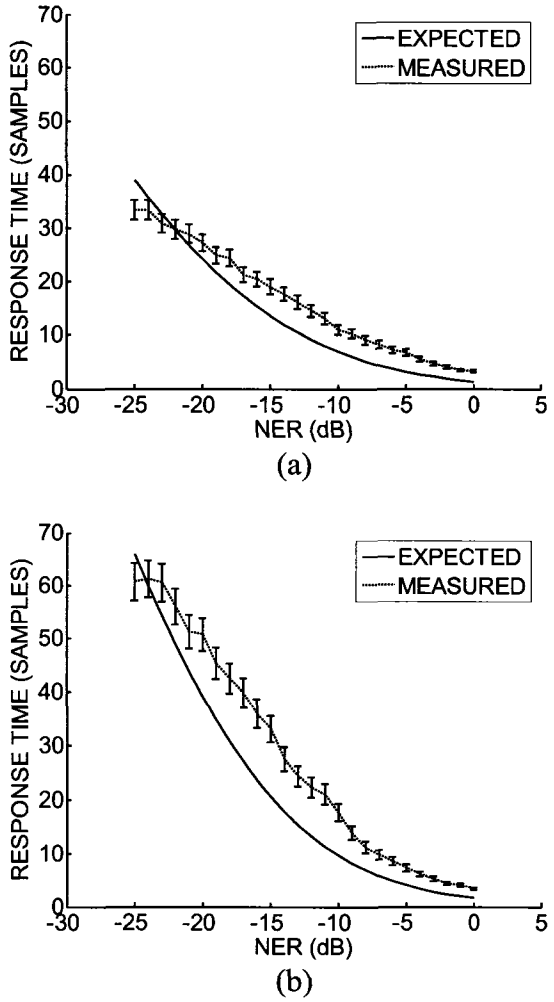


Figure 4.11 - Comparison of expected and actual doubletalk detector response time (95% C.I.) as a function of NER, with SNR = 30 dB and fixed threshold chosen for $P_F \leq 0.1$; (a) $K = 100$ samples; (b) $K = 200$ samples.

4.5 Summary

This chapter investigated the problem of doubletalk detector calibration, and in particular, the problem of selecting a suitable detection threshold in the presence of noise and time-varying signal statistics. Section 4.2 discussed implementation issues that may affect doubletalk detectors in practice, and in particular the problem of bias due to background noise. A method was proposed for selecting a detection threshold based on statistical measures such as desired maximum probabilities of false alarm (P_F) or miss

(P_M). In Section 4.3, the proposed approach was applied to an existing cross-correlation-based doubletalk detector, and models were derived for the detection statistic's PDFs in the absence and presence of doubletalk. A key finding was that the distribution of doubletalk detection statistics is dependent on the statistics of the input and noise signals and, in the presence of doubletalk, on the near-end speech signal. As a result, detection thresholds constructed based on desired maximum P_F or P_M must be adaptive to changes in the input signal and, as a result, SNR. It was also shown that these models can be used to characterize the detector's expected behavior in a variety of environments, and to provide the expected response time at the onset of near-end speech. In Section 4.4, two algorithms were proposed for constructing statistically optimal detection thresholds that are adaptive to changes in the far-end input and noise signal statistics. Use of these detection thresholds was shown to offer increased detection performance compared to a fixed detection threshold obtained using empirical methods.

Chapter 5 Perceptual Performance Limitations of Adaptive Echo Cancellers

5.1 Overview

As reviewed in Chapter 2, objective echo canceller performance measures such as ERLE are calculated using average signal power [12]. Minimum echo canceller performance requirements, such as those in ITU-T G.131 and G.168, are also given in terms of these performance measures [11], [12]. Although these guidelines are based on subjective user tolerance of echo, they assume the presence of low-level background noise at the near end. Psychoacoustic models of human hearing have long been used in audio compression by incorporating masking effects [111]. Some work has been done to determine performance limits of echo cancellers due to nonlinearity in the echo path in the absence of noise ([110]), but little work has been done to identify limits of echo canceller performance due to psychoacoustic effects. This chapter investigates the relationship between psychoacoustic aspects of human hearing and their effect on perceived echo canceller performance. It is shown that the presence of moderate-to-high background noise will have a masking effect on the audibility of residual echo, and that steady-state ERLE does not reflect these masking effects using average power measurements alone. These findings are used develop alternative echo canceller performance measures that incorporate these psychoacoustic masking effects.

The chapter is organized as follows. Section 5.2 discusses psychoacoustic aspects of human hearing in the context of echo cancellation, and demonstrates theoretically the inadequacy of ERLE as performance measure in the presence of noise. Section 5.3

extends this study by presenting listening tests of perceived echo canceller performance in the presence of background noise. Section 5.4 proposes echo canceller performance measures that incorporate the masking effects of background noise. Finally, in Section 5.5 the results and conclusions of this chapter are summarized.

5.2 Perceptual Limitations of Echo Cancellers

This chapter investigates the effects of psychoacoustic aspects of human hearing on perceived echo canceller performance. However, it is difficult to consider all of the possible interactions between far- and near-end signals. At the far-end, the talker hears his or her own voice in the form of sidetone, and it is known that near-end speech tends to mask residual echo returned to the far end [98]. Therefore, the focus is limited to the situation shown in Figure 5.1, where the error signal containing only residual echo $\delta(n)$ and moderate-to-high background noise $\eta(n)$ is returned to the far end with a long round-trip delay. After far-end speech bursts, and after long round-trip delays typical of packet-based telephony (upwards of 200 ms), residual echo is most perceivable at the far end [107]. ITU-T G.131 specifies levels of cancellation required for inaudible residual echo as a function of round-trip delay [11]. For 100 ms one-way delay, the minimum talker echo loudness rating (TELR) is approximately 70 dB, which includes echo return loss, loss pads, and echo cancellers. However, the recommendation does not consider the effects of moderate-to-high background noise.

In this section, the psychoacoustic aspects of sound pressure level, absolute threshold of hearing, and frequency masking are discussed in the context of echo cancellation. An overview is provided of the procedures in the MPEG Psychoacoustic Model for

calculating the masking threshold of an audio signal. Given this procedure for estimating masking effects, a simple echo cancellation experiment is analyzed by applying the masking threshold analysis to the background noise. It is shown through these experiments that existing average power-based echo canceller performance measures do not capture the frequency-masking effects of background noise.

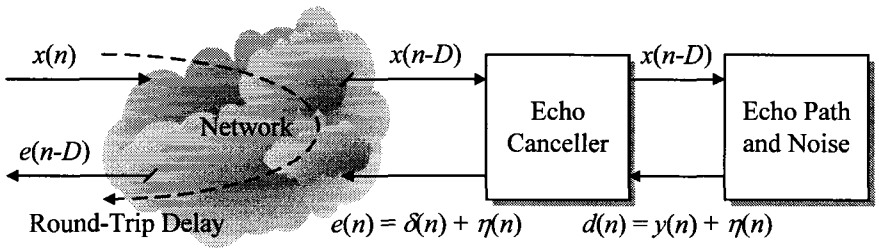


Figure 5.1 - Block diagram of test configuration used to study effects of near-end background noise on echo canceller performance; one-way delay (*D*) is assumed to be at least 100 ms.

5.2.1 Sound Pressure Level and Power Spectrum Estimation

The sound pressure level (SPL) is a standard metric used to represent the intensity of an audio signal. It is defined as the ratio in decibels (dB) of the intensity of sound pressure p relative to a reference intensity p_0, or:

$$\mathrm{SPL} = 20\log_{10}\{p / p_0\} \tag{5.1}$$

where p is the sound pressure in Newtons per square meter (N / m^2) and the reference pressure p_0 is 2×10^{-5} N / m^2 [111]. Most psychoacoustic experiments and results are specified in terms of SPL. However, it is often not possible to know beforehand what the playback levels of signals will be. Therefore, a common assumption is that the playback level will be such that the lowest possible signal power corresponds to SPL of approximately 0 dB.

In this chapter, psychoacoustic concepts are discussed in terms of the power spectra of signals, so it is convenient to represent power spectral density in terms of SPL. To that end, power spectrum estimates for a time-domain signal $x(n)$ are calculated using the Fast Fourier Transform (FFT) and a normalization procedure. In particular, the steps specified by the MPEG psychoacoustic model were employed [106]. First the input signal $x(n)$, $-1 \leq x(n) \leq 1$, is normalized by the FFT block size N_{FFT}:

$$x_{NORM}(n) = \frac{x(n)}{N_{FFT}} \tag{5.2}$$

This normalization ensures that $x_{NORM}(n)$ has a maximum power spectrum of approximately 0 dB SPL. The discrete power spectrum estimate $S_{XX}(k)$ in SPL is then obtained from a windowed version of the normalized input signal:

$$S_{XX}(k) = PN + 20\log_{10}\left|\sum_{n=0}^{N_{FFT}-1} w(n)x_{NORM}(n)e^{-j(2\pi kn/N_{FFT})}\right| \tag{5.3}$$

where $0 \leq k < N_{FFT}/2$, $w(n)$ is a window function, and PN is a fixed term equal to 90.302 dB. The normalization procedure, and in particular the fixed term PN, results in conservative estimates of the maximum playback level.

5.2.2 Equal Loudness and the Absolute Threshold of Hearing

An equal-loudness contour is a frequency-dependent measure of sound pressure which the human auditory system perceives as having constant loudness compared to a reference signal [107], [108]. Contours have been obtained by presenting subjects with a reference sinusoidal signal at 1 kHz with a given power, and by measuring the change in power required for a second sinusoid presented at another frequency to be perceived as

having the same power. These contours represent the sensitivity of human hearing across the range of audible frequencies, and as shown in Figure 5.2(a) for a family of contours, the 1 – 5 kHz range is the most sensitive.

The absolute threshold of hearing is a frequency-dependent function $T_A(f)$ showing the power required by a single sinusoidal signal so that it can be detected by an average listener in a quiet environment. The function can be approximated by the following equation (in dB SPL) [107]:

$$T_A(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 0.001(f/1000)^4 \qquad (5.4)$$

where f is the frequency of the sinusoidal signal in Hz. A plot of $T_A(f)$ is shown in Figure 5.2(b), from which it is clear again that the 1 – 5 kHz band shows the most sensitivity. The function may be written as $T_A(k)$, $0 \le k < N_{FFT} / 2$, for use with discrete power spectrum estimates by substituting $f = kf_s / N_{FFT}$, where f_s is the sampling rate in Hz.

With respect to echo cancellation, $T_A(k)$ can be interpreted as a lower bound on the audibility of residual echo in a quiet environment. That is, $\delta(n)$ is not perceivable if its power spectrum is below the absolute threshold, or $S_{\Delta\Delta}(k) < T_A(k)\ \forall k$. The equal-loudness curves also suggest that the perceived loudness of residual echo is frequency-dependent.

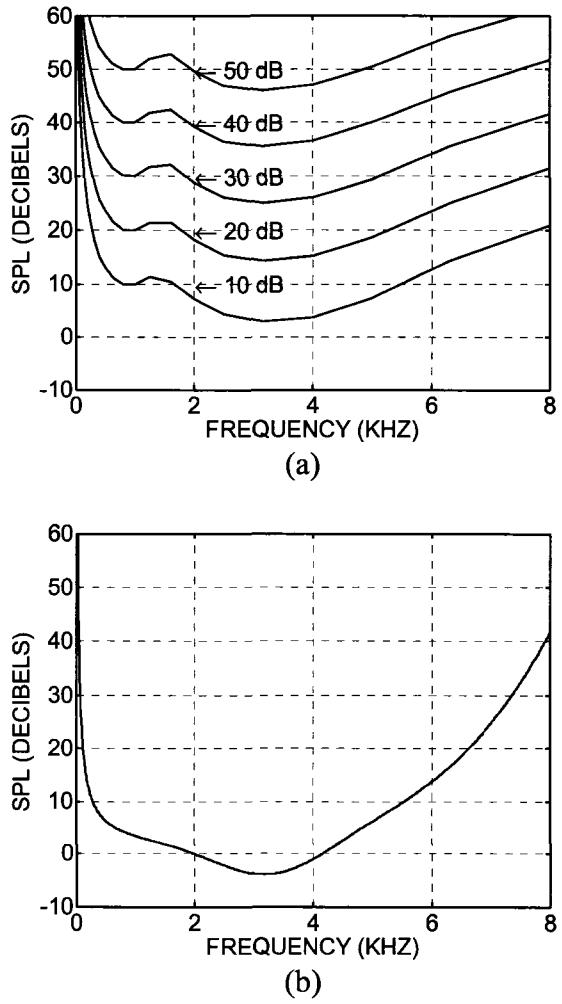


Figure 5.2 – (a) Equal-loudness curves (from [108]) and (b) absolute threshold of hearing across the range of audible frequencies, both expressed as sound pressure level (dB SPL).

5.2.3 Masking Threshold of Background Noise

The presence of background noise will have a masking effect on other signals, most importantly the residual echo signal. Intuitively this makes sense: as the background noise power increases, it will "drown out" residual echo. However, even at lower noise levels, tonal or noisy components in the background noise at frequency index k will limit the audibility of the residual echo around the same frequency. In other words, audible background noise will induce a raised masking threshold $T_M(k)$ below which the residual

echo will not be audible at the far end. The masking threshold is limited by the absolute hearing threshold, or $T_M(k) \geq T_A(k)$. In this study the masking threshold $T_M(k)$, $0 \leq k < N_{FFT} / 2$, was calculated from estimates of the background noise power spectrum $S_{NN}(k)$ using the MPEG-1 Psychoacoustic Model [106]. A thorough description of the steps required to calculate the model can be found in [111] for the original specification with f_s = 44.1 kHz, and they were modified to accommodate lower sampling rates employed in telephony (f_s = 8 – 16 kHz). In the following, the steps required to calculate the masking threshold are briefly presented.

5.2.3.1 Identification of Masking Frequency Components

Masking frequency components are identified from strong tonal and non-tonal components in the power spectrum. A tonal component represents a sinusoid that masks the audibility of other tones and narrowband noise around the same frequency. It is identified as an individual power spectrum component that is at least 7 dB higher than neighboring frequencies. As described in [111], the human auditory system can roughly be described as a non-uniform filter bank of critical bands with passband widths increasing with frequency. A non-tonal component represents a strong noise signal within a critical band that masks the audibility of tones and other noise within the band. If the total power within a critical band of the power spectrum is above the absolute threshold, then a non-tonal component is deemed to exist within that band.

5.2.3.2 Calculation of Raised Masking Thresholds

Each tonal and non-tonal component individually induces a raised masking threshold around its center frequency. A raised masking threshold represents an increase in power,

above the absolute hearing threshold, required by other spectral components in order to be audible in the presence of the masker. In the MPEG-1 psychoacoustic model, the raised masking threshold is represented by a piecewise linear function (on the Bark scale) decaying with the distance from the center frequency of the masker [106]. Let $T_{TONAL}(j, k)$ and $T_{NONTONAL}(j, k)$ represent the raised masking threshold at frequency index k corresponding to a tonal or noise masker centered at frequency j. For background noise with power spectrum $S_{NN}(k)$, the raised masking thresholds are given by the following (in dB SPL) [111]:

$$T_{TONAL}(j,k) = S_{NN}(j) - 0.275B(k) + SF(j,k) - 6.025 \tag{5.5}$$

$$T_{NONTONAL}(j,k) = S_{NN}(j) - 0.175B(k) + SF(j,k) - 2.025 \tag{5.6}$$

where $B(k)$ provides a mapping between frequency index and frequency on the Bark scale, and $SF(j, k)$ is a piecewise linear masking threshold function.

5.2.3.3 Calculation of a Global Masking Threshold

The raised masking thresholds induced by the tonal and non-tonal components are then combined to form a global masking threshold. Assume in the previous steps that N_{TONAL} tonal and $N_{NONTONAL}$ non-tonal components were identified. In the MPEG-1 psychoacoustic model it is assumed that the contributions of raised masking thresholds are cumulative. Therefore, the global masking threshold $T_M(k)$, $0 \leq k < N_{FFT} / 2$, is obtained simply by adding the contributions of individual masking thresholds to the absolute hearing threshold $T_A(k)$ as follows (in dB SPL):

$$T_M(k) = 10\log_{10}\left[10^{T_A(k)/10} + \sum_{i=1}^{N_{TONAL}} 10^{T_{TONAL}(i,k)/10} + \sum_{i=1}^{N_{NONTONAL}} 10^{T_{NONTONAL}(i,k)/10}\right] \tag{5.7}$$

5.2.3.4 Example

As an example of this procedure, background noise was recorded in the small conference room described in Section 3.1, which contained a noisy overhead air conditioning fan. Figure 5.3(a) shows the power spectrum of the background noise compared to the absolute threshold of hearing. Note that the noise has an SPL as high as 40 dB at some frequencies, and that above 4 kHz none of the background noise frequency components were above the absolute threshold. Figure 5.3(b) shows the set of tonal and non-tonal components identified from the power spectrum using the steps outlined above, along with their corresponding raised masking thresholds. Finally, Figure 5.3(c) shows the global masking threshold $T_M(k)$ compared to the original noise power spectrum. An interesting observation is that the masking threshold in mid-band frequencies (2 – 4 kHz) is up to 10 dB higher than the noise power spectrum, which implies that residual echo *above* the noise power spectrum may still not be audible.

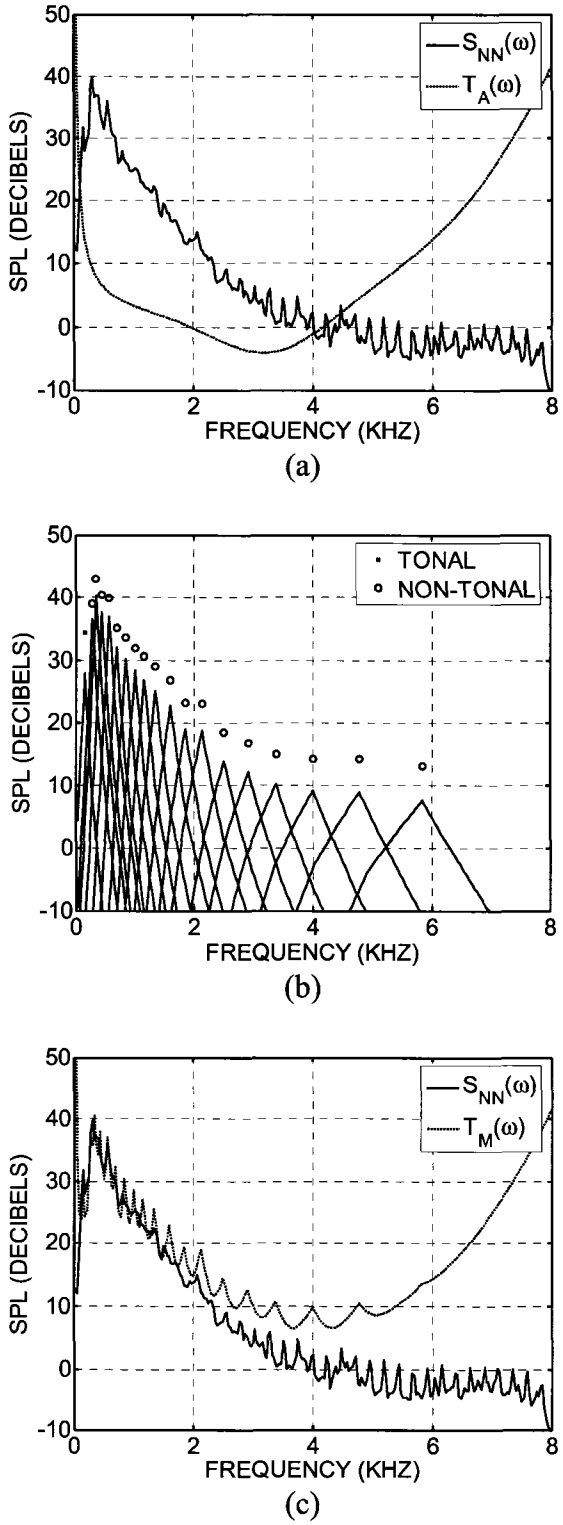


Figure 5.3 - (a) Noise power spectrum $S_{NN}(\omega)$ compared to the absolute hearing threshold $T_A(\omega)$; (b) tonal and non-tonal components and their raised thresholds; (c) background noise masking threshold $T_M(\omega)$.

5.2.4 Limitations of Echo Canceller Performance Measures in Noise

ERLE is usually calculated by averaging reference and error signal power over a window of samples at time n as given by (2.4). This assumes that the background noise $\eta(n)$ is low enough to have a negligible contribution, which may not be a valid assumption in environments with moderate-to-high levels of noise. Therefore, a more accurate measure of echo power reduction, denoted $\text{ERLE}_{\text{ACTUAL}}$, is obtained from the echo and residual echo signals themselves:

$$\text{ERLE}_{\text{ACTUAL}}(n) = 10\log_{10}\left[\frac{\sigma_y^2(n)}{\sigma_\delta^2(n)}\right] \approx 10\log_{10}\left[\frac{\sum_{k=0}^{K-1} y^2(n-k)}{\sum_{k=0}^{K-1} \delta^2(n-k)}\right] \tag{5.8}$$

where $\sigma_y^2(n)$ and $\sigma_\delta^2(n)$ are the echo and residual echo signal powers at time n, respectively, and $E[\cdot]$ is the statistical expectation operator.

This discussion is presented in the context of echo canceller performance measures obtained from an extremely simple simulation environment. An impulse response $w(n)$ was obtained from the conference room described in Section 3.1, but at a sampling rate of f_s = 16 kHz, and truncated to L = 2500 samples. An input signal $x(n)$ consisting of white Gaussian noise was convolved with the impulse response to create the echo signal $y(n)$, with the input signal calibrated to have a power of approximately 60 dB SPL. Additive background noise $\eta(n)$ from the conference room was then added to create the reference signal $d(n)$. An adaptive filter employing NLMS was used to represent a typical echo canceller operating in such an environment, with an adaptive filter of N = 2500 samples

and a step size of $\mu = 0.05$ [44]. ERLE was calculated as a function of time using (2.4) and, since the background noise signal was known *a priori*, also calculated using (5.8).

Figure 5.4(a) shows ERLE calculated using (2.4) and (5.8), and indicated are the approximate steady-state times at which additional analysis was performed. It is clear that the presence of background noise limits ERLE calculated using (2.4), which reaches a steady-state value of 35 dB at approximately $n_1 = 27500$ samples. However, when ERLE is calculated without the background noise using (5.8), a steady-state value of 50 dB is obtained at approximately $n_2 = 40000$ samples. Figure 5.4(b) shows the estimated power spectrum of the residual echo signal, $S_{\Delta\Delta}(k)$, at times n_1 and n_2 compared with the masking threshold of the background noise. From this plot it is clear that at time n_1 the residual echo is not completely below the masking threshold. As a result, frequency components between 3 – 6 kHz will still be audible at the far end. In contrast, at time n_2 the residual echo is far below the masking threshold, implying that it is completely inaudible at the far end.

Several observations can be made from these results. It is clear that for ERLE calculated with (2.4), the steady-state value may not always determine whether the residual echo signal inaudible to the far-end talker in the presence of noise. This results from the fact that average power simply does not capture frequency-dependent psychoacoustic aspects. As a result, it may be possible to have two different sets of echo canceller coefficients capable of producing the same ERLE for a given environment, but the residual echo signal from one may be *more perceivable* than the other. Another observation is that the masking threshold of background noise is frequency-dependent.

For a given environment some frequencies are perceptually more "important" than others in that the residual echo signal may have components above the masking threshold. Finally, these results also indicate that it may be possible to construct alternative performance measures based on knowledge of the background noise masking threshold, an idea proposed in Section 5.4.

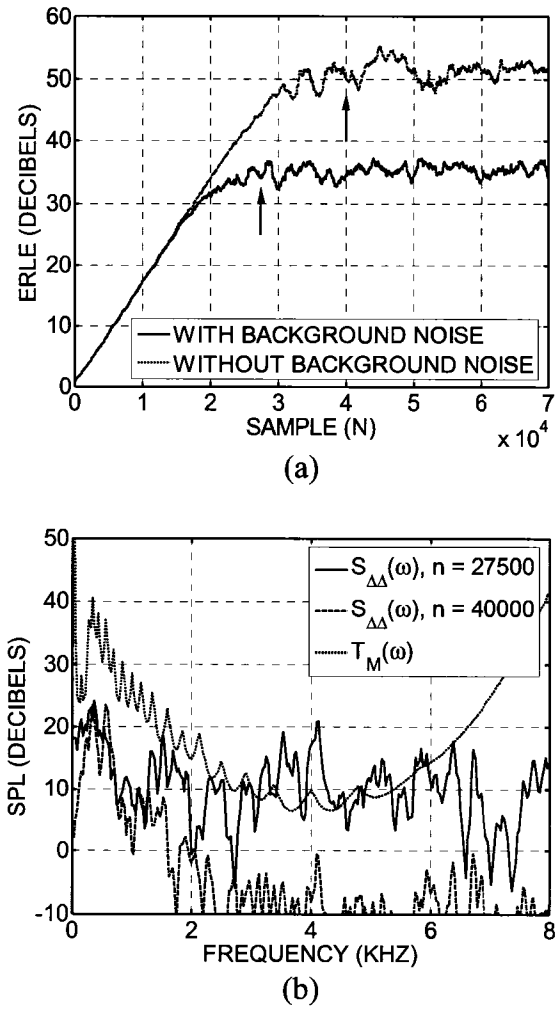


Figure 5.4 - (a) ERLE during initial convergence, calculated with and without background noise; (b) residual echo power spectrum $S_{\Delta\Delta}(\omega)$ at 27500 and 40000 samples, compared to the noise masking threshold $T_M(\omega)$.

5.3 Verification of Psychoacoustic Effects

In the previous section, the masking effects of background noise were investigated by applying an existing psychoacoustic model to the background noise signal, which was then used to determine the audibility of the residual echo signal. In this section the investigation is extended through the use of experimental results and informal listening tests. First a simple modification to NLMS is presented that minimizes a weighted error criterion instead of the usual mean squared error. A description is provided of the fixed weighting function, derived from the absolute threshold of hearing, followed by the results of simulation and simple listening tests.

5.3.1 A Perceptually Weighted Adaptation Algorithm

In the previous section it was noted that ideally one would like an adaptation algorithm that emphasizes frequencies where the residual echo signal is higher than the masking threshold of background noise. Previously NLMS was employed because of its simplicity and widespread use. However, numerous techniques exist for achieving a faster rate of convergence for adaptive filters, such as Recursive Least Squares (RLS) and Affine Projection (AP) [7]. Most of these algorithms update filter coefficients to minimize the mean square error or some other global error function. There have been attempts at tailoring subband echo canceller performance by assigning taps to subband adaptive filters based on objective or psychoacoustic properties [112], [113]. However, it is not obvious how direct knowledge of the background noise masking threshold can be directly incorporated into an echo canceller's adaptation algorithm.

The equal-loudness contour plots of Section 5.2.2 revealed that human hearing is more sensitive to frequencies between 1 – 5 kHz. In this section a simple modified version of NLMS is presented that incorporates a perceptual pre-emphasis filter based on these plots. The goal is to show through listening tests (in Section 5.3.3) that by increasing the rate of convergence in sensitive frequency bands relative to others, the "perceived" rate of convergence of an adaptive echo canceller can be increased.

A block diagram of the perceptually weighted adaptation algorithm is shown in Figure 5.5. First let $\underline{f}(n) = [f_0(n) f_1(n) \ldots f_{F-1}(n)]^T$ be the coefficient vector for an FIR filter of length F samples. The filter coefficients are written as a function of time n to indicate that the coefficients may be either fixed or slowly time-varying. Let $x_f(n)$ and $e_f(n)$ be versions of the input and error signals that have been filtered with $\underline{f}(n)$. In particular:

$$x_f(n) = \underline{f}^T(n)\underline{x}(n) \tag{5.9}$$

$$e_f(n) = \underline{f}^T(n)\underline{e}(n) \tag{5.10}$$

where $\underline{x}(n) = [x(n)\ x(n-1)\ \ldots\ x(n-F+1)]^T$ and $\underline{e}(n) = [e(n)\ e(n-1)\ \ldots\ e(n-F-1)]^T$. The NLMS coefficient update using the filtered input and error signals is then given by the following equation:

$$\underline{\hat{w}}(n+1) = \underline{\hat{w}}(n) + \mu \frac{e_f(n)\underline{x}_f(n)}{\underline{x}_f^T(n)\underline{x}_f(n) + \delta} \tag{5.11}$$

where μ is the adaptation step size, and δ is a regularization parameter. Using the filtered input and error signals of (5.9) and (5.10) in the adaptation algorithm is equivalent to minimizing a weighted cost function $J(n)$ based on the filtered error signal:

$$J(n) = E[e_f^2(n)] \tag{5.12}$$

As a result, the average error signal power is minimized as a frequency-weighted version of the original error signal:

$$\sigma_{e_f}^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{EE}(\omega)\,|F(\omega)|^2\,d\omega \tag{5.13}$$

where $S_{EE}(\omega)$ is the error signal power spectrum, and $F(\omega)$ is the magnitude response of $f(n)$. It is important to note that the proposed modification is somewhat similar to the Filtered-X NLMS algorithm employed in active noise control systems [55].

Analyses of NLMS have shown that the rate of convergence in terms of mean square error is limited by small eigenvalues of the input signal's autocorrelation matrix [54]. The choice of $f(n)$ has a significant impact on the power spectrum of $x_f(n)$, and hence on the eigenvalues of the filtered signal's autocorrelation matrix. Several authors have shown that constructing $f(n)$ to adaptively decorrelate (or whiten) the input signal can increase the overall rate of convergence [114]. However, note that the error signal $e(n)$ is also filtered and may contain background noise, and such algorithms operate under the constraint that the background noise signal is filtered and may be enhanced.

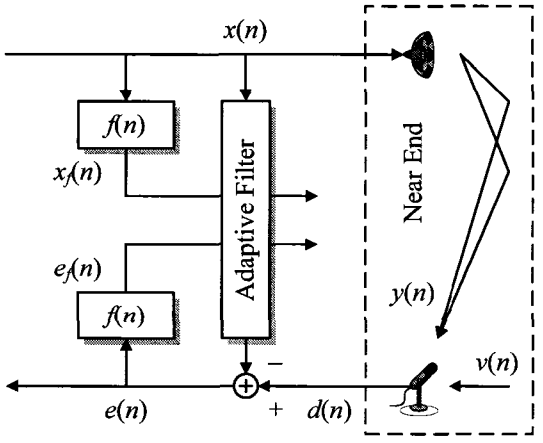


Figure 5.5 - Block diagram of NLMS employing a fixed pre-emphasis filter $f(n)$.

5.3.2 Pre-emphasis Filter Design

In this study a fixed filter $f(n)$ was designed with a magnitude response roughly approximating the inverse of the absolute threshold of hearing $T_A(f)$ of (5.4). As described in Section 5.2.2, a 256-sample discrete-frequency version of the absolute threshold function, $T_A(k)$, was first obtained using $f = kf_s / N_{FFT}$ with $N_{FFT} = 512$ samples for $0 \leq k < N_{FFT} / 2$. A target magnitude response function was obtained by first inverting the absolute threshold function, converting it to linear SPL from decibels, and then normalizing it by the maximum of the magnitude response:

$$M(k) = \frac{10^{-T_A(k)/20}}{\max\{10^{-T_A(k)/20}\}} \tag{5.14}$$

Finally, the MATLAB function *firls* was used to generate a linear phase FIR filter of length $F = 16$ samples using the samples of the target magnitude response function and a least-squares technique. A relatively short filter was used because an exact representation of the inverse absolute threshold function was not required. A plot of the filter's magnitude response is shown in Figure 5.6, and it can be seen that frequencies outside the band between 2.5 – 6 kHz are attenuated relative to frequencies within the band. In [54] it was shown that for long impulse responses ($L >> 1$), the eigenvalues of $x(n)$ and, by implication, $x_f(n)$, can be approximated by the DFT power spectrum of the signal's autocorrelation function. Therefore, choosing a pre-emphasis filter $f(n)$ in such a manner has the effect of increasing the eigenvalue spread between the 2.5 – 6 kHz band relative to the others, which should degrade the rate of convergence of the adaptive filter outside of that band.

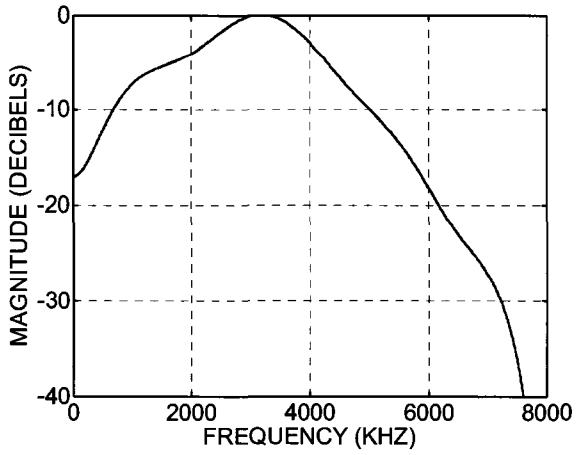


Figure 5.6 - Magnitude response of the fixed pre-emphasis filter $f(n)$ used in this study.

5.3.3 Simulation and Listening Test Results

In this section experimental results are presented for NLMS employing the perceptual pre-emphasis filter described in Section 5.3.2. The goals of these experiments are twofold. First of all, it is desired to compare the performance of standard NLMS with pre-emphasized NLMS with respect to the standard echo canceller performance measures of ERLE and system distance. The results from Section 5.2 showed that it is not possible to rely on these standard measures to determine the perceivability of residual echo signals. Therefore, listening tests are presented comparing the perceivability of residual echo signals generated by the two adaptive filter configurations during their initial convergence periods.

5.3.3.1 Experimental Setup

In these experiments a setup similar to that used in Section 5.2.4 was employed, using white noise input with a power of approximately 60 dB SPL. The same conference room impulse response and background noise obtained for the previous experiment were also used. Two configurations of adaptive filters were used, standard NLMS and NLMS with

the pre-emphasis filter described in Section 5.3.2, denoted "pre-emphasized NLMS". Both of the adaptive filters contained $N = 2500$ samples to match the length of the room impulse response, and a sampling rate of $f_s = 16$ kHz and step size of $\mu = 0.05$ were used for both algorithms. ERLE and system distance were calculated during initial convergence using (2.4) and (2.13), respectively.

For the listening tests, the filter coefficients for the two adaptive filter structures were extracted at five uniformly spaced times n_i, $1 \leq i \leq 5$, during the initial convergence period. The corresponding filter coefficient error vectors for NLMS and pre-emphasized NLMS at time n_i are denoted $\Delta\underline{w}_{NLMS}(n_i)$ and $\Delta\underline{w}_{PNLMS}(n_i)$, respectively. Four subjects were recruited, two male and two female, varying in age between 20 and 35 years. 50 sets of speech input signals from both male and female speakers were randomly selected from the TIMIT continuous speech database [92]. For each of the five times n_i during convergence, listening tests were conducted with a different subset of 10 of the 50 original speech input signals. For each of the two adaptive filter structures, a test error signal was constructed by convolving one of the 10 speech input signals with the corresponding adaptive filter error vector, and again background noise $\eta(n)$ from the conference room was added. In all cases, the speech input signals were adjusted to an average power of approximately 60 dB SPL in accordance with (5.3). In particular:

$$e_{NLMS}(n) = \Delta\underline{w}_{NLMS}^{T}(n_i)\underline{x}(n) + \eta(n) \tag{5.15}$$

$$e_{PNLMS}(n) = \Delta\underline{w}_{PNLMS}^{T}(n_i)\underline{x}(n) + \eta(n) \tag{5.16}$$

The first listening test employed a two-alternative forced choice test [107]. Subjects were fitted with a set of Jensen JF-25 headphones and allowed to listen to each of the two

corresponding error signals, repeatedly if required, and asked to decide which of the two residual echo signals overall was more perceivable. This process was repeated for each of the 10 test input signals in the subset, and in turn for each of the five sets of adaptive filter coefficients at times n_i. In such a test environment, if subjects selected the error signal from each algorithm evenly, then it would suggest that no perceivable difference exists between the echo cancellers at that time. To help reduce bias, pairs of test error signals from each of the five different sets were presented in a random order for each of the subjects.

For the second experiment the subjects were presented with the same sets of error signals as before, but were asked to adjust the power of one of the *input* signals until they were satisfied that the error signals had the same overall level of loudness. A simple user interface was used to automate the process and iterate over the set of test input signals, and is shown in Figure 5.7. Subjects were allowed to adjust the power of one of the input signals and listen to the resulting error signals until satisfied with the result. Adjustments were allowed in increments of 1 dB SPL. Speech input signals were used for the listening tests, and so there are time variations in signal strength and spectral characteristics even with a fixed average power. However, subjects were asked for decisions based on overall perceivability and loudness.

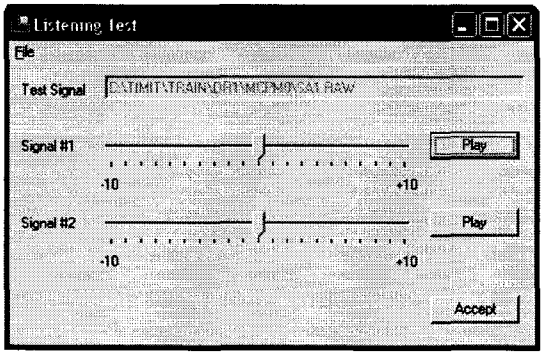


Figure 5.7 - Graphical user interface used in the second listening test. Subjects may adjust the average power of the inputs and listen to the results, accepting the settings to move onto the next test signal.

5.3.3.2 Effect of Pre-Emphasis on Echo Canceller Performance

Figure 5.8(a) shows ERLE for NLMS and pre-emphasized NLMS (PNLMS) obtained during initial convergence. In addition, Figure 5.8(b) shows a plot of the system distance calculated over the same time period. It is clear from these figures that according to the standard performance measures, the pre-emphasized NLMS algorithm has a severely degraded rate of convergence for the same input signal, background noise, and step size parameter. However, Figure 5.9(a) and Figure 5.9(b) show the magnitude responses of the adaptive filter error vectors at time indices $n = 15000$ and $n = 30000$, respectively. It is evident that pre-emphasized NLMS achieves as much as 20 dB lower misadjustment in the 1.5 – 4 kHz frequency band early on, at the expense of higher misadjustment in lower and higher bands. Although this experiment was performed with a white noise excitation, the results suggest that during convergence, the magnitude of residual echo in the 1.5 – 4 kHz band will be less using pre-emphasized NLMS than regular NLMS.

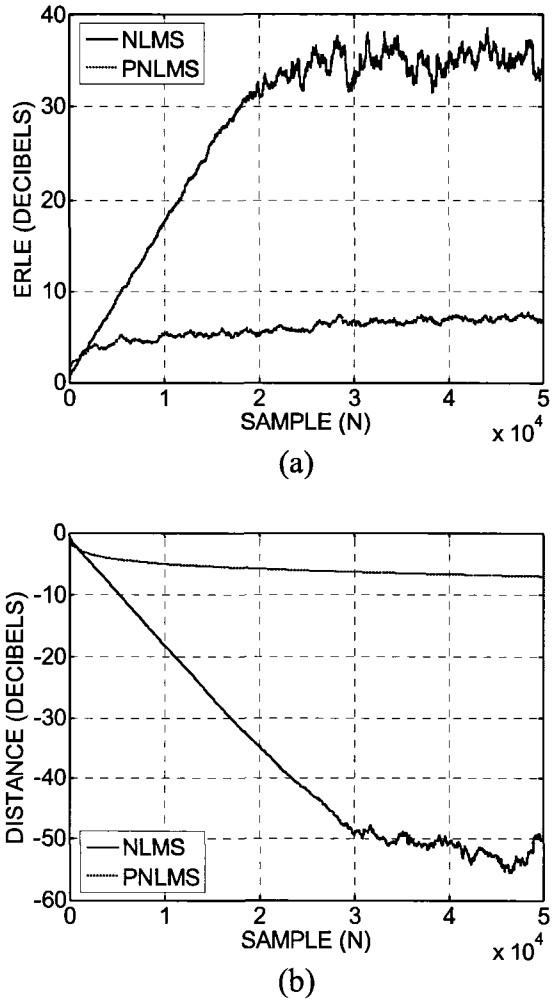


Figure 5.8 – Comparison of NLMS and pre-emphasized NLMS (PNLMS) during initial convergence period: (a) ERLE and (b) system distance.

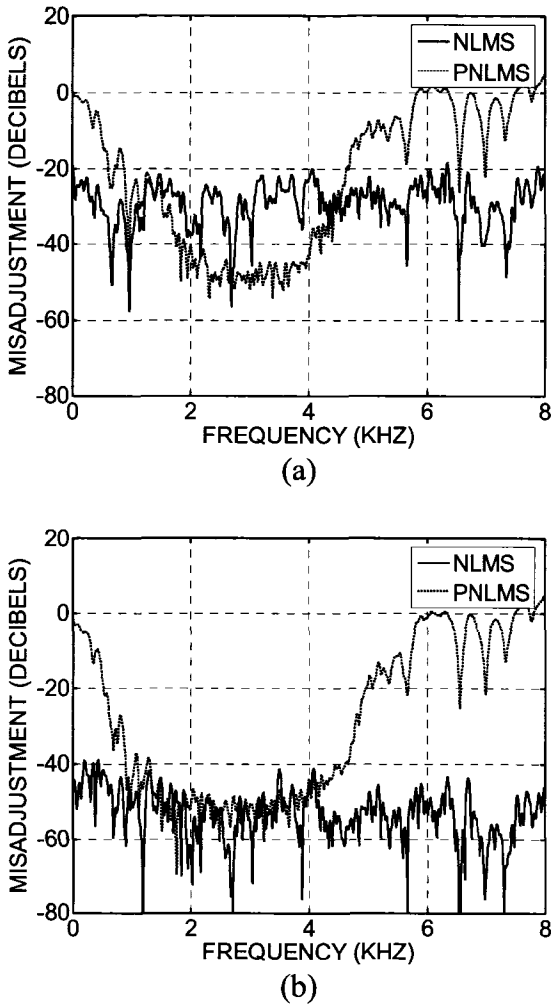


Figure 5.9 – Comparison of misadjustment during initial convergence period: magnitude response of adaptive filter error vectors at (a) n = 15000 and (b) n = 30000 samples.

5.3.3.3 Listening Test Results

Echo canceller error vectors were obtained from the simulations of Section 5.3.3.2 at the following times: n_i = {5000, 10000, 15000, 20000, 25000}. Table 5.1 shows the results of the first listening test for each of the times n_i and for each of the test subjects. Shown are the number of times (out of ten) a subject selected the error signal from NLMS as the more perceivable. It is clear that at time periods n_2 – n_4 the error signals corresponding to NLMS were judged the more perceivable of the two. Figure 5.10 shows

the results of the second listening test showing as a function of time n_i the average number of dB SPL that the input signal corresponding to the pre-emphasized NLMS must be increased so that the error signals have the same overall level of loudness. Again it is clear that at time periods $n_2 - n_4$ the test subjects found that input signal power must be increased, with a maximum increase of 5.5 dB SPL at n_3. These two sets of results are significant because in the previous section it was seen that the pre-emphasized NLMS has a *degraded* convergence overall, and one would expect that the corresponding residual echo signal would be more perceivable. Therefore, additional evidence exists that ERLE and other performance measures do not reflect the perceived performance of an echo canceller, and one cannot rely on them to differentiate between the perceivability of residual echo signals.

Table 5.1 - Listening test results showing the number of times a subject selected the error signal generated by NLMS as the more perceivable of the two test signals (out of ten).

Subject	Time Index				
Number	n_1	n_2	n_3	n_4	n_5
1	5	6	6	5	4
2	4	7	8	6	5
3	6	6	7	6	5
4	5	8	8	7	6

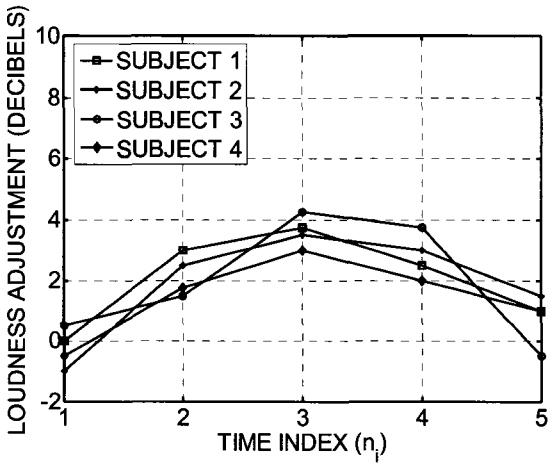


Figure 5.10 - Listening test results showing how much the input signal power for pre-emphasized NLMS must be increased to make the residual echo have the same overall loudness as that produced by NLMS.

5.4 Perceptual Performance Measures for Echo Cancellers

Sections 5.2 and 5.3 investigated the effects of background noise on the audibility of residual echo, and it was shown experimentally that standard echo canceller performance measures are not accurate in the presence of frequency-dependent masking effects of background noise. In this section, performance measures are proposed that incorporate these masking effects, along with procedures for calculating them in practice. The utility of these performance measures is demonstrated through simulation and informal listening tests.

5.4.1 Audible Echo Return Loss Enhancement

As described in Section 5.2.4, a more accurate measure of echo canceller performance is given by the ratio of echo to residual echo signal powers in (5.8), since the latter is not skewed by the presence of background noise. However, note that power is simply calculated as an average of the power spectrum over all frequencies. In terms

of (5.8), the formula for steady-state ERLE can be expanded in terms of the power spectra of the echo and residual echo signals:

$$\mathrm{ERLE}_{\mathrm{ACTUAL}} = 10\log_{10}\left[\frac{\sigma_y^2}{\sigma_\delta^2}\right] = 10\log_{10}\left[\frac{\frac{1}{2\pi}\int_{\omega=0}^{2\pi} S_{YY}(\omega)d\omega}{\frac{1}{2\pi}\int_{\omega=0}^{2\pi} S_{\Delta\Delta}(\omega)d\omega}\right] \tag{5.17}$$

where $S_{YY}(\omega)$ and $S_{\Delta\Delta}(\omega)$ denote the power spectra of the echo and residual echo signals, respectively.

In previous sections it was shown that in the presence of background noise, the audibility of echo and residual echo depends on whether their power spectra contain frequency components above the noise masking threshold, $T_M(\omega)$. Even in quiet environments, audibility of the signals is ultimately limited by the absolute threshold of hearing, $T_A(\omega)$. Therefore, it makes sense to incorporate knowledge of these hearing thresholds into echo canceller performance measures. Two such methods are proposed as follows.

5.4.1.1 Audible Echo Return Loss Enhancement ($\mathrm{ERLE_A}$)

One method of incorporating masking thresholds is to calculate echo and residual echo signal power by considering only frequencies containing audible energy, or in other words, by averaging their respective power spectra only over frequencies containing components above the masking threshold. Let ω_y and ω_δ represent the piecewise-continuous sets of frequencies over which the echo and residual echo power spectrum, respectively, are above the masking threshold:

$$\omega_y = \{\omega \mid S_{YY}(\omega) > T_M(\omega)\} \tag{5.18}$$

$$\omega_\delta = \{\omega \mid S_{\Delta\Delta}(\omega) > T_M(\omega)\} \tag{5.19}$$

Let $|\omega_y|$ and $|\omega_\delta|$ be the sums of the ranges of frequencies covered by ω_y and ω_δ, respectively. The echo and residual echo power, averaged over those frequencies, represent the audible signal power before and after cancellation, respectively:

$$\sigma^2_{y,audible} = \frac{1}{|\omega_y|}\int_{\omega_y} S_{YY}(\omega)d\omega \tag{5.20}$$

$$\sigma^2_{\delta,audible} = \frac{1}{|\omega_\delta|}\int_{\omega_\delta} S_{\Delta\Delta}(\omega)d\omega \tag{5.21}$$

Substituting (5.20) and (5.21) into (5.17) results in a performance measure reflecting the reduction in audible echo signal power, denoted audible ERLE ($ERLE_A$):

$$\mathrm{ERLE_A} = 10\log_{10}\left[\frac{\sigma^2_{y,audible}}{\sigma^2_{\delta,audible}}\right] = 10\log_{10}\left[\frac{\frac{1}{|\omega_y|}\int_{\omega_y} S_{YY}(\omega)d\omega}{\frac{1}{|\omega_\delta|}\int_{\omega_\delta} S_{\Delta\Delta}(\omega)d\omega}\right] \tag{5.22}$$

One disadvantage of (5.22) occurs if the residual echo power spectrum falls completely below the background noise masking threshold. This would imply that no audible residual echo remains, producing an infinite value of $ERLE_A$. To avoid this computational problem, the maximum $ERLE_A$ is assumed to be produced by residual echo of constant power equal to the minimum of the background noise masking threshold:

$$\mathrm{ERLE}_{\mathrm{A,MAX}} = 10\log_{10}\left[\frac{\frac{1}{|\omega_y|}\int_{\omega_y} S_{YY}(\omega)d\omega}{\min\{T_M(\omega_y)\}}\right] \tag{5.23}$$

5.4.1.2 Proportional Audible Echo Return Loss Enhancement

Both ERLE and $\mathrm{ERLE_A}$ in (5.8) and (5.22) provide broad measures of echo power reduction, which are useful for network planning and implementing echo loss plans in accordance with ITU-T G.131 [11]. An alternative formulation of echo return loss enhancement is described as follows. First define a frequency-dependent function $D(\omega)$ as the ratio between the power spectral density of the echo and residual echo signals. For a non-zero residual echo signal, the audible contribution at each frequency is the maximum of the residual echo power and the background noise masking threshold. A proportional audible ERLE ($\mathrm{ERLE_{PA}}$) performance measure is obtained by averaging $D(\omega)$ across all frequencies as follows:

$$D(\omega) = \frac{S_{YY}(\omega)}{\max\{S_{\Delta\Delta}(\omega), T_M(\omega)\}} \tag{5.24}$$

$$\mathrm{ERLE_{PA}} = 10\log_{10}\left\{\frac{1}{2\pi}\int_{\omega=0}^{2\pi} D(\omega)d\omega\right\} \tag{5.25}$$

It is important to note that although (5.8) and (5.25) appear similar, they are equivalent only for white echo and residual echo signals that are completely above the masking threshold. Since the masking threshold $T_M(\omega)$ forms a lower bound on residual echo audibility, the maximum $\mathrm{ERLE_{PA}}$ occurs when the residual echo is at or below the noise masking threshold:

$$D_{MAX}(\omega) = \frac{S_{YY}(\omega)}{T_M(\omega)} \tag{5.26}$$

$$ERLE_{PA,MAX} = 10\log_{10}\left\{\frac{1}{2\pi}\int_{\omega=0}^{2\pi} D_{MAX}(\omega)d\omega\right\} \tag{5.27}$$

5.4.2 Calculating the Audible ERLE

A block diagram of the audible ERLE calculation steps is shown in Figure 5.11. Both methods described in Section 5.4.1 require estimates of the echo and residual echo power spectrum, which are obtained from the reference and error signals $d(n)$ and $e(n)$ using spectral subtraction. The masking threshold is calculated from the background noise using the MPEG psychoacoustic model [106]. Finally, the estimates and masking threshold are used to calculate ERLE using either (5.22) or (5.25) for each block. The background noise $\eta(n)$ is assumed to be stationary and its power spectrum is estimated from the reference signal $d(n)$ during periods of quiet (no near- or far-end speech). To that end, Welch's modified periodogram method is employed with N_{FFT}-sample analysis blocks and a Hanning window applied [43]. Individual periodogram estimates are obtained from the FFT of each block, and averaged over the set of B most recent blocks. Let $S_{NN}(k)$ represent the discrete background noise power spectrum estimate, for $0 \le k < N_{FFT}/2$.

The echo and residual echo power spectrum are estimated from the reference and error signals using spectral subtraction. First the power spectra of the reference signal $d(n)$ and error signal $e(n)$ are estimated using the current windowed input block, represented by $S_{DD}(k)$ and $S_{EE}(k)$, respectively. One cannot employ averaging methods

for these signals because real speech inputs can only be assumed to be stationary within periods of 20 – 30 ms [60]. Let $S_{YY}(k)$ and $S_{\Delta\Delta}(k)$ represent the power spectra of the echo and residual echo signals, respectively, estimated as follows:

$$S_{YY}(k) = \max\{S_{DD}(k) - S_{NN}(k), 0\} \tag{5.28}$$

$$S_{\Delta\Delta}(k) = \max\{S_{EE}(k) - S_{NN}(k), 0\} \tag{5.29}$$

The masking threshold $T_M(k)$, $0 \le k < N_{FFT}/2$, is calculated from $S_{NN}(k)$ using the MPEG-1 Psychoacoustic Model 1 [106]. The model was modified to accommodate lower sampling rates (8 – 16 kHz). Finally, the audible ERLE for each block is calculated using discrete-frequency versions of (5.22) or (5.25):

$$\mathrm{ERLE}_{\mathrm{A}} = 10\log_{10}\left[\frac{\frac{1}{|k_y|}\sum_{k_y} S_{YY}(k_y)}{\frac{1}{|k_\delta|}\sum_{k_\delta} S_{\Delta\Delta}(k_\delta)}\right] \tag{5.30}$$

$$\mathrm{ERLE}_{\mathrm{PA}} = 10\log_{10}\left\{\frac{1}{K}\sum_{k=0}^{N_{FFT}/2-1} D(k)\right\} = 10\log_{10}\left\{\frac{1}{K}\sum_{k=0}^{N_{FFT}/2-1}\frac{S_{YY}(k)}{\max[S_{\Delta\Delta}(k), T_M(k)]}\right\} \tag{5.31}$$

where k_y and k_δ represent the set of coefficients for which $S_{YY}(k)$ and $S_{\Delta\Delta}(k)$, respectively, are above the masking threshold $T_M(k)$.

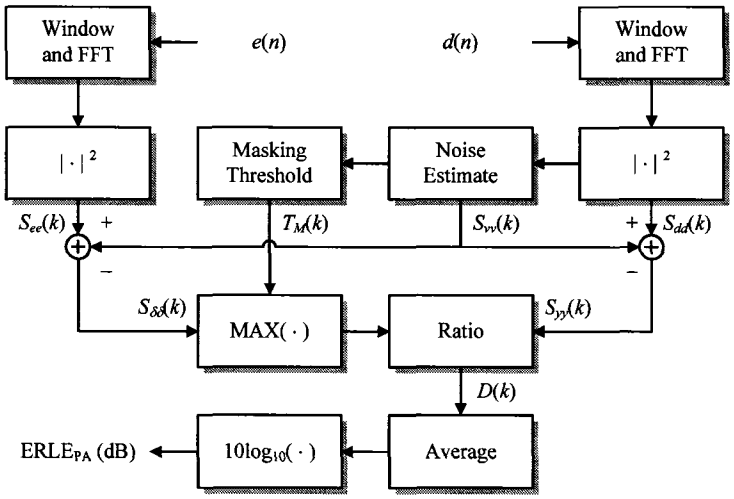


Figure 5.11 - Block diagram of audible ERLE calculation steps.

5.4.3 Simulation and Listening Test Results

5.4.3.1 Experimental Setup

In these experiments the configuration of Section 5.2 was employed. In particular, a sampling rate of f_s = 16 kHz was employed, along with the room impulse response of length L = 2500 samples corresponding to the conference room of Section 3.1. As before, recorded background noise from the room was added to the echo signal, and an echo canceller with N = 2500 samples was adapted using NLMS. Two sets of input signals were applied, the white Gaussian noise sequence of Section 5.2, and 50 continuous speech sequences of 4-6 seconds in length drawn from the TIMIT database. All input signals were calibrated to produce an average echo power of 60 dB SPL in accordance with (5.3). To calculate ERLE_A and ERLE_PA in accordance with (5.30) and (5.31), power spectra of the reference and error signals, $S_{DD}(k)$ and $S_{EE}(k)$, were first measured with (5.3) using the FFT over Hanning-windowed blocks of N_{FFT} = 512 samples. The echo and residual echo power spectra were estimated using (5.28) and

(5.29) with the background noise power spectrum $S_{NN}(k)$ estimated offline. For comparison, ERLE and $\text{ERLE}_{\text{ACTUAL}}$ were calculated using (2.4) and (5.8), respectively.

5.4.3.2 Simulation Results

Figure 5.12(a) shows a comparison of ERLE_{A}, ERLE_{PA}, and ERLE for the white noise input signal during convergence with a step size of $\mu = 0.05$. The maximum levels of cancellation obtainable with (5.29) and (5.30), $\text{ERLE}_{\text{A,MAX}}$ and $\text{ERLE}_{\text{PA,MAX}}$, are indicated for the given noise masking threshold, as are the approximate steady-state times of the algorithms. As observed before in Figure 5.4(a), standard ERLE achieves a steady-state value of 35 dB at approximately $n_1 = 27500$ samples, whereas ERLE_{A} and ERLA_{PA} achieve a maximum reduction in audible echo signal power of 52.5 dB and 47 dB, respectively, at approximately $n_2 = 35000$ samples. Figure 5.12(b) shows the power spectrum of the residual echo signal at the two steady-state times indicated in Figure 5.12(a) compared to the background noise masking threshold. The results are similar to Figure 5.4, but note that the residual echo power spectrum is just at the background noise masking threshold at the steady-state times of ERLE_{A} and ERLE_{PA}.

Figure 5.13(a) shows a four-second speech input signal applied to the test environment, and Figure 5.13(b) shows the standard ERLE calculated using (2.4). Figure 5.14(a) shows a comparison of ERLE_{A} and $\text{ERLE}_{\text{A,MAX}}$ during initial convergence with a step size of $\mu = 0.2$, and Figure 5.14(b) shows a similar comparison for ERLE_{PA} and $\text{ERLE}_{\text{PA,MAX}}$. In this example, both ERLE_{A} and ERLE_{PA} reach their maximum levels after approximately 2.5 seconds of adaptation. The maximum observed ERLE, calculated using (2.4), was approximately 25 dB, which corresponds to the average echo

to noise power ratio. In both of these simulations, it is clear that in the presence of moderate background noise, standard ERLE is a coarse performance measure that may underestimate the amount of cancellation achieved by an echo canceller. In the context of telecommunication network planning, this may prompt the planner to unnecessarily apply a nonlinear processor or additional elements in the loss plan to achieve a desired TELR [11], [12]. In contrast, $ERLE_A$ and $ERLE_{PA}$ give more accurate measures of audible echo power reduction.

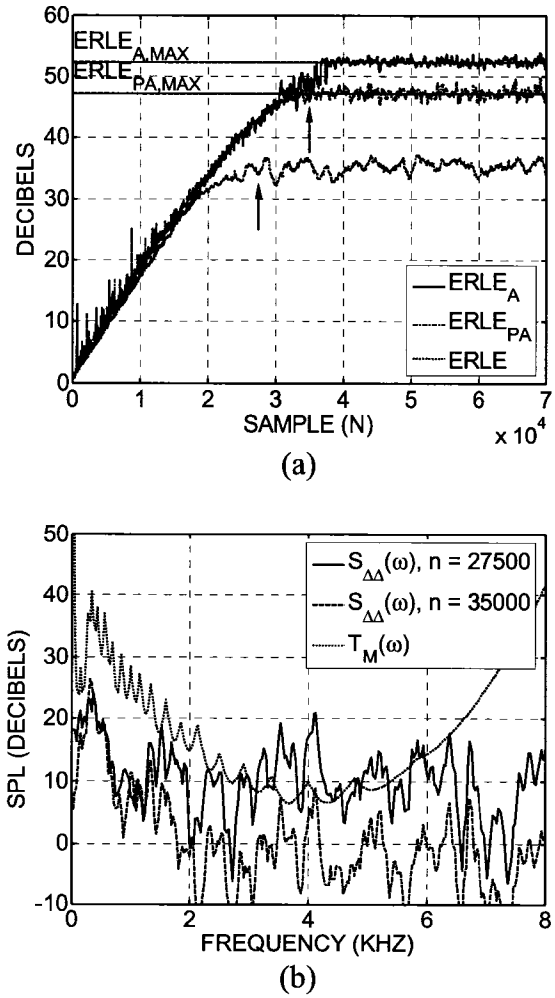


Figure 5.12 – (a) $ERLE_A$ and $ERLE_{PA}$ during convergence with white noise input, compared to standard ERLE; (b) residual echo power spectrum at steady-state times, compared to the background noise masking threshold.

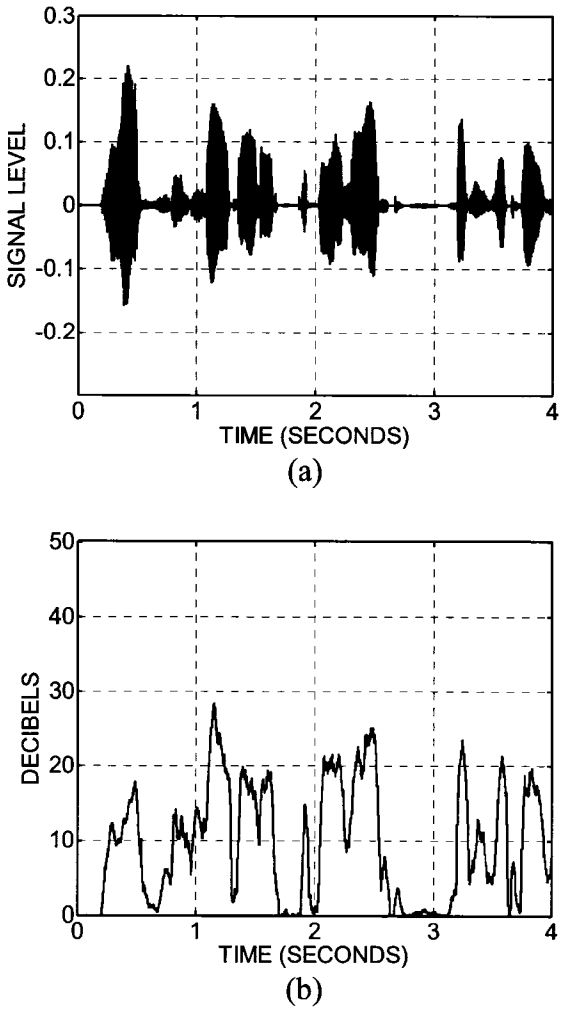


Figure 5.13 – (a) Test speech input signal, and (b) corresponding ERLE measured during initial convergence period in the presence of stationary background noise.

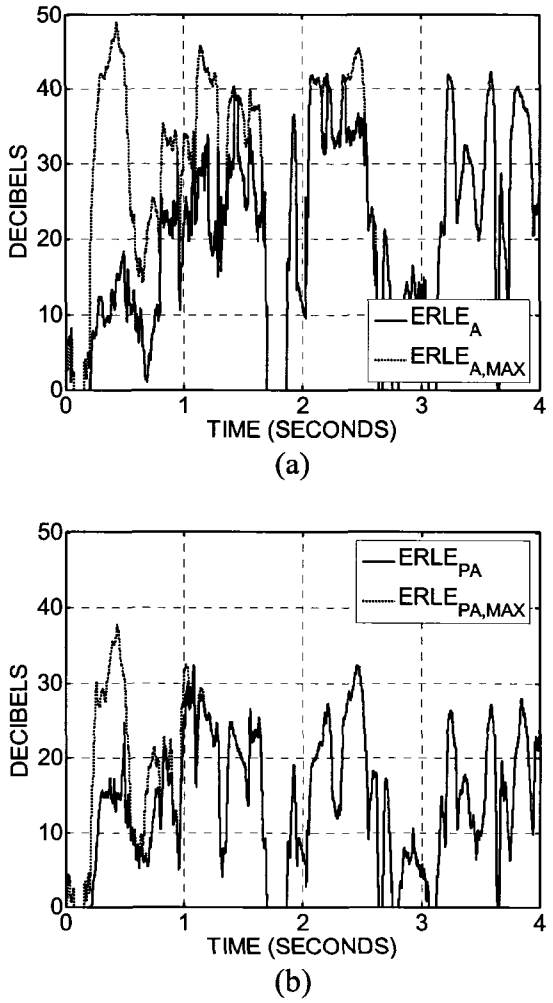


Figure 5.14 – (a) $ERLE_A$ compared to $ERLE_{A,MAX}$, and (b) $ERLE_{PA}$ compared to $ERLE_{PA,MAX}$ during initial convergence with the test speech input signal and stationary background noise.

5.4.3.3 Listening Test Results

Informal tests were conducted with the same panel of four subjects recruited for Section 5.3. Subjects were asked to listen to the echo canceller error signal produced by each of the 50 speech input signals applied to the test environment during their initial convergence periods, repeatedly if necessary, and asked to estimate the convergence time as the earliest time for which the echo signal is no longer audible against the background noise. The convergence time was also estimated manually for each of the input signals

by identifying the earliest time at which $ERLE_A = ERLE_{A,MAX}$ after which no more than ten percent of the remaining $ERLE_A$ values are less than the maximum. Figure 5.15 shows a scatter plot of the observed convergence times averaged from the four subject responses for each of the 50 speech input signals, compared to the manually estimated convergence times. A correlation exists between the estimated convergence times and those reported from the test subjects (ρ = 0.727), which suggests that $ERLE_A$ and $ERLE_{PA}$ may be used to estimate approximate echo canceller convergence times.

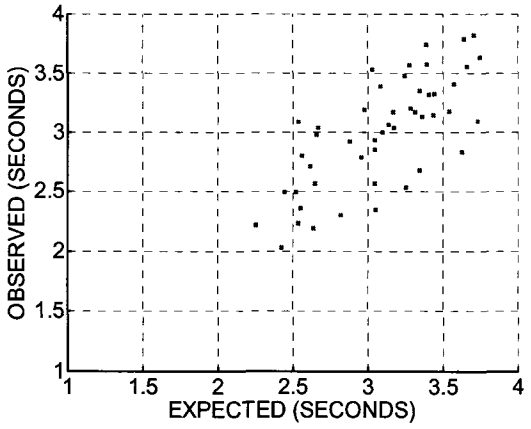


Figure 5.15 - Scatterplot of observed convergence times averaged from four subject responses for 50 speech input signals, compared to expected times obtained manually using $ERLE_A$.

5.5 Summary

This chapter investigated perceived echo canceller performance limitations based on psychoacoustic aspects of human hearing. Echo canceller performance is usually expressed as the reduction in average signal power before and after cancellation (ERLE), which is useful for estimating talker echo loudness rating (TELR) in network planning. However, it is known that moderate-to-high levels of background noise in the environment will limit ERLE to the ratio of average echo to noise power. Section 5.2 presented simulation results illustrating this limit, and it was shown that the audibility of

residual echo is limited by the masking threshold induced by background noise. In Section 5.3 this study was extended to demonstrate that perception of residual echo is frequency-dependent as well, with the highest sensitivity within the 1-5 kHz band. The results of these two sections suggest that performance measures based on average power alone cannot be used to determine the audibility of residual echo, particularly in the presence of background noise. Therefore, in Section 5.4, two echo canceller performance measures were proposed, audible ERLE ($ERLE_A$) and proportional audible ERLE ($ERLE_{PA}$), which estimate the reduction in average audible echo power provided by an echo canceller in the presence of noise. Maximum values of $ERLE_A$ and $ERLE_{PA}$ were derived based on the background noise masking threshold, and used to determine the approximate audible convergence time of an echo canceller. Simulation results and informal listening tests confirmed the validity of the proposed performance measures, and showed that standard ERLE may over- or under-estimate the amount of cancellation provided by an echo canceller.

Chapter 6 Adaptation and Control Algorithms for Critically Sampled Subband Echo Cancellers

6.1 Overview

In Chapter 4 it was shown that statistically optimal doubletalk detection thresholds, and the resulting probability of miss they produce, are dependent on the echo signal to noise ratio (SNR) in the reference signal. Chapter 5 showed that perceived echo canceller performance is dependent on the frequency-dependent effects of the absolute hearing threshold and, in the presence of background noise, on the masking threshold induced by the noise. In many echo cancellation applications, the far-end input and background noise signals are time-varying and not spectrally flat. Therefore, a natural question addressed in this chapter is whether it is possible to apply more sophisticated adaptation and control algorithms to echo cancellers based on subband adaptive filters. Critically sampled subband adaptive filter (CS-SBAF) structures are of interest because they offer a maximal reduction in computational complexity due to decimation. However, they require unique structures to reduce aliasing from non-ideal filter banks. Affine Projection (AP) and cross-correlation-based doubletalk detection have been applied successfully to oversampled subband adaptive echo cancellers [69], [85]. In this chapter it is shown that per-subband AP, cross-correlation-based doubletalk detection, and post-filtering algorithms can be applied to the CS-SBAF structure of [68] reviewed in Section 2.3.2.

The sections of this chapter are organized as follows. Section 6.2 presents a subband Affine Projection algorithm and a convergence analysis of the algorithm. In Section 6.3 the cross correlation-based doubletalk detector of [16] is adapted to provide per-subband doubletalk estimates. Finally, in Section 6.4 the primary results and conclusions of this chapter are summarized.

6.2 Affine Projection in CS-SBAF Structures

Affine Projection (AP) employs an adaptive filter update vector constructed from a P-dimensional projection of the current and $P - 1$ previous tap input vectors onto the *a priori* error signal vector [58]. The projection operation results in an update vector forcing to zero the set of $P - 1$ previous error signals calculated using the *a posteriori* (after update) adaptive filter coefficients. AP reduces to the NLMS algorithm for a projection order of $P = 1$, and so for this reason AP is often viewed as a generalization of the latter. In this section a subband version of Affine Projection, denoted subband AP, is described as a replacement for subband NLMS in the critically sampled subband adaptive filter structure of Section 2.3.2. A convergence analysis of the algorithm is presented, and the algorithm is also analyzed in terms of its computational complexity and implementation considerations.

6.2.1 Algorithm Description

Recall the critically sampled filter bank reviewed in Section 2.3.2. First of all, the subband input signals $x_{i,j}(m)$ and subband reference signals $d_i(m)$ are constructed in accordance with (2.28) and (2.29), respectively. A $P \times 1$ vector of *a priori* error signal

values for the i^{th} subband, $\underline{e}_i(m)$, is constructed by estimating the subband echo signal using the current and adjacent adaptive filter coefficient vectors at time m. In particular:

$$\underline{e}_i(m) = \underline{d}_i(m-d) - \underline{\hat{y}}_i(m) \tag{6.1}$$

$$\underline{d}_i(m) = [d_i(m) \quad \cdots \quad d_i(m-P+1)]^T \tag{6.2}$$

$$\underline{\hat{y}}_i(m) = \underline{X}_{i,i-1}^T(m)\underline{\hat{w}}_{i-1}(m) + \underline{X}_{i,i}^T(m)\underline{\hat{w}}_i(m) + \underline{X}_{i,i+1}^T(m)\underline{\hat{w}}_{i+1}(m) \tag{6.3}$$

$$\underline{X}_{i,i-1}(m) = [\underline{x}_{i,i-1}(m) \quad \cdots \quad \underline{x}_{i,i-1}(m-P+1)] \tag{6.4}$$

$$\underline{X}_{i,i}(m) = [\underline{x}_{i,i}(m) \quad \cdots \quad \underline{x}_{i,i}(m-P+1)] \tag{6.5}$$

$$\underline{X}_{i,i+1}(m) = [\underline{x}_{i,i+1}(m) \quad \cdots \quad \underline{x}_{i,i+1}(m-P+1)] \tag{6.6}$$

where $\underline{x}_{i,j}(m) = [x_{i,j}(m) \quad x_{i,j}(m-1) \quad \cdots \quad x_{i,j}(m-N_D+1)]^T$ represents a tap-input vector of length N_D samples. An estimate of the gradient vector for each subband is constructed such that it forces the *a posteriori* error signal vector to zero. The resulting adaptive filter update for the i^{th} subband is given by the following equations:

$$\underline{\hat{w}}_i(m+1) = \underline{\hat{w}}_i(m) + \mu\underline{\nabla J}_i(m) \tag{6.7}$$

$$\begin{aligned}\underline{\nabla J}_i(m) = {} & \underline{X}_{i,i-1}(m)\underline{C}_i^{-1}(m)\underline{e}_{i-1}(m) + \underline{X}_{i,i}(m)\underline{C}_i^{-1}(m)\underline{e}_i(m) \\ & + \underline{X}_{i,i+1}(m)\underline{C}_i^{-1}(m)\underline{e}_{i+1}(m)\end{aligned} \tag{6.8}$$

$$\underline{C}_i(m) = \underline{X}_{i,i-1}^T(m)\underline{X}_{i,i-1}(m) + \underline{X}_{i,i}^T(m)\underline{X}_{i,i}(m) + \underline{X}_{i,i+1}^T(m)\underline{X}_{i,i+1}(m) + \delta\underline{I} \tag{6.9}$$

where $\underline{C}_i(m)$ forms an estimate of the input signal autocorrelation matrix, and μ is the step-size parameter. Equation (6.9) has a small scalar regularization parameter δ to avoid potential stability problems while calculating the inverse of the matrix $\underline{C}_i(n)$. An alternative viewpoint is that the update equation at each time interval minimizes the per-sample change in adaptive filter coefficient values subject to the P^{th}-order constraints of

(6.1) – (6.6). As before, for subbands $i = 1$ and $i = M$, (6.3) – (6.9) are adjusted to remove terms involving subbands $i - 1$ and $i + 1$, respectively. Note that for $P = 1$, subband AP reduces to the original subband NLMS algorithm of Section 6.2.2.

6.2.2 Convergence Analysis

In this section a convergence analysis of subband AP is presented for the case of $M =$ 2 subbands. First of all, let $\underline{\varepsilon}_1(m)$ and $\underline{\varepsilon}_2(m)$ represent the $N_D \times 1$ error vectors between the optimal and estimated adaptive filter vectors at time m for subbands 1 and 2, respectively:

$$\underline{\varepsilon}_1(m) = \underline{w}_1(m) - \underline{\hat{w}}_1(m) \tag{6.10}$$

$$\underline{\varepsilon}_2(m) = \underline{w}_2(m) - \underline{\hat{w}}_2(m) \tag{6.11}$$

Substituting (6.10) and (6.11) into (6.7) yields the adaptive filter update equations in terms of the error vectors as follows:

$$\underline{\varepsilon}_1(m+1) = \underline{\varepsilon}_1(m) - \mu \underline{\nabla} J_1(m) \tag{6.12}$$

$$\underline{\varepsilon}_2(m+1) = \underline{\varepsilon}_2(m) - \mu \underline{\nabla} J_2(m) \tag{6.13}$$

Writing the subband error signal vectors in terms of (6.1) and (6.3), the adaptive filter error vectors of (6.12) and (6.13), and the steady-state prediction error vectors $\underline{E}_1(m)$ and $\underline{E}_2(m)$ gives:

$$\underline{e}_1(m) = \underline{X}_{1,1}^T \underline{\varepsilon}_1(m) + \underline{X}_{1,2}^T \underline{\varepsilon}_2(m) + \underline{E}_1(m) \tag{6.14}$$

$$\underline{e}_2(m) = \underline{X}_{1,2}^T \underline{\varepsilon}_1(m) + \underline{X}_{2,2}^T \underline{\varepsilon}_2(m) + \underline{E}_2(m) \tag{6.15}$$

Now substitute (6.14) and (6.15) into the gradient estimate vector of (6.8) and, in turn, (6.12) and (6.13). This results in the following adaptive filter coefficient error update equations:

$$\begin{aligned}\underline{\varepsilon}_1(m+1) = \underline{\varepsilon}_1(m) \\ -\mu\{\underline{X}_{1,1}(m)\underline{C}_1^{-1}(m)[\underline{X}_{1,1}^T(m)\underline{\varepsilon}_1(m)+\underline{X}_{1,2}^T(m)\underline{\varepsilon}_2(m)+\underline{E}_1(m)]\} \\ -\mu\{\underline{X}_{1,2}(m)\underline{C}_1^{-1}(m)[\underline{X}_{1,2}^T(m)\underline{\varepsilon}_1(m)+\underline{X}_{2,2}^T(m)\underline{\varepsilon}_2(m)+\underline{E}_2(m)]\}\end{aligned} \tag{6.16}$$

$$\begin{aligned}\underline{\varepsilon}_2(m+1) = \underline{\varepsilon}_2(m) \\ -\mu\{\underline{X}_{1,2}(m)\underline{C}_2^{-1}(m)[\underline{X}_{1,1}^T(m)\underline{\varepsilon}_1(m)+\underline{X}_{1,2}^T(m)\underline{\varepsilon}_2(m)+\underline{E}_1(m)]\} \\ -\mu\{\underline{X}_{2,2}(m)\underline{C}_2^{-1}(m)[\underline{X}_{1,2}^T(m)\underline{\varepsilon}_1(m)+\underline{X}_{2,2}^T(m)\underline{\varepsilon}_2(m)+\underline{E}_2(m)]\}\end{aligned} \tag{6.17}$$

From this point on the independence assumption is employed, in particular by assuming that the input signal $x(n)$ and optimal adaptive filter vectors $\underline{w}_1(m)$ and $\underline{w}_2(m)$ are stationary, and that the steady-state prediction error vectors $\underline{E}_1(m)$ and $\underline{E}_2(m)$ are zero-mean random processes independent of the input signal [44]. With the assumptions above, taking the expected value of equations (6.16) and (6.17) yields:

$$\begin{bmatrix}E\{\underline{\varepsilon}_1(m+1)\}\\ E\{\underline{\varepsilon}_2(m+1)\}\end{bmatrix} = [\underline{I}_{2L} - \mu\underline{\Phi}]\cdot\begin{bmatrix}E\{\underline{\varepsilon}_1(m)\}\\ E\{\underline{\varepsilon}_2(m)\}\end{bmatrix} \tag{6.18}$$

$$\underline{\Phi} = E\left\{\begin{bmatrix}\underline{A} & \underline{B}\\ \underline{C} & \underline{D}\end{bmatrix}\right\} \tag{6.19}$$

$$\underline{A} = \underline{X}_{1,1}(m)\underline{C}_1^{-1}(m)\underline{X}_{1,1}^T(m) + \underline{X}_{1,2}(m)\underline{C}_1^{-1}(m)\underline{X}_{1,2}^T(m) \tag{6.20}$$

$$\underline{B} = \underline{X}_{1,1}(m)\underline{C}_1^{-1}(m)\underline{X}_{1,2}^T(m) + \underline{X}_{1,2}(m)\underline{C}_1^{-1}(m)\underline{X}_{2,2}^T(m) \tag{6.21}$$

$$\underline{C} = \underline{X}_{1,2}(m)\underline{C}_2^{-1}(m)\underline{X}_{1,1}^T(m) + \underline{X}_{2,2}(m)\underline{C}_2^{-1}(m)\underline{X}_{1,2}^T(m) \tag{6.22}$$

$$\underline{D} = \underline{X}_{1,2}(m)\underline{C}_2^{-1}(m)\underline{X}_{1,2}^T(m) + \underline{X}_{2,2}(m)\underline{C}_2^{-1}(m)\underline{X}_{2,2}^T(m) \tag{6.23}$$

From (6.18) it is clear that the rate of convergence of the expected adaptive filter error vectors depends on the eigenvalue spread of the matrix $\underline{\Phi}$. For the case where the original input signal $x(n)$ consists of white noise, $\underline{\Phi}$ reduces to a $2N_D \times 2N_D$ diagonal matrix with components equal to the one-half of the input signal variance. In general, the sub-matrices $\underline{B}$ and $\underline{C}$ are non-zero, which implies that the convergence of the two subband coefficient vectors is dependent on each other. In [87] an analysis of AP for the single-channel (fullband) case showed that for small step size ($\mu << 1$) and projection order P, the progression of the mean square error can be approximated by the following:

$$E[e^2(n)] = \sigma_y^2 \sum_{i=1}^{N} (1-\alpha\beta_i)^n \frac{\lambda_i}{tr(\underline{R}_{xx})} \tag{6.24}$$

$$\alpha = \mu(2-\mu) \tag{6.25}$$

$$\beta_i = 1-[1-\lambda_i / tr\{\underline{R}_{xx}\}]^P \tag{6.26}$$

where σ_y^2 is the variance of the echo signal, N is the length of the system, $\underline{R}_{xx}$ is the $N \times N$ covariance matrix of the input signal, and λ_i are the eigenvalues of $\underline{R}_{xx}$. For subband AP, experimentally it was found that in many cases the matrix $\underline{\Phi}$ is dominated by the "in-band" terms, and so the sub-matrices $\underline{B}$ and $\underline{C}$ are relatively small compared to the sub-matrices $\underline{A}$ and $\underline{D}$. In other words:

$$\underline{A} \approx \underline{X}_{1,1}(m)\underline{C}_1^{-1}(m)\underline{X}_{1,1}^T(m) \tag{6.27}$$

$$\underline{B}, \underline{C} \approx 0 \tag{6.28}$$

$$\underline{D} \approx \underline{X}_{2,2}(m)\underline{C}_2^{-1}(m)\underline{X}_{2,2}^T(m) \tag{6.29}$$

After substituting the equations above into (6.18), (6.24) – (6.26) can be used to approximate the mean square error convergence of subbands 1 and 2 by replacing $\underline{R}_{xx}$ with the autocorrelation matrices of $x_{1,1}(m)$ and $x_{2,2}(m)$, respectively.

6.2.3 Computational Complexity

A straightforward implementation of AP for the single-channel (fullband) case requires approximately $2NP + P_{INV}P^2$ multiplications per sample, where N is the length of the fullband adaptive filter, P is the projection order, and P_{INV} is a constant term representing the cost of calculating a $P \times P$ matrix inverse [61]. Employing a similar implementation for subband AP requires approximately $N_D P(3M - 2) + N_D(3M - 2) + P_{INV}P^2(3M - 2)$ multiplications per sample, where N_D is the subband adaptive filter length and M is the number of subbands. However, it is important to note that processing is performed on *downsampled* data. If the length of the system is large compared to the analysis and synthesis filter lengths ($N >> N_H$, N_F), then subband AP requires approximately $\frac{N(P+1)(3M-2)}{M^2} + \frac{P_{INV}P^2(3M-2)}{M}$ multiplications per fullband sample, which is a significant reduction from the fullband case. Figure 6.1 shows the complexity in multiplications per sample for fullband NLMS / AP compared to subband NLMS / AP as a function of fullband adaptive filter length. As outlined in the introduction, the use of a filter bank employing critical sampling allows a maximal reduction in computational complexity compared to using an oversampled filter bank as in [69]. For example, the complexity of subband AP with projection order of $P = 4$ using $M = 4$ subbands is less than that of fullband AP with projection order of $P = 4$. Techniques are available to further reduce the complexity of subband AP, such as estimating the correlation matrixes

$\underline{C}_i(m)$ using a sliding window [69]. Although not explored further in this section, it may also be possible to derive a "fast" version of subband AP similar to the fast affine projection algorithm and its variants [61].

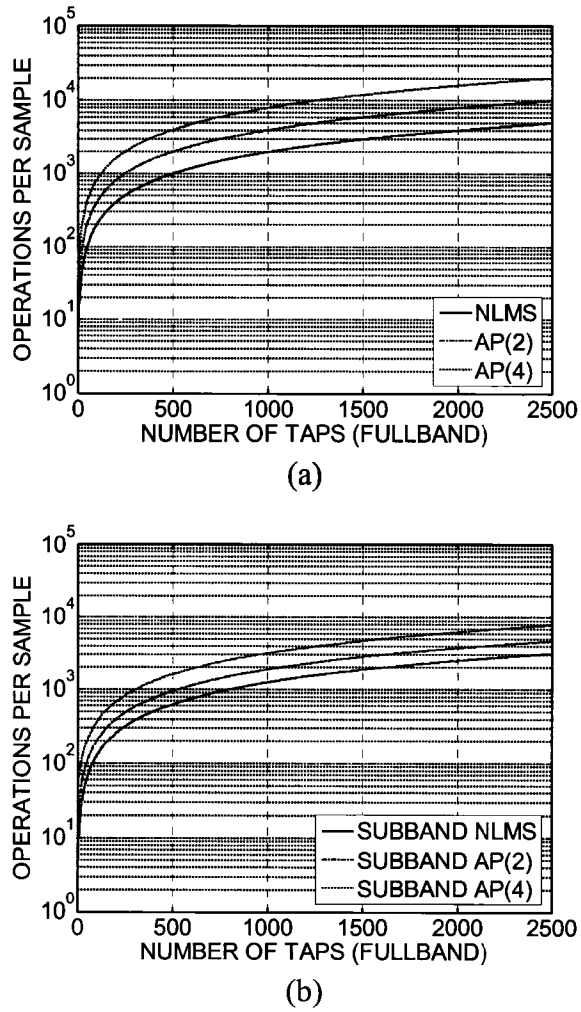


Figure 6.1 – Complexity of fullband and subband adaptation algorithms as a function of fullband adaptive filter length for (a) fullband NLMS / AP and (b) subband NLMS / AP ($M = 4$ subbands).

6.2.4 Simulation Results

6.2.4.1 Simulation Setup

In this study an M-channel pseudo-QMF filter bank was employed for the adaptive filter structure [88]. FIR, linear phase lowpass prototype filters $p_0(n)$ were designed

using the Parks-McClellan algorithm, with a stopband edge at $\omega_S = \pi / M$ and a passband edge adjusted to minimize the following objective function [43]:

$$\phi = \max_{0<\omega<\pi/M} \left[\left| P_0(e^{j\omega}) \right|^2 + \left| P_0(e^{j(\omega-\pi/M)}) \right|^2 - 1 \right] \tag{6.30}$$

Prototype filters were constructed for M = 2, 4, and 8 subbands, and for each filter the length N_P was chosen to achieve a stopband attenuation of 100 dB. Figure 6.2 shows the magnitude responses of the prototype filters. The analysis and synthesis filters for the i^{th} subband, $h_i(n)$ and $f_i(n)$, were obtained by cosine modulation of the lowpass prototype filter in accordance with [88]:

$$h_i(n) = 2p_0(n)\cos[(i+\tfrac{1}{2})(n-\tfrac{N_P}{2})\tfrac{\pi}{M} + (-1)^i \tfrac{\pi}{4}] \tag{6.31}$$

$$f_i(n) = 2p_0(n)\cos[(i+\tfrac{1}{2})(n-\tfrac{N_P}{2})\tfrac{\pi}{M} - (-1)^i \tfrac{\pi}{4}] \tag{6.32}$$

The simulated echo path of length N = 500 samples from Figure 3.3 was employed in the following simulations. The echo signal $y(n)$ was constructed by convolving the input

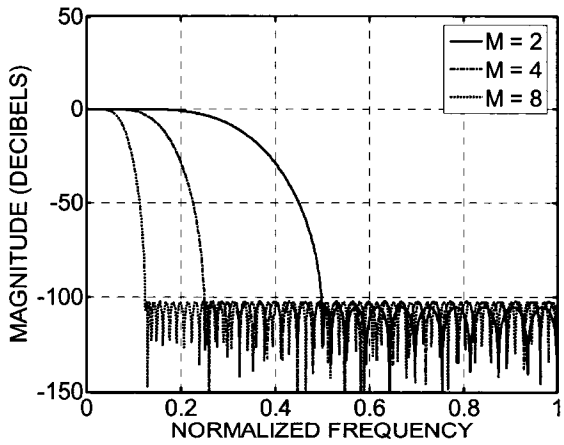


Figure 6.2 - Magnitude responses of lowpass prototype filters for M = 2, 4, and 8 subbands.

signal $x(n)$ with the impulse response, to which background noise was added to produce the reference signal $d(n)$. Throughout this study a sampling rate of f_s = 8 kHz was used. Two sets of input signals were used, colored noise and speech. The noise input signal was a stationary 10^{th}-order autoregressive (AR) process whose power spectrum is shown in Figure 6.3(a). The AR coefficients were {-0.888, 0.558, -1.374, 0.533, -1.137, 0.683, 0.055, 0.057, -0.383, 0.067}. In addition, a 40-second input signal consisting of continuous speech from a male speaker was obtained, as shown in Figure 6.3(b). Two configurations of the critically sampled subband adaptive filter structure were compared: the subband NLMS algorithm from Section 6.2.2 and the subband AP from Section 6.3.1. In the simulations projection orders of P = 2, 3, and 4 were used. Performance was measured using the mean square error (MSE) of the fullband error signal obtained at the synthesis filter bank output. For comparison, results were also obtained for fullband implementations of NLMS and AP.

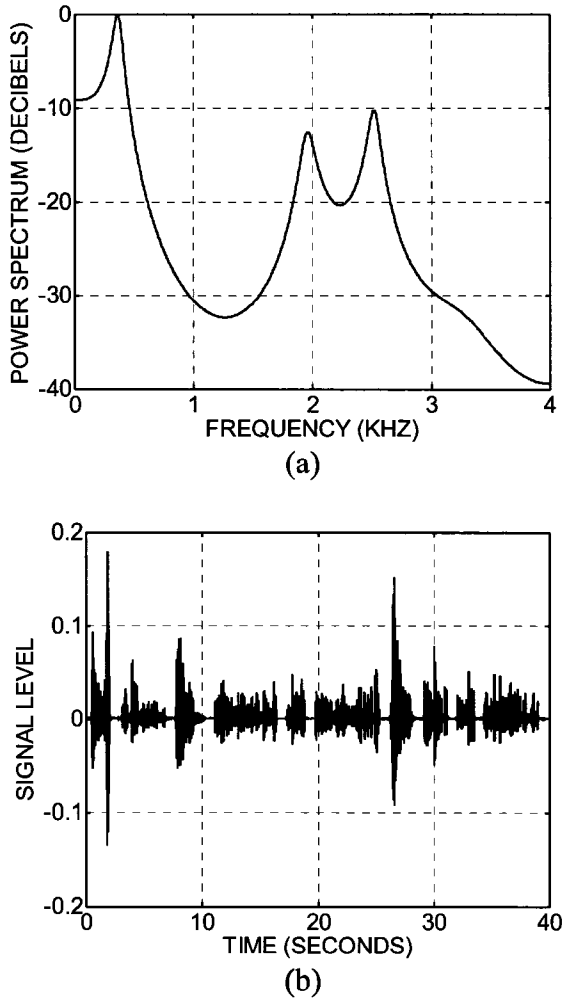


Figure 6.3 – (a) Power spectrum of autoregressive noise input signal; (b) test input signal consisting of continuous speech from a male speaker.

6.2.4.2 Convergence and Tracking

Figures 6.4, Figure 6.5, and Figure 6.6 show the MSE as a function of time for $M = 2$, $M = 4$, and $M = 8$ subbands, respectively, compared to subband NLMS ($P = 1$). For comparison, Figure 6.7 shows the MSE obtained using fullband implementations of NLMS and AP. In this simulation the colored noise input signal, a step size of $\mu = 0.5$, and white background noise of -60 dB power were used. From these results several important observations can be made. First of all, in theory increasing the number of

subbands should whiten the signal in each subband and increase the rate of convergence of the subband NLMS algorithm of Section 2.3.2. However, this effect depends on the input signal spectrum, and in this experiment there is little difference in the MSE between $M = 2$ and $M = 4$ cases for the subband NLMS algorithm. It is not until the number of subbands is increased further, to $M = 8$, that an increase in performance is observed for subband NLMS. In addition, it is quite clear from Figures 6.4 and 6.5 that the subband AP of Section 6.2.1 introduces a considerable improvement in convergence rate even for a relatively small number of subbands and for low projection orders of $P = 2$ and $P = 3$. One possible reason is that the whitening effect of employing a subband decomposition is of more assistance to the AP algorithm than NLMS. It is known that AP with a projection order of P can completely decorrelate an autoregressive input signal process of the same order [59]. By partially decorrelating the input signal in each subband, this may reduce the "effective" order of the input signal process further. Finally, the results also reveal that the performance improvement diminishes as the projection order increases. This is in agreement with previously published experiments with Affine Projection [87].

In applications such as acoustic echo cancellation, the echo path impulse response is often time-varying, so it is important to compare algorithms with respect to re-convergence due to a change in the unknown system. Figure 6.8 shows the MSE as a function of time for the case of $M = 8$ subbands with a change in all of the impulse response coefficients occurring at $n = 25000$ samples. For comparison, Figure 6.9 shows the outcome of the same experiment using fullband implementations of NLMS and AP.

From these results it is clear that the performance improvement obtained by using subband AP remains even during a re-convergence of the adaptive filter coefficients.

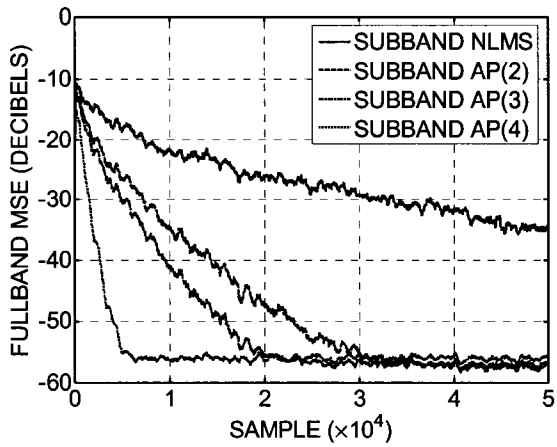


Figure 6.4 – Fullband mean square error during initial convergence for $M = 2$ subbands using subband NLMS and subband AP ($P = 2, 3, 4$).

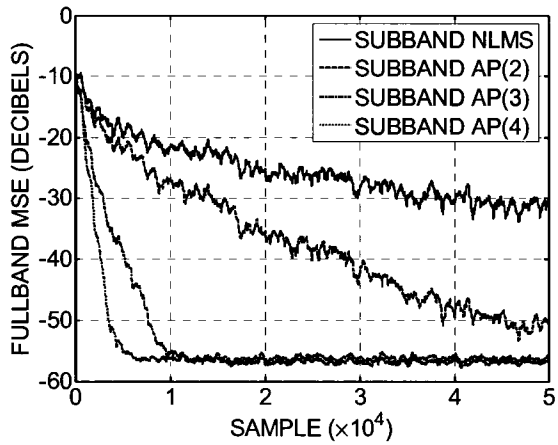


Figure 6.5 – Fullband mean square error during initial convergence for $M = 4$ subbands using subband NLMS and subband AP ($P = 2, 3, 4$).

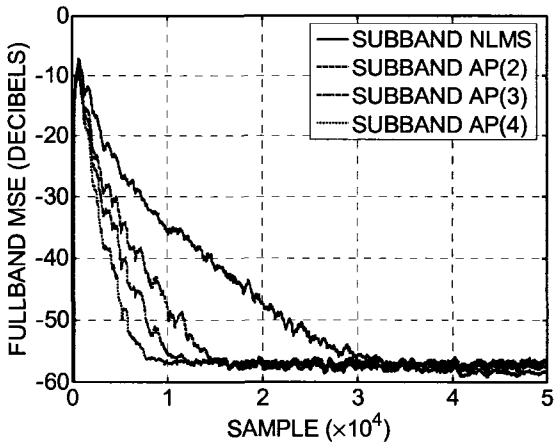


Figure 6.6 – Fullband mean square error during initial convergence for $M = 8$ subbands using subband NLMS and subband AP ($P = 2, 3, 4$).

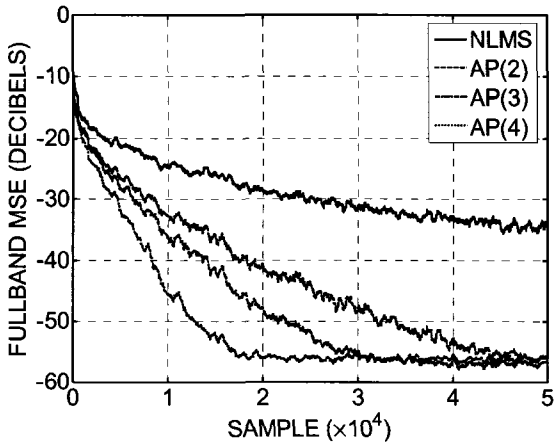


Figure 6.7 – Mean square error during initial convergence using fullband NLMS and AP ($P = 2, 3, 4$).

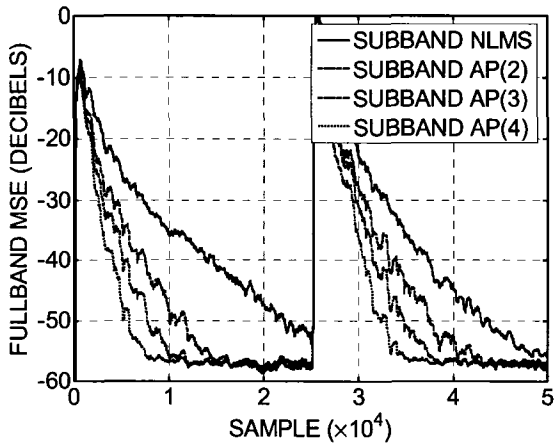


Figure 6.8 – Fullband mean square error in the presence of an echo path change at n = 25000 samples for M = 8 subbands using subband NLMS and subband AP (P = 2, 3, 4).

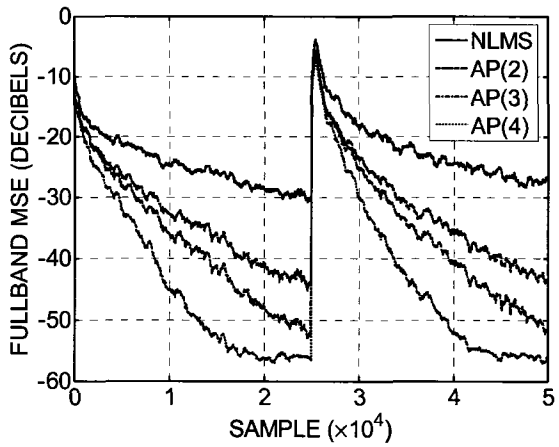


Figure 6.9 – Mean square error in the presence of an echo path change at n = 25000 samples using fullband NLMS and AP (P = 2, 3, 4).

6.2.4.3 Experimental Results

To verify these results in a physical environment, the speech input signal was played through a Tannoy Reveal loudspeaker in the conference room described in Section 3.1. The input and reference signals were fed into the subband NLMS and subband AP algorithms, both with M = 4 subbands and a step size of μ = 0.25. For subband AP a projection order of P = 4 was used. For comparison, fullband NLMS and AP (P = 4) were applied to the same input signal and echo path. The noise floor of the room was

measured beforehand to be approximately -50 dB, and Figure 6.10 shows the measured room impulse response for $N = 2500$ samples obtained using subband AP. Figure 6.11 shows the MSE from both fullband and subband configurations. In this case it is clear that subband AP converges significantly faster than fullband NLMS / AP and slightly faster than subband NLMS, even using low projection order, resulting in 5 – 10 dB lower average MSE over the course of the simulation.

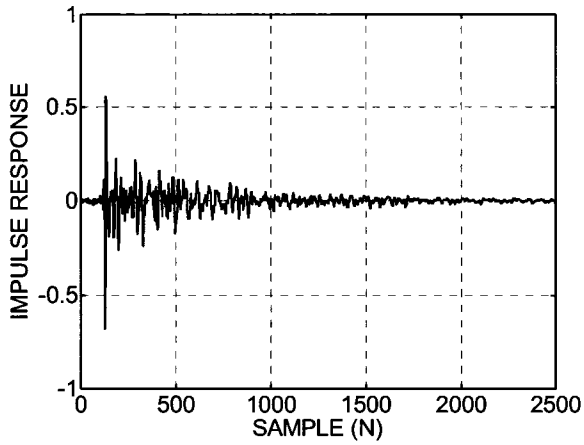


Figure 6.10 - Room impulse response measured using subband AP with a speech input signal ($M = 4$ subbands, $P = 4$).

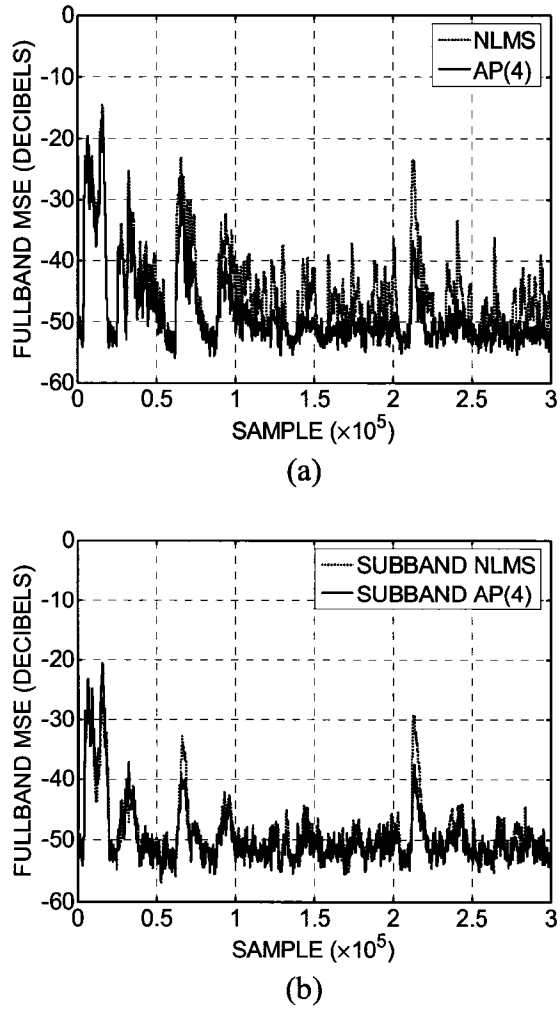


Figure 6.11 – Comparison of mean square error performance of speech input signal using (a) fullband NLMS / AP ($P = 4$); (b) subband NLMS / AP NLMS ($M = 4$ subbands, $P = 4$).

6.2.4.4 Verification of Convergence Analysis Model

To verify the convergence analysis model presented in Section 6.2.2, the experiment conducted in Section 6.2.4.2 was re-run for $M = 2$ subbands using subband AP with a projection order of $P = 2$. The same echo path impulse response, colored noise input signal and background noise conditions were employed, but with a smaller adaptation step size of $\mu = 0.1$. Figure 6.12 shows the MSE as a function of time for subbands 1 and 2, compared to the theoretical MSE evolution of the two subbands calculated using (6.24)

– (6.26) under the assumptions of (6.27) – (6.29). It is clear in this case that the theoretical MSE provides a good approximation to the actual MSE convergence.

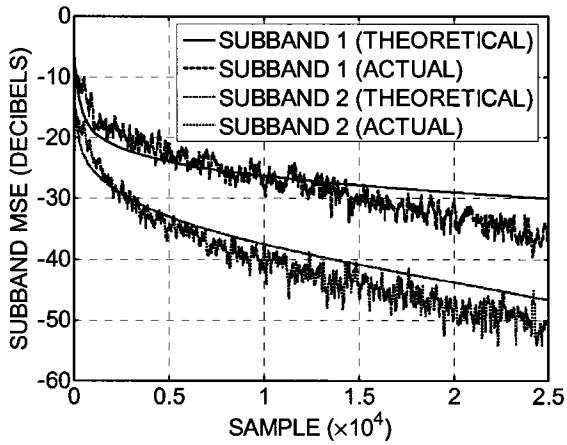


Figure 6.12 – Theoretical mean square error calculated using (6.24) for subbands 1 and 2 ($M = 2$ subbands), compared to measured mean square error as a function of time.

6.3 Doubletalk Detection in CS-SBAF Structures

In Chapter 4 it was shown that statistically optimal doubletalk detection thresholds, and the resulting performance, are dependent upon the echo to noise ratio (SNR). Since in practical echo canceller implementations the input and noise signals are not spectrally flat, a natural question is whether doubletalk detector performance can be improved by incorporating frequency-dependent doubletalk decisions. To address this question, in this section it is shown that the cross-correlation-based doubletalk detector of [16] can be derived for use in the critically sampled subband adaptive filter structure of [68]. The result is a per-subband doubletalk detector that is responsive to near-end speech within each subband, and capable of independently controlling adaptation of each of the subband adaptive filters. Methods for adaptive calibration and bias compensation, presented in Chapter 4, are shown to be easily adapted for use in the proposed subband

doubletalk detector. Computer simulations are presented to investigate the performance of the structure in terms of doubletalk detection.

6.3.1 Algorithm Derivation

The algorithm described in this section is based on the normalized cross-correlation between the subband input and subband reference signals. It is an extension of the doubletalk detector introduced in [16] for fullband echo cancellers, and the derivation follows similarly. First of all, assume that each subband echo signal can be perfectly constructed using the true subband echo path impulse responses $\underline{w}_i(m)$ and subband input signals, and also assume that the near-end speech is uncorrelated with the echo signal. The reference signal in the i^{th} subband, $d_i(m)$, can be obtained from (2.31):

$$d_i(m) = y_i(m) + v_i(m) = \underline{x}_{i,i-1}^T(m)\underline{w}_{i-1} + \underline{x}_{i,i}^T(m)\underline{w}_i + \underline{x}_{i,i+1}^T(m)\underline{w}_{i+1} + v_i(m) \qquad (6.33)$$

where $\underline{x}_{i,j}(m)$ represents the $N_D \times 1$ subband input signal vector generated by (2.28), $\underline{w}_i$ is the $N_D \times 1$ true echo path impulse response vector for the i^{th} subband, and $v_i(m)$ is the subband near-end speech signal. Since the analysis filters are assumed to have high stopband attenuation in non-adjacent subbands, the expected subband echo signal variance can be written as follows:

$$\sigma_{y,i}^2 = E[y_i^2(m)] = \underline{w}_{i-1}^T \underline{R}_{i,i-1}\underline{w}_{i-1} + \underline{w}_i^T \underline{R}_{i,i}\underline{w}_i + \underline{w}_{i+1}^T \underline{R}_{i,i+1}\underline{w}_{i+1} \qquad (6.34)$$

where $\underline{R}_{i,j} = E[\underline{x}_{i,i}(m)x_{i,j}(m)]$ is the $N_D \times N_D$ cross-correlation matrix between the subband input signal vectors $x_{i,i}(m)$ and $x_{i,j}(m)$, for $1 \le i, j \le M$. Now consider the cross-correlation vector between the subband input signal vector $\underline{x}_{i,i}(m)$ and the scalar subband reference signal $d_i(m)$, denoted $\underline{r}_{i,i}$. Assuming that the near-end speech in the subband, $v_i(m)$, is

uncorrelated with the input signal, substituting (6.33) into the cross-correlation vector equation yields the following:

$$\begin{aligned} \underline{r}_{i,i} &= E[\underline{x}_{i,i}(m)d_i(m)] \\ &= E\{\underline{x}_{i,i}(m)[y_i(m)+v_i(m)]\} \\ &= E\{\underline{x}_{i,i}(m)[\underline{x}_{i,i-1}^T(m)\underline{w}_{i-1}+\underline{x}_{i,i}^T(m)\underline{w}_i+\underline{x}_{i,i+1}^T(m)\underline{w}_{i+1}+v_i(m)]\} \\ &= \underline{R}_{i,i-1}\underline{w}_{i-1}+\underline{R}_{i,i}\underline{w}_i+\underline{R}_{i,i+1}\underline{w}_{i+1} \end{aligned} \tag{6.35}$$

Assuming that $\underline{R}_{i,i-1}$ and $\underline{R}_{i,i+1}$ are small compared to $\underline{R}_{i,i}$ then $\underline{w}_i$ can be approximated by inverting the autocorrelation matrix $\underline{R}_{i,i}$. Following similar arguments one can construct approximations for $\underline{w}_{i-1}$ and $\underline{w}_{i+1}$ as well:

$$\underline{w}_{i-1} \approx \underline{R}_{i,i-1}^{-1}\underline{r}_{i,i-1} \tag{6.36}$$

$$\underline{w}_i \approx \underline{R}_{i,i}^{-1}\underline{r}_{i,i} \tag{6.37}$$

$$\underline{w}_{i+1} \approx \underline{R}_{i,i+1}^{-1}\underline{r}_{i,i+1} \tag{6.38}$$

Substituting (6.36) – (6.38) into (6.34) provides an estimate of the subband echo signal variance in terms of the subband input signal cross-correlation matrices, and the cross-correlation vectors between the subband input and reference signals:

$$\sigma_{y,i}^2 = \underline{r}_{i,i-1}^T\underline{R}_{i,i-1}^{-1}\underline{r}_{i,i-1}+\underline{r}_{i,i}^T\underline{R}_{i,i}^{-1}\underline{r}_{i,i}+\underline{r}_{i,i+1}^T\underline{R}_{i,i+1}^{-1}\underline{r}_{i,i+1} \tag{6.39}$$

Finally, a normalized doubletalk detection statistic can be obtained by dividing the right-hand side of (6.39) by the measured subband reference signal variance and taking the square root of the result:

$$\xi_i = \sqrt{\frac{\underline{r}_{i,i-1}^T\underline{R}_{i,i-1}^{-1}\underline{r}_{i,i-1}+\underline{r}_{i,i}^T\underline{R}_{i,i}^{-1}\underline{r}_{i,i}+\underline{r}_{i,i+1}^T\underline{R}_{i,i+1}^{-1}\underline{r}_{i,i+1}}{\sigma_{d,i}^2}} \tag{6.40}$$

Equation (6.40) represents the ratio of the expected to actual subband reference signal variances. Clearly $\xi_i = 1$ in the absence of doubletalk, otherwise the denominator contains an additional term from the subband near-end speech signal $v_i(m)$, resulting in $\xi_i < 1$.

6.3.2 Implementation Considerations

The input and reference signals in (6.40) vary with time, so a practical implementation of the doubletalk detection statistic is a time-varying function $\xi_i(m)$ obtained by estimating the parameters at each sample:

$$\xi_i(m) = \sqrt{\frac{\underline{\hat{r}}_{i,i-1}^T(m)\underline{\hat{R}}_{i,i-1}^{-1}(m)\underline{\hat{r}}_{i,i-1}(m) + \underline{\hat{r}}_{i,i}^T(m)\underline{\hat{R}}_{i,i}^{-1}(m)\underline{\hat{r}}_{i,i}(m) + \underline{\hat{r}}_{i,i+1}^T(m)\underline{\hat{R}}_{i,i+1}^{-1}(m)\underline{\hat{r}}_{i,i+1}(m)}{\hat{\sigma}_{d,i}^2(m)}} \tag{6.41}$$

The detection statistic in (6.41) requires estimates of three $N_D \times 1$ cross-correlation vectors between the subband input signals and the subband reference signal. As shown in Section 2.4, these parameters can be estimated by employing an estimation window of K samples over which they are averaged. Similarly, the reference signal variance in each subband can be estimated using a standard unbiased variance estimator over a window of K samples:

$$\underline{\hat{r}}_{i,i-1}(m) \approx \frac{1}{K}\sum_{k=0}^{K-1}\underline{x}_{i,i-1}(m-k)d_i(m-k) \tag{6.42}$$

$$\underline{\hat{r}}_{i,i}(m) \approx \frac{1}{K}\sum_{k=0}^{K-1}\underline{x}_{i,i}(m-k)d_i(m-k) \tag{6.43}$$

$$\underline{\hat{r}}_{i,i+1}(m) \approx \frac{1}{K}\sum_{k=0}^{K-1}\underline{x}_{i,i+1}(m-k)d_i(m-k) \tag{6.44}$$

$$\hat{\sigma}_{d,i}^2(m) = \frac{1}{K-1}\sum_{k=0}^{K-1}\left[d_i(m-k) - \frac{1}{K}\sum_{j=0}^{K-1} d_i(m-j) \right]^2 \tag{6.45}$$

However, it is also necessary to estimate three $N_D \times N_D$ cross-correlation matrices of the subband input signals as well as their inverses. These are expensive operations, particularly if a doubletalk decision statistic is required for each of the M subbands. One way to avoid this problem arises once the subband adaptive filters have converged. In this case it is possible to utilize the adaptive filter coefficient vectors to approximate (6.36) – (6.38), which in turn can be substituted into (6.41). Following this procedure yields a simplified doubletalk detection statistic:

$$\xi_i(m) = \sqrt{\frac{\underline{\hat{r}}_{i,i-1}^T(m)\underline{\hat{w}}_{i-1}(m) + \underline{\hat{r}}_{i,i}^T(m)\underline{\hat{w}}_i(m) + \underline{\hat{r}}_{i,i+1}^T(m)\underline{\hat{w}}_{i+1}(m)}{\hat{\sigma}_{d,i}^2(m)}} \tag{6.46}$$

6.3.3 Robust Subband Doubletalk Detector Calibration

Since the detection statistic's parameters must be estimated in practice, there will be variability in the resulting value of (6.46). Chapter 4 showed that one can construct statistically optimal doubletalk decisions by comparing the detection statistic to thresholds adaptive to changes in the SNR. This approach is now extended by comparing each subband detection statistic $\xi_i(m)$ to a per-subband adaptive detection threshold. First of all, in Section 4.2 it was shown that the bias effect of background noise $\eta(n)$ in the environment may be compensated by adding an estimate of the noise variance to the numerator of the doubletalk detection statistic. Since background noise is typically not spectrally flat, it makes sense to apply an estimate of the noise variance within each subband to the corresponding detection statistic as follows:

$$\xi_i(m) = \sqrt{\frac{\hat{\underline{r}}_{i,i-1}^T(m)\hat{\underline{w}}_{i-1}(m) + \hat{\underline{r}}_{i,i}^T(m)\hat{\underline{w}}_i(m) + \hat{\underline{r}}_{i,i+1}^T(m)\hat{\underline{w}}_{i+1}(m) + \hat{\sigma}_{\eta,i}^2}{\hat{\sigma}_{d,i}^2(m)}} \approx 1 \qquad (6.47)$$

Section 4.4 presented two algorithms for constructing adaptive doubletalk detection thresholds based on either a desired maximum probability of false alarm (P_F) or probability of miss (P_M) based on (4.2) and (4.3). It is straightforward to adapt these algorithms to provide per-subband adaptive doubletalk detection thresholds, $T_{PF,i}(m)$ and $T_{PM,i}(m)$, respectively, by following the same procedures outlined in Section 4.4:

1. Assuming the background noise is stationary or slowly time-varying, estimate the noise variance in each subband from the reference signal during quiet periods:

$$\hat{\sigma}_{d,i}^2(m) = \frac{1}{K-1}\sum_{k=0}^{K-1}\left[d_i(m-k) - \frac{1}{K}\sum_{j=0}^{K-1} d_i(m-j)\right]^2 \qquad (6.48)$$

$$\hat{\sigma}_{\eta,i}^2(m) = \lambda\hat{\sigma}_{\eta,i}^2(m-1) + (1-\lambda)\hat{\sigma}_{d,i}^2(m) \qquad (6.49)$$

2. Estimate the near-end speech variance within each subband, $\sigma_{y,i}^2(m)$, from the estimated echo signal within each subband:

$$\hat{\sigma}_{\hat{y},i}^2(m) = \frac{1}{K-1}\sum_{k=0}^{K-1}\left[\hat{y}_i(m-k) - \frac{1}{K}\sum_{j=0}^{K-1} \hat{y}_i(m-j)\right]^2 \qquad (6.50)$$

$$\hat{\sigma}_{y,i}^2(m) = \lambda\hat{\sigma}_{y,i}^2(m-1) + (1-\lambda)\hat{\sigma}_{\hat{y},i}^2(m) \qquad (6.51)$$

3. Use the estimated noise and near-end speech variance to construct the probability density functions of the per-subband detection statistics using (4.24) or (4.33), and select the detection threshold giving the desired maximum P_F or P_M:

$$T_{PF,i}(m) = T(m) \text{ such that } P[\xi_i(m) < T(m) \mid H_1] = P_F \rightarrow \int_{\xi_i(m)=0}^{T(m)} f_{\Xi}[\xi_i(m) \mid H_1] d\xi = P_F \quad (6.52)$$

$$T_{PM,i}(m) = T(m) \text{ such that } P[\xi_i(m) > T(m) \mid H_0] = P_M \rightarrow \int_{\xi_i(m)=T(m)}^{\infty} f_{\Xi}[\xi(m) \mid H_0] d\xi = P_M \quad (6.53)$$

As noted in Section 4.4, it is possible to use a single lookup table to provide pre-computed values of $T_{PF,i}(m)$ or $T_{PM,i}(m)$ as a function of the per-subband SNR and a given maximum P_F or P_M.

6.3.4 Simulation Results

6.3.4.1 Simulation Setup

The same M-channel cosine-modulated filter bank used in Section 6.2 was employed for these sets of simulation results [88]. The prototype filter $p_0(n)$ was designed for M = 16 subbands with a stopband attenuation of approximately 100 dB. A plot of the filter's magnitude response is shown in Figure 6.13. The simulated echo path impulse response of length N = 500 samples in Figure 3.2 was employed, and normalized such that $\underline{w}^T\underline{w} = 1$ to equate the echo signal power to the far-end input signal power. Ten sequences of input and near-end speech sequences were obtained from the TIMIT database, with their average powers adjusted to obtain a desired near-end to echo signal power ratio (NER). Stationary, white background noise with an average SNR of 30 dB was added to the reference signal, and assumed to be of known variance $\sigma_{\eta,i}^2$ in each subband. The noise-compensated subband doubletalk detector of (6.47) was evaluated compared to the fullband doubletalk detector of (4.9). In both cases it was assumed that the adaptive filter had converged, and variability in the echo paths was modeled by applying white noise

modulation with variance 0.01. Cross-correlation vectors and reference signal variances were estimated using K = 15 samples for each subband, and K = 200 samples for the fullband. Adaptive detection thresholds for each subband were constructed using (6.53) for $P_F \leq 0.1$ by estimating the near-end speech variance using (6.50) and (6.51) with a smoothing factor of $\lambda = 0.98$.

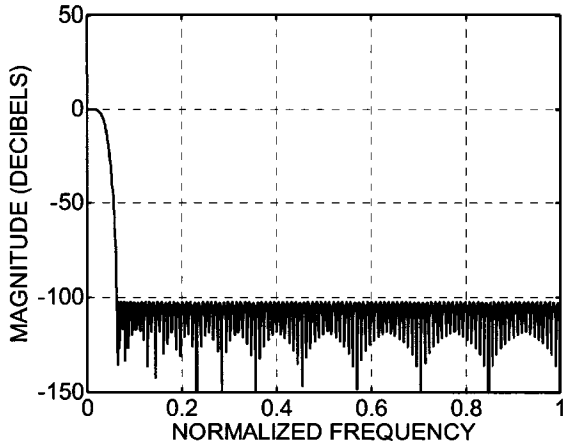


Figure 6.13 - Magnitude response of lowpass prototype filter for M = 16 subbands.

6.3.4.2 Comparison of Subband and Fullband Doubletalk Detector Performance

Figure 6.14 shows the probability of miss (P_M) as a function of the fullband NER averaged over all doubletalk periods for the ten pairs of input and near-end speech sequences. For the subband doubletalk detector, P_M is shown for the detection statistics in the first four subbands. At low-to-moderate levels of near-end speech (NER in the range of -25 to -10 dB), the doubletalk detectors for subbands 2 and 3 (250 – 750 Hz) provide a lower P_M than the fullband doubletalk detector. However, the P_M for subbands 1 and 4 (0 – 250 Hz and 750 Hz – 1 kHz, respectively) are considerably higher than the fullband doubletalk detector. Section 4.3 showed that the expected P_M is dependent on both the SNR of the echo signal, and the relative near-end speech power, which implies

that these marked differences in performance can be explained by examining the relationships between echo, noise, and near-end speech signals in each of the subbands. Figure 6.15(a) shows the ensemble average of the ten echo signal power spectra compared to the power spectrum of the background noise. The ensemble average of the ten near-end speech power spectra is shown in Figure 6.15(b) for NER = 0 dB. Although the SNR of the echo signal is 30 dB on average, the per-subband SNR differs considerably, up to 40 dB in the 500 Hz – 1 kHz bands, down to nearly 20 dB in the 3 – 4 kHz bands. Another explanation for the differences is that the near-end speech power spectrum varies between -2.5 to 7 dB from the echo signal. In practice, the near-end speech power spectrum is generally not known. Therefore, the results of Figure 6.14 and Figure 6.15 suggest that in general, better overall doubletalk detector performance may be obtained by calculating a detection statistic in subband(s) enjoying the highest SNR.

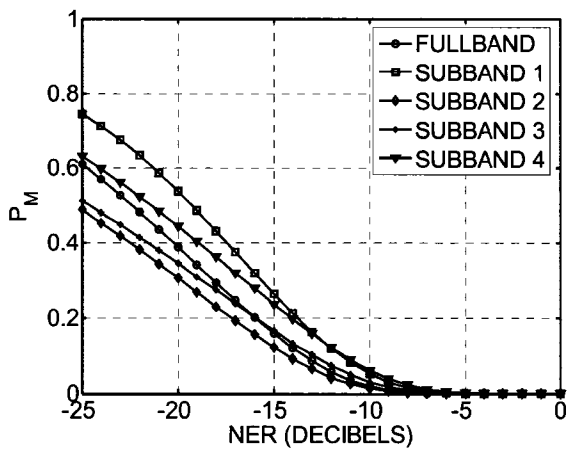


Figure 6.14 – Comparison of fullband and subband doubletalk detection statistics (4 out of 16 subbands) employing adaptive detection thresholds: probability of miss (P_M) as a function of fullband NER for $P_F \leq 0.1$.

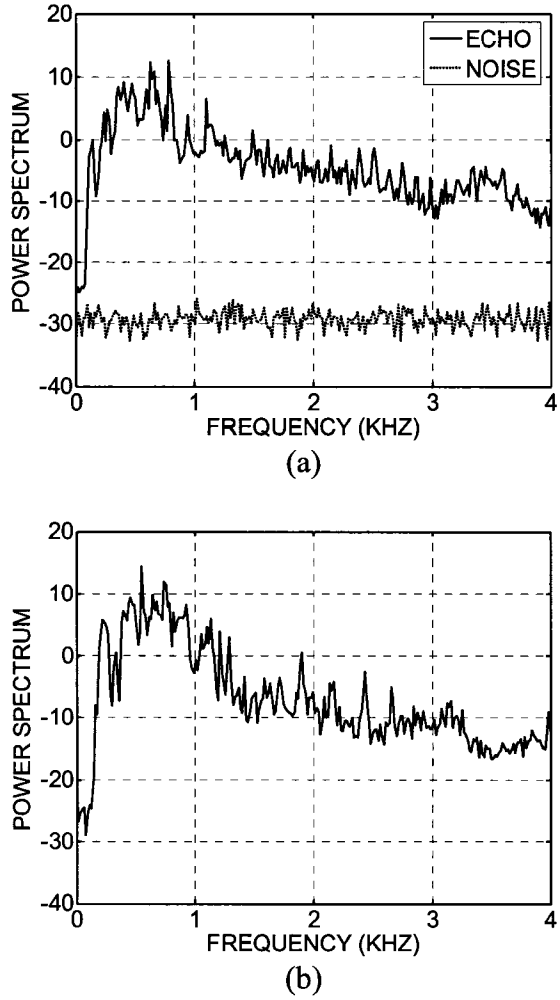


Figure 6.15 – Comparison of echo, background noise, and near-end speech power spectra; (a) ensemble average of echo signal power spectra compared to noise power spectrum; (b) ensemble average of near-end speech power spectra.

6.4 Summary

This chapter presented adaptation and control algorithms for echo cancellers employing the recently proposed critically sampled subband adaptive filter (CS-SBAF) of [68]. Section 6.2 presented a subband AP algorithm suitable for use in the structure. A convergence and computational complexity analysis of the algorithm was presented, which revealed that the convergence rate of each subband is dependent on that of adjacent subbands. This is in contrast with oversampled structures for which the

adaptation of each subband is relatively independent. In addition, it was shown through simulation results that even with filter banks employing a small number of subbands, the subband AP algorithm can provide an improved rate of convergence for correlated signals than fullband AP. Section 6.3 presented a subband normalized cross-correlation-based doubletalk detector algorithm suitable for use in the CS-SBAF structure. It was shown that robust doubletalk detector calibration methods, proposed in Chapter 4, may be easily applied to subband doubletalk detectors. Simulation results comparing the structure with a fullband doubletalk detector showed that per-subband detection statistics can produce a lower probability of miss due to differences in SNR between subbands.

Chapter 7 Echo Canceller Structures for VoIP

7.1 Overview

As reviewed in Chapter 2, a current trend is the delivery of voice services over packet-switched networks, or VoIP. Packet-based networks exacerbate the echo cancellation problem by introducing longer round-trip delays, as well as the possibility of vocoder distortion in echo paths. Typical speech codecs employed in VoIP include ITU-T G.729, G.723.1 (narrowband), and G.722.2 (wideband) [18], [19], [102]. In a typical configuration, speech decoders and encoders are employed at locations where time-domain speech signals are generated or reconstructed. In an IP-based speakerphone, compressed speech frames from the network are decoded into time-domain samples before playback over the loudspeaker. Similarly, at a VoIP / PSTN gateway as shown in Figure 2.9, compressed speech frames are decoded to linear PCM and then to ITU-T G.711 for entry into the PSTN. In both situations, speech signals are reconstructed from a parametric representation, and after processing, are re-compressed into parametric form. A natural question, addressed in this chapter, is whether echo canceller adaptation, doubletalk detection, and post-filtering algorithms can be enhanced, either by increasing performance or reducing computational complexity, by employing information available from LPC-based representations of speech input signals.

The sections of this chapter are organized as follows. Section 7.2 describes a simple NLMS-based adaptation algorithm for echo cancellation, and a normalized cross-correlation-based doubletalk detection algorithm, both of which take advantage of

intermediate signals available from LPC-based speech decoders. Section 7.3 introduces a post-filtering algorithm for suppressing residual echo resulting from vocoder distortion along the echo path. Finally, in Section 7.4 the primary results and conclusions of this chapter are summarized.

7.2 Decorrelated NLMS and Doubletalk Detection Using LPC-Based Speech Parameters

In this section an adaptation algorithm for echo cancellers is presented based on combining a low-bit-rate LPC-based speech decoder and NLMS, reviewed in Sections 2.6.1 and 2.3.2, respectively. A block diagram of the echo canceller structure is shown in Figure 7.1 and briefly described as follows. Compressed speech frames are received at the network interface and used to reconstruct the input signal $x(n)$ in accordance with (2.53) and (2.54). Echo is estimated and subtracted from the reference signal to produce an error signal $e(n)$. However, the adaptation algorithm employs the short-term excitation signal from the speech decoder as an input signal, and preprocesses the error signal with a bank of filters constructed from the current and previous sets of LPC synthesis filter coefficients from the speech decoder. This has the effect of decorrelating the input signal, resulting in an increased rate of convergence in the presence of correlated input signals such as speech. The system is described in more detail in the following subsections.

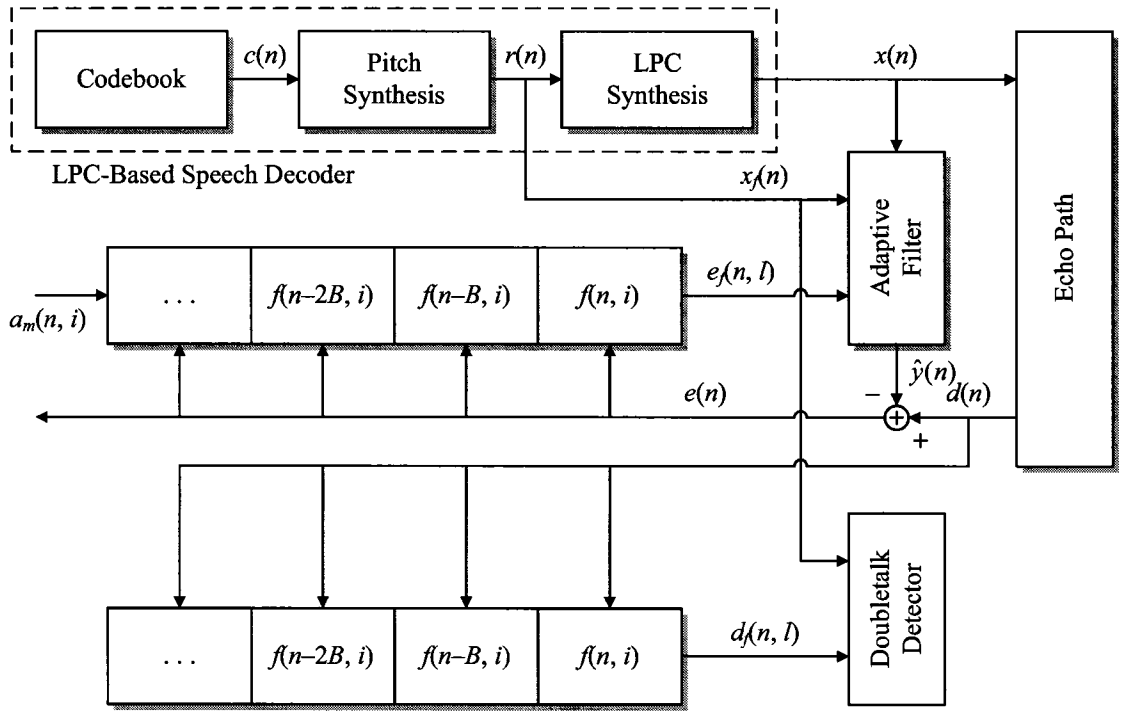


Figure 7.1 - Block diagram of the decorrelated NLMS and doubletalk detection algorithms employing LPC-based speech decoder parameters.

7.2.1 Decorrelated NLMS Algorithm Derivation

The adaptation algorithm presented in this section is based on a reformulated version of the Filtered-X LMS algorithm described in [7] and [91], as shown in Figure 7.1. First of all, assume that the reference signal in Figure 7.1 contains only echo $y(n)$ and background noise $\eta(n)$ signals. Recall from Section 2.6.1 that the input signal $x(n)$ is reconstructed from the excitation signal, pitch period, and LPC parameters from compressed speech frames using (2.53) and (2.54). Assuming an FIR adaptive filter of length N samples, let $\Delta w_j(n) = w_j(n) - \hat{w}_j(n)$ be the adaptive filter error between the jth true and estimated impulse response coefficients at time n, for $0 \leq j \leq N-1$. Combining this with (2.2) gives the error signal as the convolution between the input signal and the adaptive filter error:

$$e(n) = \sum_{j=0}^{N-1} \Delta w_j(n) x(n-j) + \eta(n) \tag{7.1}$$

Substituting (2.54) for $x(n)$ in (7.1) gives an expression of the error signal as a function of the current and previous sets of LPC synthesis filter coefficients over the duration of the target impulse response, and the short-term excitation signal $r(n)$:

$$e(n) = \sum_{j=0}^{N-1} \Delta w_j(n) \sum_{i=1}^{M} a_m(n-j,i) x(n-i-j) + r(n-j) + \eta(n) \tag{7.2}$$

In [91] a time-varying FIR decorrelation filter $f(n,i)$ is applied to the input $x(n)$ and the error $e(n)$ to form filtered input and error signals, respectively:

$$x_f(n) = \sum_{i=0}^{F-1} f(n,i) x(n-i) \tag{7.3}$$

$$e_f(n) = \sum_{i=0}^{F-1} f(n,i) e(n-i) \tag{7.4}$$

where F is the length of $f(n,i)$. These filtered signals are employed in the normalized LMS filter coefficient update instead of the original input and error signals as follows:

$$\hat{w}_l(n+1) = \hat{w}_l(n) + \frac{\mu e_f(n) x_f(n-l)}{\sum_{k=0}^{N-1} x_f^2(n-k) + \delta} \tag{7.5}$$

where l is the adaptive filter coefficient to update, for $0 \le l \le N-1$, and μ is the adaptation step size. Ideally $x_f(n)$ represents a decorrelated version of the input signal, in which case $e_f(n)$ would be proportional to the convolution of the filtered input and the desired impulse response:

$$e_f(n) \approx \sum_{j=0}^{N-1} \Delta w_j(n) x_f(n-j) + \sum_{i=0}^{F-1} f(n,i) \eta(n-i) \tag{7.6}$$

In this case the expected value of the instantaneous gradient estimate in an NLMS adaptation would be directly proportional to the adaptive filter error vector:

$$\begin{aligned} E[e_f(n)x_f(n-l)] &= E\left[\left\{\sum_{j=0}^{N-1}\Delta w_j(n)x_f(n-j)+\sum_{i=0}^{F-1}f(n,i)\eta(n-i)\right\}x_f(n-l)\right] \\ &= E\left[\sum_{j=0}^{N-1}\Delta w_j(n)x_f(n-j)x_f(n-l)\right] \\ &= \sum_{j=0}^{N-1}\Delta w_j(n)E[x_f(n-j)x_f(n-l)] \\ &= \sum_{j=0}^{N-1}\Delta w_j(n)\sigma_{x_f}^2\delta(j-l) \\ &= \Delta w_l(n)\sigma_{x_f}^2 \end{aligned} \tag{7.7}$$

where $\sigma_{x_f}^2$ is the variance of $x_f(n)$, and $0 \le l \le N-1$. This is equivalent to filtering uncorrelated noise, and the convergence rate is no longer dependent upon the eigenvalue spread of the input signal's correlation matrix [44].

From (2.54) it is clear that one possible choice for the decorrelation filter is the inverse of the all-pole LPC synthesis filter at time n:

$$f(n,0)=1; \quad f(n,k)=-a_m(n,k) \quad 1 \le k \le M \tag{7.8}$$

Substituting (2.54) and (7.8) into (7.3) reveals that using the LPC synthesis filter coefficients produces a filtered input signal $x_f(n)$ that is simply $r(n)$, the short-term excitation signal from the speech decoder:

$$\begin{aligned} x_f(n) &= \sum_{i=0}^{F-1}f(n,i)\left[\sum_{j=1}^{M}a_m(n-i,j)x(n-i-j)+r(n-i)\right] \\ &= \sum_{j=1}^{M}[a_m(n,j)-a_m(n,j)]x(n-j)+r(n) \\ &= r(n) \end{aligned} \tag{7.9}$$

Recall that in model LPC-based speech coders, the short-term excitation signal is formed from a pseudorandom codebook signal $c(n)$ and a single-pole pitch synthesis filter [81]. Therefore, it is predicted that $r(n)$ will represent a less correlated version of the speech input signal than $x(n)$, an assumption investigated in Section 7.2.3. Even if $r(n)$ is uncorrelated, to achieve some benefit from using the LPC synthesis filter coefficients as a time-varying decorrelation filter, it is also required that (7.6) holds, or in this case:

$$e_f(n) \approx \sum_{j=0}^{N-1} \Delta w_j(n) r(n-j) + \sum_{i=0}^{F-1} f(n,i)\eta(n-i) \tag{7.10}$$

Expanding the filtered error signal of (7.4) in terms of (7.2) yields the following expression for $e_f(n)$:

$$\begin{aligned} e_f(n) = &\sum_{i=0}^{F-1} f(n,i) \sum_{j=0}^{N-1} \Delta w_j(n-i) \sum_{k=1}^{M} a_m(n-i-j,k)x(n-i-j-k) + r(n-i-j) \\ &+ \sum_{i=0}^{F-1} f(n,i)\eta(n-i) \end{aligned} \tag{7.11}$$

From (7.11) it can be seen that $e_f(n)$ is a function of the current and previous LPC synthesis filter coefficients, which may vary considerably in time for particularly long echo paths such as those common in acoustic echo cancellation ($N \geq 2000$). Practical LPC-based speech coders typically employ a 10^{th}-order LPC model ($M = 10$) and fix the LPC synthesis filter coefficients for the duration of a frame or subframe. Since the decorrelation filter length ($F = M$) is relatively short, then the LPC synthesis filter coefficients can be assumed to be approximately stationary in the short term, or $a_m(n-i-j, k) \approx a_m(n-j, k)$ for $0 \leq i \leq F-1$. Furthermore, if a small adaptation step size is used ($\mu << 1$), then the adaptive filter error vector is approximately steady in the short term, or

$\Delta w_j(n-i) \approx \Delta w_j(n)$ for $0 \le i \le F-1$. Incorporating these approximations into (7.11) allows the expression to be re-organized as follows:

$$\begin{aligned} e_f(n) &= \sum_{j=0}^{N-1} \Delta w_j(n) \sum_{i=0}^{F-1} f(n,i) \sum_{k=1}^{M} a_m(n-j,k)x(n-i-j-k) + r(n-i-j) \\ &\quad + \sum_{i=0}^{F-1} f(n,i)\eta(n-i) \\ &= \sum_{j=0}^{N-1} \Delta w_j(n) \left[x(n-j) - \sum_{i=1}^{M} a_m(n,i) \sum_{k=1}^{M} a_m(n-j,k)x(n-i-j-k) + r(n-i-j) \right] \\ &\quad + \sum_{i=0}^{F-1} f(n,i)\eta(n-i) \end{aligned}$$

(7.12)

Careful examination of (7.12) reveals that if the decorrelation filter of (7.8) is applied, then the approximation of (7.10) will hold only if the LPC synthesis filter coefficients are stationary over the duration of the echo path, or $a_m(n, k) = a_m(n-j, k)$ for $0 \le j \le N-1$. In general this is not the case, and when (7.12) is applied to the NLMS update equation in (7.5), the ideal instantaneous gradient estimate of (7.7) will only approximately hold, and only for lower-order filter coefficients.

Therefore, define a new filtered error signal $e_f(n, l)$ that is a function of both the current time n and the desired filter tap l to be updated in the NLMS coefficient update equation, for $0 \le l \le N-1$. In this case, a bank of decorrelation filters is constructed by applying the inverse of the all-pole LPC synthesis filter employed by the decoder at time $n-l$:

$$f(n,0) = 1; \quad f(n,k) = -a_m(n-l,k) \quad 1 \le k \le M \tag{7.13}$$

Substituting (7.13) into (7.4) gives the new filtered error signal $e_f(n, l)$ for use in the NLMS filter coefficient update equation:

$$e_f(n,l) = \sum_{i=0}^{F-1} f(n-l,i)e(n-i) \tag{7.14}$$

$$\hat{w}_l(n+1) = \hat{w}_l(n) + \frac{\mu e_f(n,l)r(n-l)}{\sum_{k=0}^{N-1} r^2(n-k) + \delta} \tag{7.15}$$

Following the same process and assumptions used to determine (7.12), the filtered error signal of (7.14) can be represented by the following expansion:

$$\begin{aligned} e_f(n,l) = & \sum_{j=0}^{N-1} \Delta w_j(n) \cdot \\ & \left[x(n-j) - \sum_{i=1}^{M} a_m(n-l,i) \sum_{k=1}^{M} a_m(n-j,k)x(n-i-j-k) + r(n-i-j) \right] \\ & + \sum_{i=0}^{F-1} f(n-l,i)\eta(n-i) \end{aligned} \tag{7.16}$$

By filtering the error signal with decorrelation filter coefficients in effect at time $n - l$, the error signal may become more like the ideal case presented in (7.10). It is difficult to model changes in LPC synthesis filter coefficients for time-varying signals such as speech. Another complication from (7.16) is that the background noise signal $\eta(n)$ will be filtered by a bank of decorrelation filters, each potentially contributing a different noise gain in the gradient estimate of (7.15). Therefore, in Section 7.2.4 simulations will be used to investigate the proposed algorithm of (7.14) – (7.15).

It is important to note that the proposed adaptation algorithm differs from conventional decorrelated NLMS in that the presence of LPC synthesis filter coefficients at the speech decoder avoids the need to periodically calculate decorrelation filter coefficients [7]. In addition, a bank of decorrelation filters is employed to produce a coefficient-varying filtered error signal for use in the NLMS coefficient update equation.

7.2.2 Decorrelated Doubletalk Detector Algorithm Derivation

Let $f(n, i)$ again denote the time-varying decorrelation filter of (7.13). Applying $f(n, i)$ to the reference signal yields the following:

$$d_f(n) = \sum_{i=0}^{F-1} f(n,i)d(n-i) \tag{7.17}$$

Ideally $x_f(n)$ is completely decorrelated, resulting in a corresponding autocorrelation matrix that is diagonal and thus easily invertible. In this case, ideally $d_f(n)$ becomes proportional to the convolution of $x_f(n)$ and the room impulse response:

$$d_f(n) \approx \sum_{j=0}^{N-1} w_j(n)x_f(n-j) + \sum_{i=0}^{F-1} f(n,i)[v(n-i) + \eta(n-i)] \tag{7.18}$$

Recall the cross-correlation-based doubletalk detection statistic of Section 2.4, and in particular the time-varying version of (2.42) formed from estimates of the cross-correlation vector and reference signal variance. Assuming that $x_f(n)$ represents a completely decorrelated version of the input signal, and that near-end speech and background noise are uncorrelated with the input signal, then the expected value of the cross-correlation vector between the filtered input and reference signals becomes a function of the filtered input signal variance:

$$\begin{aligned} E[\underline{x}_f(n)d_f(n)] &= E\left[\underline{x}_f(n)\left\{\sum_{j=0}^{N-1} w_j(n)x_f(n-j) + \sum_{i=0}^{F-1} f(n,i)[v(n-i)+\eta(n-i)]\right\}\right] \\ &= E\left[\underline{x}_f(n)\sum_{j=0}^{N-1} w_j(n)x_f(n-j)\right] \\ &= \sum_{j=0}^{N-1} w_j(n)E[\underline{x}_f(n)x_f(n-j)] \\ &= \sigma_{x_f}^2 \sum_{j=0}^{N-1} w_j(n)\delta(j) \\ &= \sigma_{x_f}^2 \underline{w}(n) \end{aligned} \tag{7.19}$$

Assuming the adaptive filter has converged, replacing the input and reference signals in (2.42) with the filtered input and reference signals of (7.3) and (7.18) yields a simplified doubletalk detection statistic as follows:

$$\xi(n) = \sqrt{\frac{\hat{\underline{r}}_{x_f d_f}^T(n)\hat{\underline{w}}(n)}{\hat{\sigma}_{d_f}^2(n)}} = \sqrt{\frac{[\sigma_{x_f}^2 \underline{w}(n)]^T \hat{\underline{w}}(n)}{\hat{\sigma}_{d_f}^2(n)}} \approx \sqrt{\frac{\hat{\sigma}_{x_f}^2(n)\hat{\underline{w}}^T(n)\hat{\underline{w}}(n)}{\hat{\sigma}_{d_f}^2(n)}} \tag{7.20}$$

Equation (7.20) represents a simplified version of the detection statistic of (2.42) in that the estimate of the cross-correlation vector has been replaced with an estimate of the filtered input signal variance, $\sigma_{x_f}^2$. As in (2.40), the filtered input and reference signal variances can both be estimated over a window of K samples:

$$\hat{\sigma}_{x_f}^2(n) = \frac{1}{K-1}\sum_{k=0}^{K-1}\left[x_f(n-k) - \frac{1}{K}\sum_{j=0}^{K-1}x_f(n-j)\right]^2 \tag{7.21}$$

$$\hat{\sigma}_{d_f}^2(n) = \frac{1}{K-1}\sum_{k=0}^{K-1}\left[d_f(n-k) - \frac{1}{K}\sum_{j=0}^{K-1}d_f(n-j)\right]^2 \tag{7.22}$$

7.2.3 Computational Complexity

If the short term excitation signal $r(n)$ can be extracted directly from the speech decoder, then no computation is required to compute $x_f(n)$ from the input signal. Calculating $e_f(n,l)$ would require NF multiplications per sample if the LPC synthesis filter coefficients were time varying in n. In practice the synthesis filter coefficients are constant for the duration of a subframe and interpolated within frames of B samples, so an approximation is to use the per-frame LPC coefficients to calculate $e_f(n, l)$ in B-sample blocks. In this case only $ceil(N / B)$ filters are necessary to calculate $e_f(n, l)$ per sample period, each with a cost of F multiplications. For example, ITU-T G.729 employs 80-

sample frames and $F = M = 10$. For an echo path of length $N = 2000$ samples, the decorrelated NLMS algorithm requires $(2000 / 80) \times 10 = 250$ multiplications per sample to calculate $e_f(n, l)$, increasing the NLMS complexity by approximately six percent. After decorrelation, the doubletalk detection statistic of (7.20) requires approximately $2K + N$ multiplications per sample, whereas the original detection statistic of (2.42) requires $(N+1)K + N$ multiplications per sample.

7.2.4 Simulation Results

7.2.4.1 Simulation Setup

The proposed algorithms were implemented using the LPC-based speech coder defined in ITU-T G.729A [18]. The reference code was modified to extract the short-term excitation signal $r(n)$ and LPC synthesis filter coefficients $a_m(n, i)$ from the decoder, and the speech signal $x(n)$ was reconstructed with post-filtering disabled. Tests were conducted using the echo path impulse responses shown in Figure 3.4 consisting of $N = 2000$ samples (250 ms). The performance of the decorrelated NLMS algorithm was compared to NLMS and to decorrelated NLMS employing a fifth-order decorrelation filter [7]. A step size of $\mu = 0.1$ was employed for all algorithms, and performance measured using ERLE and system distance. Two test signals were employed, a stationary 10^{th}-order autoregressive process driven by white Gaussian noise, and concatenated continuous speech sequences from the TIMIT database downsampled to 8 kHz [92]. Both signals were compressed using the reference ITU-T G.729A encoder, and white noise was added to the resulting echo signals to produce an average SNR of 40 dB in the reference signal. The performance of the doubletalk detection algorithm was

compared to a full-complexity implementation using an adaptive decorrelation filter $f(n, i)$ constructed for each frame using the Levinson-Durbin algorithm and autocorrelation function estimated over a window centered on the current frame with 50 percent overlap.

7.2.4.2 Statistics of Short-Term Excitation Signals

Section 7.2.2 assumed that the short-term excitation signal from the decoder, $r(n)$, is significantly less correlated than the reconstructed signal $x(n)$. This assumption was investigated by examining the condition number of the autocorrelation matrices for $r(n)$ and $x(n)$, which for a Toeplitz matrix is the ratio of maximum to minimum eigenvalues [44]. It is known that for a diagonal autocorrelation matrix, corresponding to a completely uncorrelated signal, the condition number is one. The autocorrelation was estimated for $M = 10$ lags over frames of 160 samples (20 ms at $f_s = 8$ kHz) over four seconds. Figure 7.2(a) and Figure 7.2(b) show the per-frame condition numbers for the reconstructed AR(10) and speech signals, respectively, along with the those for the corresponding short-term excitation signals from the ITU G.729A decoder. The condition numbers for the excitation signals are low for both test signals, averaging 2.88 for the AR(10) signal and 9.96 for the speech signal. In contrast, the autocorrelation matrices for the reconstructed AR(10) and speech signals have much higher condition numbers, which suggests that the excitation signal $r(n)$ represents a good choice for use as a filtered input signal.

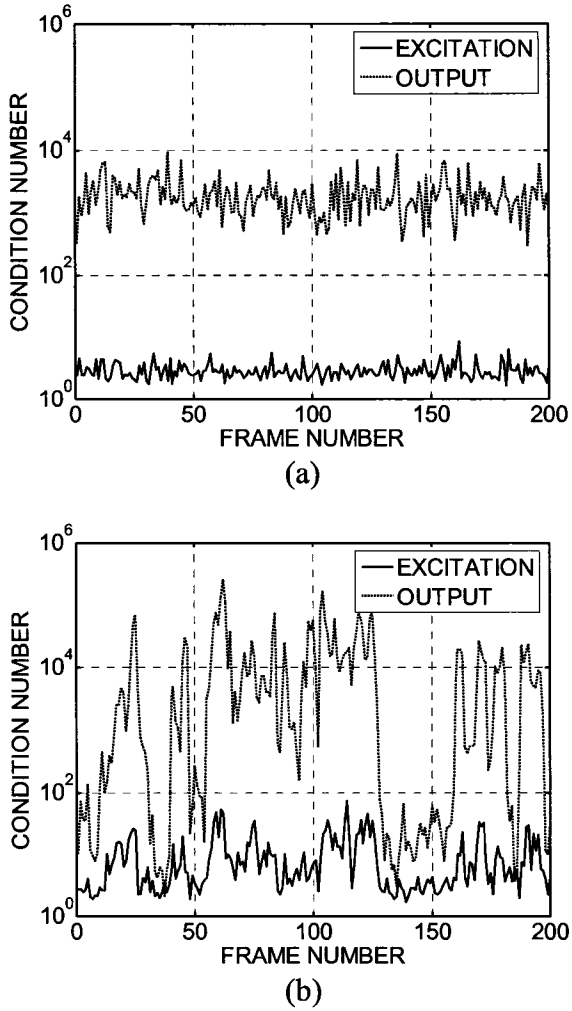


Figure 7.2 – Autocorrelation matrix condition number for excitation and output signals from ITU-G.729A decoder, calculated over 20 ms frames with $M = 10$ lags, for (a) stationary AR(10) and (b) speech input signals.

7.2.4.3 Decorrelated NLMS Convergence and Tracking Performance

Figure 7.3(a) and Figure 7.3(b) show ERLE and system distance during initial convergence using the echo path impulse response of Figure 3.4(a) for the AR(10) input signal. Similarly, Figure 7.4(a) and Figure 7.4(b) show the same measurements for the speech input signal. For both the AR(10) and speech input signals, it is clear that the proposed LPC-based decorrelated NLMS produces a faster and more constant rate of convergence with respect to system distance, which is in agreement with the theoretical

performance shown in (7.6). For the AR(10) input signal, after six seconds of adaptation the proposed algorithm achieves a steady-state ERLE approximately 20 dB higher than NLMS, and 10 dB higher than NLMS with fifth-order decorrelation. For speech input the ERLE was approximately 7.5 dB higher than both NLMS and decorrelated NLMS averaged over the 12-second signal duration. Figure 7.5(a) and Figure 7.5(b) show ERLE and system distance for speech input signals after switching from the echo path impulse response of Figure 3.4(a) to the one in Figure 3.4(b) after ten seconds. It is clear that the depth and rate of convergence of the proposed algorithm is similar during both initial convergence and after the change in echo path. In addition, the improvement in ERLE over NLMS remains approximately 7.5 dB on average during re-convergence.

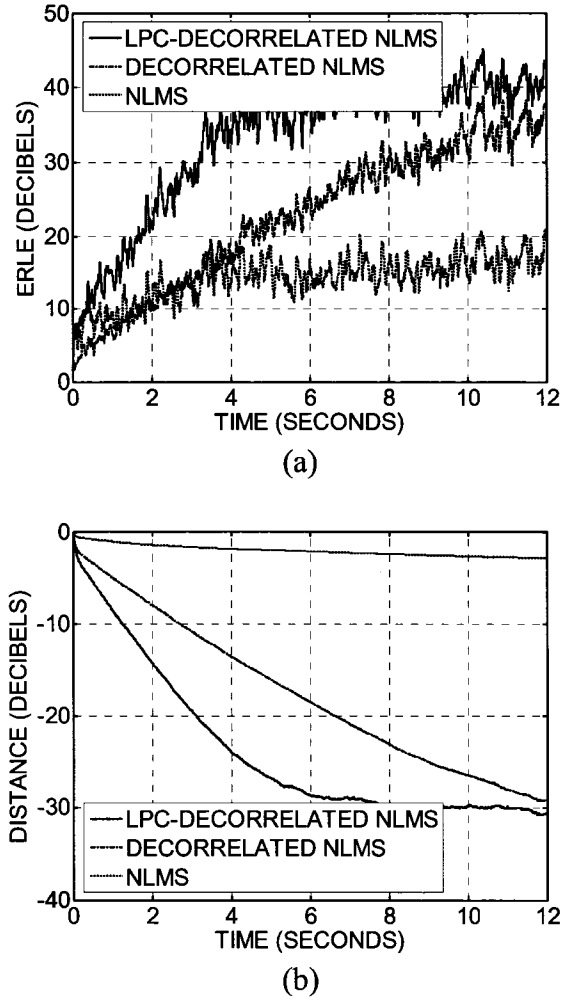


Figure 7.3 – Convergence rate of LPC-based decorrelated NLMS algorithm for AR(10) input signal, compared to NLMS and NLMS with fifth-order decorrelation; (a) ERLE and (b) system distance.

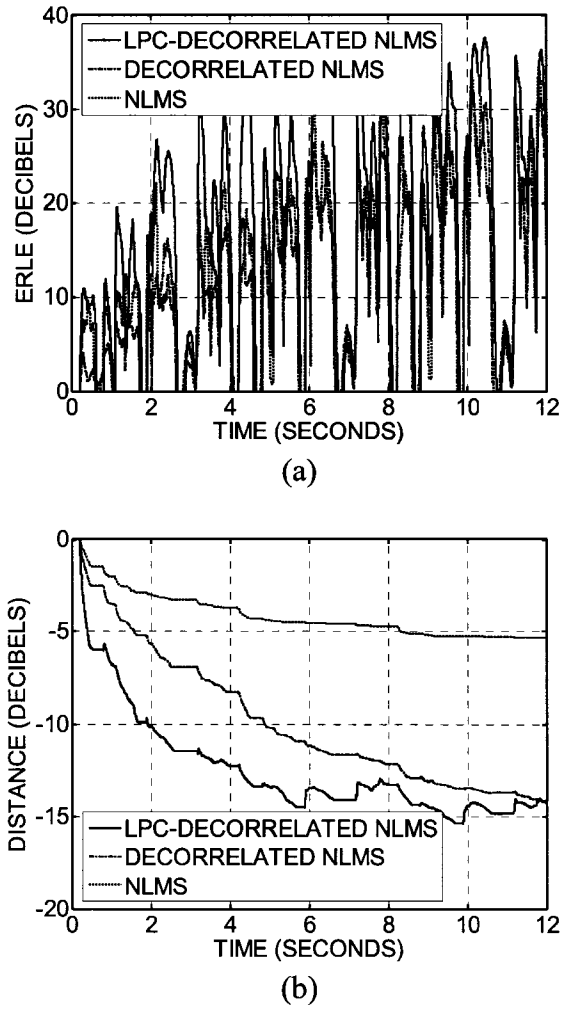


Figure 7.4 - Convergence rate of LPC-based decorrelated NLMS algorithm for speech input signal, compared to NLMS and NLMS with fifth-order decorrelation; (a) ERLE and (b) system distance.

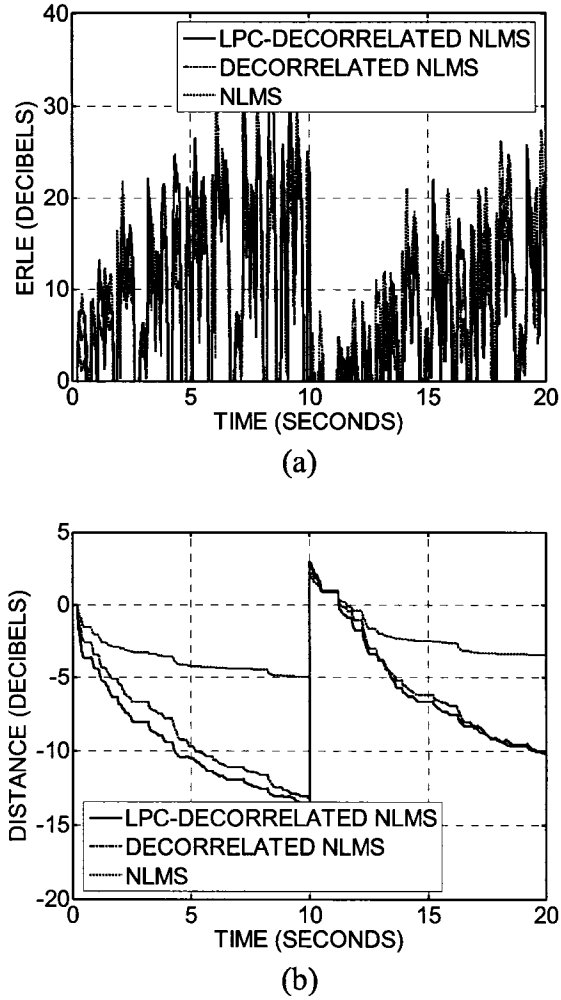


Figure 7.5 – Tracking performance of LPC-based decorrelated NLMS for speech input signal with echo path change after 10 seconds, compared to NLMS and NLMS with fifth-order decorrelation; (a) ERLE and (b) system distance.

7.2.4.4 Effect of Decorrelation Order on Doubletalk Detection Probability

Figure 7.6(a) shows a plot of the probability of miss (P_M) as a function of NER for the proposed algorithm, and for a full-complexity implementation with decorrelation filter orders of L = 1, 2, 5, and 10. The results were obtained by averaging doubletalk tests over ten sets of input and near-end speech signals from the TIMIT database for $P_F \leq 0.1$ with K = 200 samples. The proposed low-complexity algorithm has the same doubletalk detection probability as the full-complexity implementation for the same decorrelation

filter order (F = 10). Low-order decorrelation filters result in a degraded detection probability, likely because they cannot sufficiently whiten $x(n)$ and the assumptions of (7.10) do not hold. Figure 7.6(b) shows the probability of miss for the proposed algorithm compared to the full implementation of (2.42) as a function of NER for K = 200 samples. At lower NER the miss probability is higher for the proposed algorithm, which suggests that there is higher variance in parameter estimates.

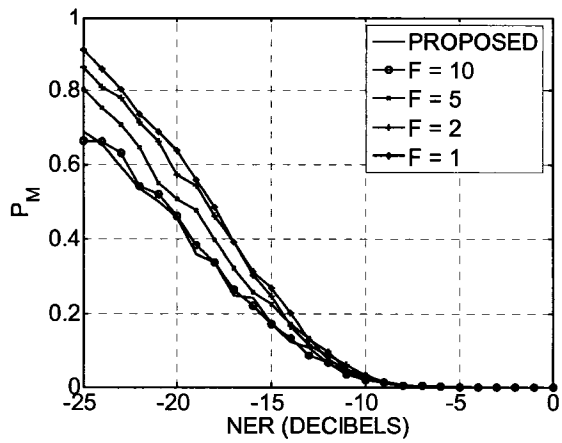


Figure 7.6 - Probability of miss (P_M) as a function of NER using K = 200 samples, $P_F \leq 0.1$ for LPC-based decorrelated doubletalk detector compared to (a) doubletalk detector using decorrelation filters of length F = 1, 2, 5 and 10 samples, and (b) full-complexity doubletalk detector.

7.3 Post-Filtering in the Presence of Vocoder Distortion

One problem in VoIP and mobile networks is the potential for nonlinear vocoder distortion in the echo path, a problem reviewed in Section 2.6. Frequency-domain post-filtering algorithms are capable of suppressing residual echo to enhance near-end speech, but a key problem is obtaining estimates of the residual echo power spectrum. In this section, the effects of vocoder distortion on echo canceller convergence are investigated, along with the power spectrum properties of the resulting residual echo signal. A method is presented for estimating the residual echo power spectrum arising from nonlinear vocoder distortion, and compared with existing approaches in terms of near-end speech quality after post-filtering.

7.3.1 Effect of Vocoder Distortion in the Echo Path

Low-bit-rate speech encoders introduce nonlinear distortion into speech signals through quantization of LPC coefficients, pitch period and gains, and excitation signal parameters [18], [19], [21]. In addition, decoders may employ a harmonic post-filter stage which introduces further nonlinearity. In the configuration of Figure 2.12, vocoders introduce distortion into the input signal $x(n)$ and reference signal $d(n)$ entering the network. Assuming negligible background noise and, optionally, the presence of near-end speech $v(n)$, the reconstructed signals can be written as the sum of the original and nonlinear distortion signals:

$$\tilde{x}(n) = x(n) + x_{NL}(n) \tag{7.23}$$

$$\tilde{d}(n) = d(n) + d_{NL}(n) = [x(n) + x_{NL}(n)] \otimes w(n) + v(n) + d_{NL}(n) \tag{7.24}$$

where $x_{NL}(n)$ and $d_{NL}(n)$ represent nonlinear distortion in the input and reference signals, respectively, and $\otimes$ denotes convolution. From (2.2) and assuming a linear echo path, the error and residual echo signals can be represented as follows:

$$e(n) = x(n) \otimes [w(n) - \hat{w}(n)] + x_{NL}(n) \otimes w(n) + v(n) + d_{NL}(n) = v(n) + \delta(n) \quad (7.25)$$

$$\delta(n) = x(n) \otimes [w(n) - \hat{w}(n)] + x_{NL}(n) \otimes w(n) + d_{NL}(n) \quad (7.26)$$

In [105] it was proposed to model the nonlinear distortion signals as autoregressive processes independent of the original speech signals. If no near-end speech is present in (7.25) and the adaptive filter is allowed to adapt, the signal distortions $x_{NL}(n)$ and $d_{NL}(n)$ introduce noise components into the error signal that are uncorrelated with the original input signal used for adaptation. Figure 7.7 shows ERLE and system distance during initial convergence for a stationary AR(10) input signal applied to the configuration of Figure 2.12 under three conditions: no vocoder distortion, an ITU-T G.729A encoder / decoder pair affecting the input signal only, and vocoders affecting both the input and reference signals. Figure 7.8 shows the same performance measures for a speech input signal of 4 seconds duration. Both simulations employed the echo path of Figure 3.3 (N = 500 samples) with background noise of 40 dB SNR and no near-end speech, and NLMS with a step size of μ = 0.1. It is clear that the presence of vocoder distortion severely degrades the echo canceller performance in both cases, producing an average steady-state ERLE of 6.5 dB for the AR(10) signal with a vocoder pair along the send path, dropping to 5.4 dB when the receive path vocoder pair is introduced. Some improvement was observed for the speech input signal, achieving a maximum ERLE of 10-15 dB, dropping by 3-4 db with the addition of the receive path vocoder pair.

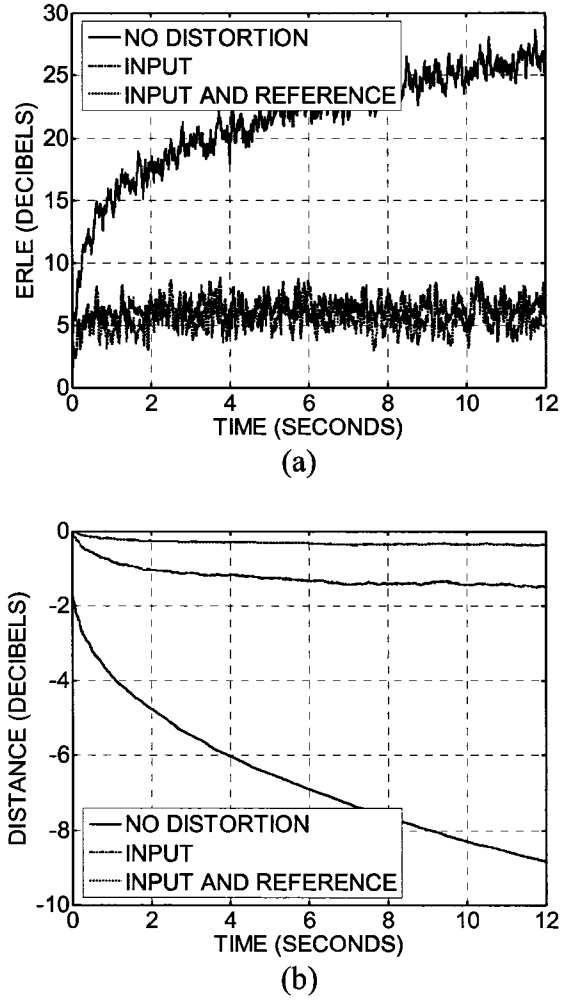


Figure 7.7 – Convergence performance of NLMS with echo path distortion from ITU-T G.729A for AR(10) input signal with no distortion, vocoder pair on the input signal, and on both input and reference signals; (a) ERLE and (b) system distance.

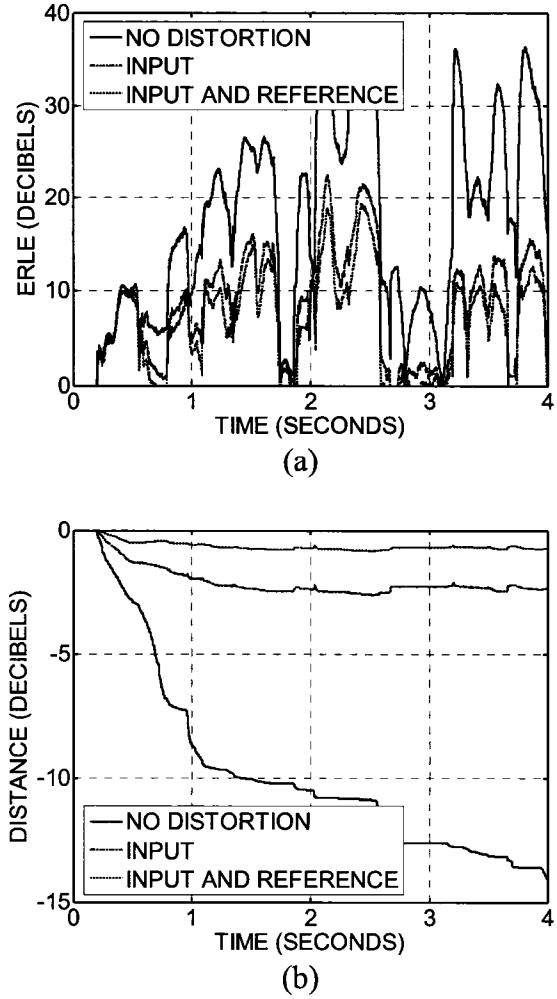


Figure 7.8 - Convergence performance of NLMS with echo path distortion from ITU-T G.729A for speech input signal with no distortion, vocoder pair on the input signal, and on both input and reference signals; (a) ERLE and (b) system distance.

7.3.2 Residual Echo Power Spectrum Estimation

Even if the adaptive filter were able to converge completely, or $\underline{\hat{w}}(n) \approx \underline{w}(n)$, from (7.26) it is clear that the residual echo signal will possess spectral properties similar to that of the input signal. To effectively employ post-filtering methods to further suppress the residual echo, it is necessary to estimate the power spectrum for the residual echo signal of (7.26), a task complicated by the encoder / decoder complexity. To investigate the residual echo properties, signals resulting from the simulations of Figure 7.8 were

analyzed further. Figure 7.9(a) and Figure 7.9(b) show the estimated power spectra of each of the three terms in (7.26) for voiced and unvoiced speech segments, respectively. The power spectrum of the first term, the "true" residual echo produced by the echo canceller, is approximately 10-15 dB below that of the second and third terms. Even with the low suppression achieved by the echo canceller, the residual echo power spectrum appears to be dominated by the nonlinear distortion of the input and reference signals:

$$\delta(n) \approx x_{NL}(n) \otimes w(n) + d_{NL}(n) \tag{7.27}$$

In [105] reconstructed speech signals were modeled as purely autoregressive processes, with vocoder distortion introduced as white quantization noise in the excitation signal. Let σ_X^2, $\sigma_{\Delta X}^2$, and $A_X(\omega)$ represent the short-term excitation signal power, quantization noise power, and LPC synthesis filter coefficients for the input signal $x(n)$, and let σ_D^2, $\sigma_{\Delta D}^2$, and $A_D(\omega)$ be similar variables for the reference signal $d(n)$. Following the approach in [105], the power spectra of (7.23) and (7.24) can be modeled with the original signal power spectrum and the signal to quantization noise ratio of the reconstructed input signal (SNR_X) and reference signal (SNR_D):

$$S_{\tilde{X}\tilde{X}}(\omega) \approx \frac{\sigma_X^2 + \sigma_{\Delta X}^2}{|A_X(\omega)|^2} = S_{XX}(\omega) + \frac{\sigma_{\Delta X}^2}{|A_X(\omega)|^2} = S_{XX}(\omega)\left(\frac{\mathrm{SNR_X} + 1}{\mathrm{SNR_X}}\right) \tag{7.28}$$

$$S_{\tilde{D}\tilde{D}}(\omega) \approx \frac{\sigma_D^2 + \sigma_{\Delta D}^2}{|A_D(\omega)|^2} = S_{DD}(\omega) + \frac{\sigma_{\Delta D}^2}{|A_D(\omega)|^2} = S_{DD}(\omega)\left(\frac{\mathrm{SNR_D} + 1}{\mathrm{SNR_D}}\right) \tag{7.29}$$

Following a similar approach, the power spectrum of the residual echo in (7.27) can be approximated by the power spectra of the input and reference signals:

$$S_{\Delta\Delta}(\omega) \approx \frac{S_{XX}(\omega)|W(\omega)|^2}{\mathrm{SNR_X}} + \frac{S_{DD}(\omega)}{\mathrm{SNR_D}} \tag{7.30}$$

where $W(\omega)$ is the frequency response of the echo path. From (7.24), the reference signal power spectrum $S_{DD}(\omega)$ consists of the echo signal and the near-end speech, which are assumed to be uncorrelated with each other:

$$S_{DD}(\omega) = S_{\tilde{X}\tilde{X}}(\omega)|W(\omega)|^2 + S_{VV}(\omega) \approx S_{XX}(\omega)|W(\omega)|^2 \left(\frac{\mathrm{SNR_X}+1}{\mathrm{SNR_X}}\right) + S_{VV}(\omega) \quad (7.31)$$

Equations (7.30) and (7.31) show that the reference signal distortion is a function of both the echo and the near-end speech power spectra. However, the near-end speech distortion will have similar spectral properties to the actual near-end speech, and may be omitted as part of the residual echo to be suppressed. This results in a simplified approximation for the residual echo power spectrum:

$$S_{\Delta\Delta}(\omega) \approx S_{XX}(\omega)|W(\omega)|^2 \left(\frac{\mathrm{SNR_X}+\mathrm{SNR_D}+1}{\mathrm{SNR_X}\mathrm{SNR_D}}\right) \quad (7.32)$$

In [105] it was assumed that the signal to quantization noise ratio produced by LPC-based speech encoder / decoder pairs is uniform across all frequencies. However, encoders minimize a weighted error function that allows greater error in high-energy (formant) regions corresponding to peaks in the LPC spectrum, and lower error in non-formant regions [81]. The weighting filter $W_P(z)$ is constructed from the LPC spectrum with:

$$W_P(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} \quad (7.33)$$

where $A(z)$ is the z-transform of the LPC spectrum and γ_1 and γ_2 are parameters controlling the weighting, with typical values of $\gamma_1 = 1$ and $\gamma_2 = 0.9$ employed by ITU-T G.729A [18]. As a result, it is expected that the SNR of the reconstructed input signal

(7.23) at the near end will be lower at peak frequencies of the input signal LPC spectrum $A_X(\omega)$. Similarly, the SNR of the reconstructed reference signal (7.24) is expected to be lower at peaks in $A_D(\omega)$.

One proposed way method of incorporating this weighting function is to weight the average SNR by normalized versions of the weighting functions in the frequency domain. Let $W_X(\omega)$ and $W_D(\omega)$ be weighting functions corresponding to the input and reference signal LPC spectra, respectively. Define $\mathrm{SNR_X}(\omega)$ and $\mathrm{SNR_D}(\omega)$ as frequency-dependent weighted estimates of the signal to quantization noise ratio as follows:

$$\mathrm{SNR_X}(\omega) \approx \mathrm{SNR}\frac{|W_X(\omega)|^2}{\frac{1}{2\pi}\int_{\omega=0}^{2\pi}|W_X(\omega)|^2\,d\omega} \tag{7.34}$$

$$\mathrm{SNR_D}(\omega) \approx \mathrm{SNR}\frac{|W_D(\omega)|^2}{\frac{1}{2\pi}\int_{\omega=0}^{2\pi}|W_D(\omega)|^2\,d\omega} \tag{7.35}$$

Substituting the two weighted SNR estimates into (7.32) results in the following expression for the estimated residual echo signal due to vocoder distortion in the input and reference signals:

$$S_{\Delta\Delta}(\omega) \approx S_{XX}(\omega)\,|W(\omega)|^2\left[\frac{\mathrm{SNR_X}(\omega)+\mathrm{SNR_D}(\omega)+1}{\mathrm{SNR_X}(\omega)\mathrm{SNR_D}(\omega)}\right] \tag{7.36}$$

Both the input signal power spectrum $S_{XX}(\omega)$ and LPC synthesis filter $A_X(\omega)$ can be estimated from $x(n)$. If the adaptive filter has converged as much as possible, then the residual echo from the input signal distortion can be approximated using the estimated echo produced by the echo canceller. If the near-end speech $v(n)$ is significantly higher

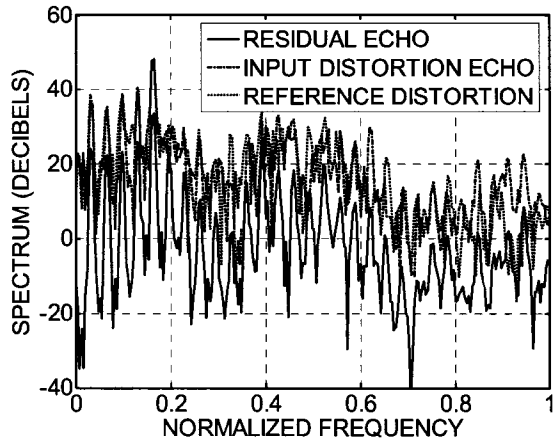
than the echo, then the LPC synthesis filter used to reconstruct the reference signal will be dominated by the near-end speech spectrum, or $A_D(\omega) \approx A_V(\omega)$. Applying these assumptions results in a simplified approximation for the residual echo power spectrum:

$$S_{\Delta\Delta}(\omega) \approx S_{XX}(\omega)\,|\hat{W}(\omega)|^2 \left[\frac{\mathrm{SNR}_X(\omega) + \mathrm{SNR}_V(\omega) + 1}{\mathrm{SNR}_X(\omega)\mathrm{SNR}_V(\omega)}\right] \tag{7.37}$$

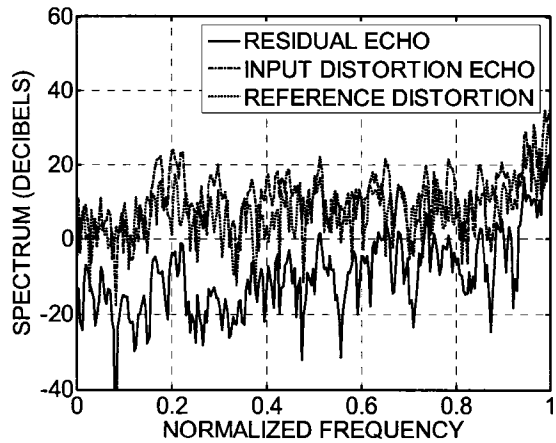
Since the near-end speech is assumed to be higher than the residual echo, an estimate of the near-end speech may be obtained via spectral subtraction using the power spectrum of the estimated echo produced by the echo canceller output:

$$S_{VV}(\omega) \approx S_{EE}(\omega) - S_{\hat{Y}\hat{Y}}(\omega) \tag{7.38}$$

Finally, it should be noted that although post-filtering is generally used during doubletalk periods to enhance near-end speech, during singletalk periods it may be possible to employ post-filtering to further suppress residual echo by employing estimates of the background noise power spectrum $S_{NN}(\omega)$ as the near-end speech estimate, or $S_{VV}(\omega) \approx S_{NN}(\omega)$.



(a)



(b)

Figure 7.9 - Power spectra of the three components of (7.26): "true" residual echo, echo produced by input signal distortion, and reference signal distortion, for (a) voiced and (b) unvoiced speech.

7.3.3 Simulation Results

7.3.3.1 Simulation Setup

Experimental results of the proposed residual echo power spectrum estimation method were obtained by incorporating it into the psychoacoustic post-filter of (2.52) constructed using the preliminary near-end speech power spectrum estimate of (7.38). For comparison, the method of [105] was implemented by employing simply the average signal to quantization to noise ratio for both the input and reference signals, or $SNR_X(\omega)$

= $SNR_D(\omega)$ = SNR. For ITU-T G.729A, the average SNR was measured to be approximately 12 dB. The analysis and synthesis stages were implemented using the FFT on 256-sample blocks with 50 percent overlap and zero-padded to 512 samples. The masking threshold for (2.52) was calculated using the MPEG-1 Psychoacoustic Model 1 modified for a sampling rate of 8 kHz [106]. Pairs of input and near-end speech signals were obtained from the TIMIT database [92]. Fixed adaptive filter coefficients were obtained from Figure 7.8 corresponding to vocoder distortion in both the send and receive paths, and near-end speech power was adjusted to an average of 10 dB NER. The power spectrum estimation algorithm was evaluated by measuring the spectral distortion between the original and estimated near-end speech signals:

$$SD^2 = \frac{1}{2\pi} \int_{\omega=0}^{2\pi} [10 \log_{10} S_{VV}(\omega) - 10 \log_{10} S_{\hat{V}\hat{V}}(\omega)]^2 \, d\omega \tag{7.39}$$

where $S_{\hat{V}\hat{V}}(\omega)$ is the estimated near-end speech power spectrum for the current block after applying the post-filter. The quality of the estimated near-end speech was also evaluated using the mean opinion score estimate provided by ITU-T P.862 [42]. To find the maximum improvement possible with post-filtering, also considered was the (ideal) case where the power spectrum of the residual echo signal is known.

7.3.3.2 Results and Discussion

Table 7.1 shows the average spectral distortion and estimated mean opinion score (MOS) for near-end speech after post-filtering with the three test configurations. The proposed method results in an average 0.98 dB lower spectral distortion and an average 0.5 higher estimated MOS, which is clearly an improvement over the fixed scaling factor.

As an example of the improvement afforded by post-filtering, Figure 7.10(a) and Figure 7.10(b) show spectrograms of the original near-end speech signal, and the echo canceller error signal containing residual echo caused by vocoders along both the send and receive paths. Figure 7.11(a) and Figure 7.11(b) show the post-filtered error signal using residual echo power spectrum estimates constructed using fixed and adaptive signal to quantization noise ratios, respectively. Low-frequency residual echo can still be observed between 2 and 2.5 seconds with the fixed SNR, which is almost completely removed by the configuration employing signal-adaptive SNR estimates.

Figure 7.12 shows ERLE measured during singletalk conditions (only far-end speech), again with vocoder distortion present in both the send and receive paths. In this example, the psychoacoustic post-filter of (2.52) was again applied using the adaptive and fixed SNR residual echo estimates, compared to ERLE obtained with no post-filtering. From the figure it is clear that applying post-filtering improves the ERLE by at least 5 – 10 dB, with an additional 2 – 3 dB improvement using the signal-adaptive quantization noise estimates.

Table 7.1 - Average spectral distortion and estimated MOS of near-end speech enhanced with psychoacoustic post-filtering using residual echo power spectrum estimates produced using signal-adaptive and fixed SNR estimates, compared to the ideal case with known residual echo

Test Pair	Proposed Adaptive SNR		Fixed SNR		Ideal	
	SD	MOS	SD	MOS	SD	MOS
1	2.21	3.39	2.86	3.05	1.22	3.96
2	3.34	2.95	4.63	2.21	2.11	3.17
3	3.49	2.89	4.60	2.45	2.24	3.23
4	2.68	3.27	3.25	3.05	1.70	3.59
5	2.87	2.96	3.82	2.58	1.79	3.27
6	3.17	2.71	4.15	2.29	1.97	3.15
7	2.26	3.25	3.04	3.10	1.38	3.47
8	3.74	3.19	3.97	2.86	2.29	3.38
9	2.75	2.91	3.95	2.51	1.83	3.29
10	2.42	3.18	3.47	2.86	1.74	3.45

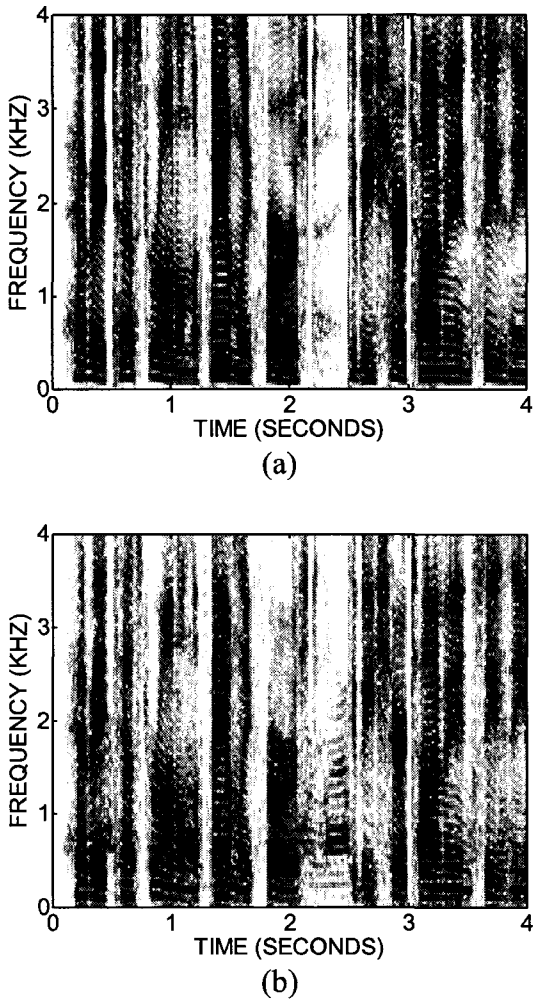


Figure 7.10 – Example of post-filtering effects: (a) original near-end speech spectrogram; (b) echo canceller error signal containing near-end speech and residual echo due to vocoder distortion.

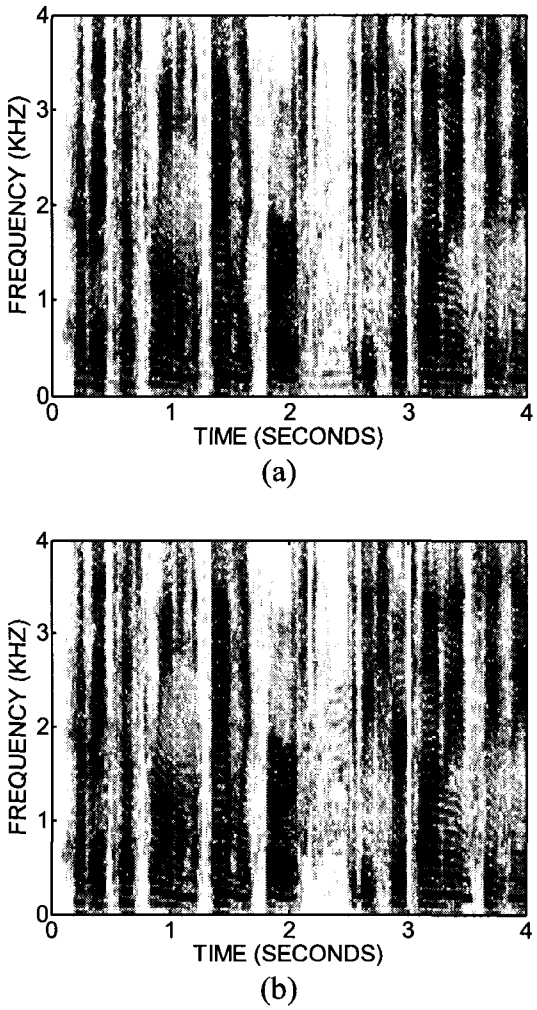


Figure 7.11 – Example of post-filtering effects (cont.): near-end speech spectrogram after psychoacoustic post-filtering with (a) fixed and (b) frequency-dependent SNR estimates.

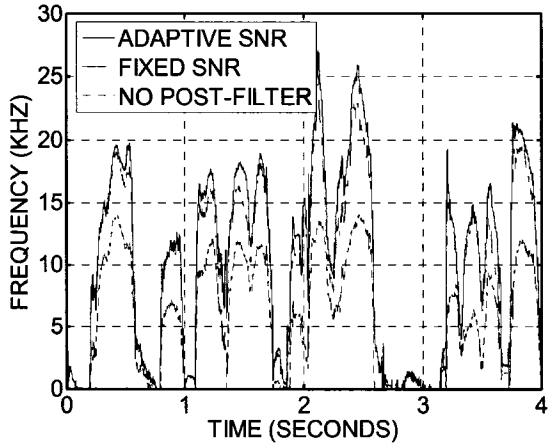


Figure 7.12 – ERLE for speech input signal during singletalk conditions with no post-filtering, compared to post-filtering using residual echo estimated using fixed and signal-adaptive SNR.

7.4 Summary

This chapter presented adaptation and control algorithms for echo cancellers that are focused on some of the unique problems inherent in VoIP networks. Section 7.2 presented low-complexity decorrelated NLMS and cross-correlation-based doubletalk detector algorithms for use at VoIP gateways and in IP-enabled phones. A typical implementation of such algorithms would require constructing decorrelation filters using, for example, the Levinson-Durbin algorithm and estimates of the autocorrelation function. The proposed algorithms employ knowledge of the speech signals contained within the compressed speech frames available at the decoder, achieving the same result but at a considerable reduction in complexity. Section 7.3 addressed the problem of residual echo after echo cancellation due to the presence of nonlinear vocoder distortion in the echo path. A method was presented for estimating the residual echo power spectrum for applying post-filtering to enhance near-end speech. The proposed method improves upon previous residual echo models from the literature by incorporating the fact that the speech encoding process allows higher quantization error at spectral peaks, which

results in higher residual echo at those frequencies. Simulation results confirmed that the proposed method provides an improvement in near-end speech quality.

Chapter 8 Conclusions and Future Research

8.1 Summary of Research

The primary objective of this thesis was the investigation of structures and adaptation algorithms capable of operating in harsh environments, with specific application to those relevant to the echo cancellation problem: doubletalk conditions, background noise, and the technologies introduced by VoIP networks. A secondary objective was maintaining a low computational complexity in the resulting algorithms. Five relevant research questions were identified in Chapter 1:

1. What is the behavior of doubletalk detectors in the presence of background noise and time-varying signal statistics, and how can an appropriate detection threshold be constructed given these influences?
2. What are the perceived limits of echo canceller performance given psychoacoustic aspects of human hearing, and how does the presence of background noise affect perceived echo canceller performance?
3. Given that additional structures are required to reduce aliasing in critically sampled subband filters, is it possible to incorporate more sophisticated adaptation and doubletalk detection algorithms into echo cancellers based on these structures?

4. Is it possible to incorporate parametric representations of speech, utilized by low-bit-rate speech encoders in VoIP networks, in the design of echo canceller structures for VoIP gateways and IP-based terminals?
5. What is the effect of vocoder distortion on echo canceller performance, in particular due to the presence of encoder / decoder pairs on the send and / or receive paths? Are there ways to mitigate the presence of this distortion?

A review of echo canceller structures, adaptation algorithms, and control algorithms in Chapter 2 indicated that these research questions have not been thoroughly addressed in the literature. In particular, it was found:

- Many doubletalk detectors have been proposed in the literature, but no work exists addressing the important problem of doubletalk detector calibration.
- Background noise places an upper limit on echo canceller performance measured using MSE or ERLE. However, little work exists in the literature to quantify the frequency-dependent masking effects of moderate-to-high levels of background noise on residual echo perception.
- No work exists in the literature for incorporating adaptation and control algorithms into echo cancellers based on critically sampled subband adaptive filter banks.
- Frequency-domain post-filtering has been successfully applied to enhance near-end speech in the presence of residual echo, including that caused by nonlinear

vocoder distortion in the echo path. However, such approaches require accurate estimates of the residual echo signal power spectrum.

Chapter 4 proposed an approach to doubletalk detector calibration based on desired statistical properties of desired maximum probability of false alarm (P_F) or probability of miss (P_M). Selecting a threshold based on those criteria requires the probability density function (PDF) of the detection statistic. This approach was applied by deriving expressions for the PDF of the normalized cross-correlation-based doubletalk detector, which forms a detection statistic from the ratio of expected to actual reference signal powers. This information was used to construct detection thresholds adaptive to changes in signal statistics. In addition, a number of conclusions can be made:

- The detection statistic's PDF is a function of the parameter estimation window size, echo-to-noise ratio (SNR) and, in the presence of doubletalk, the near-end speech power. As a result, detection thresholds constructed using statistical criteria must be adaptive to changes in the input and noise statistics.
- It follows from the above that for a given detection threshold, the probability of miss (P_M) is dependent not only on the near-end speech to echo power ratio (NER), but also on the SNR during doubletalk conditions.
- Previous authors have proposed to compensate for the presence of noise by estimating the background noise power and applying it to the detection statistic. It was shown that although this compensates for bias, lower SNR still increases the likelihood of false alarm and miss.

Chapter 5 investigated the effect of noise on echo canceller performance measures, and in particular the masking effects of background noise, in an attempt to identify psychoacoustic limits of echo canceller performance. The masking effects of background noise were identified by applying the MPEG psychoacoustic model to the estimated noise power spectrum. Two echo canceller performance measures were presented that calculate the amount of audible echo power reduction based on power spectrum estimates of the echo and residual echo signals compared to the masking threshold induced by background noise. Simulation and informal listening tests were used to verify these effects. In addition, a number of conclusions can be made:

- Psychoacoustic limits of human hearing, namely the absolute hearing threshold and the masking threshold of background noise, induce an upper bound on echo canceller performance in terms of maximum audible echo reduction. This is in addition to other known limiting effects, such as transducer nonlinearity and finite-word-length effects.
- It is possible to have echo and / or residual echo frequency components that are above or below the average signal echo power, and also above or below the absolute hearing threshold. Therefore, average-power-based performance measures such as ERLE cannot incorporate frequency-dependent limitations of human hearing, particularly those due to the masking presence of noise. As a result, ERLE may under- or over-estimate the amount of audible echo cancellation provided by an echo canceller.

Chapter 6 investigated methods of incorporating adaptation and control algorithms into echo cancellers employing recently proposed critically sampled subband adaptive filters (CS-SBAF). One particular structure investigated in this chapter avoids the requirement of "cross-adaptive" filters to compensate for aliasing. An Affine Projection (AP) algorithm was presented for use in the CS-SBAF structure, along with a convergence analysis of the algorithm. The normalized cross-correlation-based doubletalk detector was derived to provide a detection statistic for each subband. It was shown how the detection thresholds from Chapter 4 can be applied to doubletalk detectors in individual subbands, and made adaptive to per-subband differences in SNR and NER. In addition, a number of conclusions can be made:

- Per-subband AP, even with a small number of subbands, can obtain a faster rate of convergence than fullband AP employing the same projection order.
- Background noise is not spectrally flat in practice, so having estimates of the echo to background noise ratio (SNR) in each subband allows per-subband detection thresholds to be constructed.
- The statistical model presented in Chapter 4 revealed that the best detection (lowest P_M) occurs under conditions of high SNR. Therefore, doubletalk detection rates can be improved over a fullband doubletalk detector by using detection statistics in subbands that have SNR higher than the fullband SNR.

Chapter 7 focused on adaptation and control algorithms for echo cancellers deployed in VoIP networks, and in particular at IP-PSTN gateways and in IP telephones. It was shown that LPC-based speech parameters used to reconstruct signals at speech decoders are the same parameters as those that would be calculated in existing algorithms based on decorrelation of the input and error signals. This led to the development of decorrelated NLMS and normalized cross-correlation-based doubletalk detector algorithms employing parameters tapped from the speech decoder. The resulting algorithms are functionally similar, but avoid the computational cost of constructing decorrelation filter coefficients. Another issue addressed was the problem of vocoder distortion in send and / or receive paths of the echo canceller. In this type of configuration, the resulting nonlinearity in the echo path severely limits the achievable ERLE. A residual echo power spectrum estimation algorithm was proposed that improves upon existing algorithms by incorporating frequency weighting characteristics of quantization noise in LPC-based encoder structures. In addition, a number of conclusions can be made:

- Simulation results showed that even with the presence of a pitch synthesis filter, the short-term excitation signal tapped from an LPC-based speech decoder represents a significantly more decorrelated version of the reconstructed input signal.
- For real speech signals, LPC synthesis filter coefficients are not stationary over the entire duration of echo paths typical of acoustic environments. As a result, modifications must be made to decorrelated NLMS / doubletalk detector algorithms to compensate for this.

- Even with the low amount of cancellation provided by a linear adaptive filter, the residual echo power spectrum is dominated by vocoder distortion from the send and receive paths. In addition, the quantization noise produced by encoders is higher at peaks in the LPC spectrum corresponding to formant frequencies.

8.2 Summary of Contributions

The focus of this thesis was on the development of echo canceller structures, adaptation algorithms, and control algorithms capable of operating in harsh environments. The subsequent investigations, developments, and simulation results provide several contributions to the field of adaptive filter theory applied to network and acoustic echo cancellation. These contributions are briefly summarized as follows:

1. A statistical model was developed for characterizing the behavior of the normalized cross-correlation-based doubletalk detector from the literature in the absence and presence of doubletalk. The model is simple and characterizes the detection statistic as a function of the parameter estimation window size (K), echo to noise ratio (SNR), and the near-end speech to echo ratio (NER).
2. Two doubletalk detector calibration algorithms were developed for constructing optimal detection thresholds adaptive to changes in the input and noise signal statistics. The thresholds are based on desirable probabilities of false alarm (P_F) and miss (P_M), are simple to implement using look-up tables, and offer improved performance over fixed, empirically determined thresholds.

3. Two algorithms were proposed for measuring audible echo canceller performance based on psychoacoustic limitations of human hearing. The algorithms are an improvement on existing average power measures in that they incorporate the masking effects of moderate-to-high levels of background noise obtained from the proven MPEG psychoacoustic model.
4. A convergence analysis was presented for the subband AP algorithm in a recently proposed critically sampled subband adaptive filter (CS-SBAF) structure. Because of aliasing present in the signals, the convergence behaviour of individual subbands is dependent upon those in adjacent subbands.
5. A subband doubletalk detector was derived for use in the CS-SBAF structure based on normalized cross-correlation. Combining the doubletalk detector with signal-adaptive detection thresholds, in particular per-subband estimates of the SNR, allows increased detection probability over fullband doubletalk detectors.
6. Two methods were proposed for implementing decorrelated NLMS and normalized cross-correlation-based doubletalk detector algorithms in VoIP gateways or IP-based telephones. The structures achieve a computational savings by taking incorporating LPC-based speech parameters available at decoders.
7. A power spectrum estimation algorithm was proposed for modeling residual echo resulting from nonlinear vocoder distortion in the echo path. The algorithm was incorporated into a frequency-domain post-filter, and offers an improvement in near-end speech quality compared to existing estimation techniques from the literature.

8. Publication of several refereed conference and journal papers which report on the research results outlined above: [26], [27], [28], [29], [30], [31], [32], [33], [34], [35].

8.3 Suggestions for Future Research

Finally, during the course of this work several issues arose which merit further research. These are summarized as follows:

1. Combining psychoacoustic aspects of human hearing into the design of subband echo canceller structures and control algorithms. Chapter 6 showed that there is value in having per-subband estimates of SNR to calibrate adaptation and control algorithms in terms of improved convergence and doubletalk detection rates. Therefore, it seems reasonable to employ per-subband NLP or post-filtering algorithms incorporating psychoacoustic limitations of human hearing. One interesting approach appears in [64] that approximates frequency-domain post-filtering using subband signals, but it does not incorporate any psychoacoustic model of hearing.
2. Variable step-size control for AP / NLMS in CS-SBAF structures. It is known that adaptive filter misadjustment can be reduced by decreasing the adaptation step size after a period of fast initial convergence [115], [116]. In addition, AP is subject to higher MSE as the projection order increases, an effect observed in [87] and in Chapter 6. Therefore, it would seem beneficial to have a similar step-size control mechanism available for the CS-SBAF structure. An interesting variable-

step-size algorithm employs the correlation of AP projection vectors as a measure of convergence and doubletalk [117].

3. Comparison of decorrelated NLMS with other low-complexity adaptation algorithms. Many variations of NLMS-type algorithms exist in the literature, such as those based on partial coefficient updates [118], [119]. More recently, algorithms have been proposed that offer better convergence performance for NLMS and network echo paths with a small increase in cost, such as proportionate NLMS and its variants [120], [121], [122]. It would be useful to compare these algorithms with the decorrelated NLMS algorithms of Chapter 7 (both the existing and proposed algorithms) with respect to convergence rate and complexity. In addition, it would be interesting to determine whether the use of LPC-based speech parameters can be incorporated into these algorithms as well.
4. Applying the statistical modeling approach presented in Chapter 4 to other doubletalk detectors from the literature for calibration and as a means of performance comparison. The doubletalk detector statistical modeling approach was applied to the normalized cross-correlation-based doubletalk detector of [16] to investigate its usefulness for calibration. It would be beneficial to apply similar modeling to other doubletalk detectors in the literature as a means of comparing their performance with respect to probability of miss in the presence of noise and as a function of NER.
5. Compare and incorporate the statistical modeling of doubletalk detectors of Chapter 4 with other approaches. It was shown in Chapter 4 that a time-varying

detection threshold can improve detection rates over a fixed threshold, but this adaptive threshold is still used to produce a "hard" doubletalk decision (i.e., present or not). Variable step-size control algorithms exist in the literature to provide "soft" decisions on the severity of doubletalk, and to differentiate between doubletalk and changes in echo path impulse response [9], [117]. It may be useful to investigate the relationship, if any, that exists between these approaches.

References

[1] P. J. Smith *et al.*, "Tandem-free VoIP conferencing: a bridge to next-generation networks", *IEEE Commun. Mag.*, vol. 41, no. 5, pp. 136 – 145, May 2003.

[2] E. Hänsler, "The hands-free telephone problem – an annotated bibliography," *Signal Process.*, vol. 27, no. 3, pp. 259 – 271, 1992.

[3] Q. Bi, G. I. Zysman, and H. Menkes, "Wireless mobile communications at the start of the 21st century," *IEEE Commun. Mag.*, vol. 39, no. 1, pp. 110 – 116, Jan. 2001.

[4] T. J. Kostas *et al.*, "Real-time voice over packet-switched networks," *IEEE Network Mag.*, vol. 12, no. 1, pp. 18 – 27, Jan. – Feb. 1998.

[5] B. Goode, "Voice over Internet Protocol (VoIP)," *Proc. IEEE*, vol. 90, no. 9, pp. 1495 – 1517, Sep. 2002.

[6] V. G. Cerf, "On the evolution of Internet technologies," *Proc. IEEE*, vol. 92, no. 9, pp. 1360 – 1370, Sep. 2004.

[7] C. Breining *et al.*, "Acoustic echo control: An application of very-high-order adaptive filters," *IEEE Signal Process. Mag.*, vol. 16, no. 4, pp. 42 – 69, Jul. 1999.

[8] P. Heitkamper, "An adaptation control for acoustic echo cancellers," *IEEE Signal Process. Lett.*, vol. 4, no. 6, pp. 170 – 172, Jun. 1997.

[9] T. Gänsler *et al.*, "Double-talk robust fast converging algorithms for network echo cancellation", *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 656 – 663, Nov. 2000.

[10] A. N. Birkett and R. A. Goubran, "Limitations of hands-free acoustic echo cancellers due to nonlinear loudspeaker distortion and enclosure vibration effects," in *Proc. IEEE WASPAA*, Oct. 1995, pp. 103 – 106.

[11] International Telecommunication Union, *ITU-T G.131: Talker echo and its control*, ITU 2003.

[12] International Telecommunication Union, *ITU-T G.168: Digital network echo cancellers*, ITU 2002.

[13] D. L. Duttweiler, "A twelve-channel digital echo canceller," *IEEE Trans. Commun.*, vol. 26, pp. 647 – 653, May 1978.

[14] H. Ye and B. Wu, "A new double-talk detection algorithm based on the orthogonality theorem," *IEEE Trans. Commun.*, vol. 39, no. 11, pp. 1542 – 1545, Nov. 1991.

[15] T. Gänsler *et al.*, "A double-talk detector based on coherence," *IEEE Trans. Commun.*, vol. 44, no. 11, pp. 1421 – 1427, Nov. 1996.

[16] J. Benesty, D. R. Morgan, and J. H. Cho, "A new class of doubletalk detectors based on cross-correlation," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 168 – 172, Mar. 2000.

[17] International Telecommunication Union, *ITU-T G.711: Pulse code modulation (PCM) of voice frequencies*, ITU 1988.

[18] International Telecommunication Union, *ITU-T G.729: Coding of speech at 8 kbit/s using conjugate-structure algebraic code-excited linear prediction (CS-ACELP)*, ITU 1996.

[19] International Telecommunication Union, *ITU-T G.723.1: Dual rate speech coder for multimedia communications transmission at 5.3 and 6.3 kbit/s*, ITU 1996.

[20] A. P. Markopoulou, F. A. Tobagi, and M. J. Karam, "Assessing the quality of voice communications over Internet backbones," *IEEE/ACM Trans. Netw.*, vol. 11, no. 5, pp. 747 – 760, Oct. 2003.

[21] Y. Huang, *Effects of Vocoder Distortion and Packet Loss on Network Echo Cancellation*, M.Sc. thesis, Carleton University, Jan. 2000.

[22] R. Chandran and D. J. Marchok, "Compressed domain noise reduction and echo suppression for network speech enhancement," in *Proc. IEEE MWSCAS*, Aug. 2000, vol. 1, pp. 10 – 13.

[23] A. Gatherer *et al.*, "DSP-based architectures for mobile communications: past, present and future," *IEEE Commun. Mag.*, vol. 38, no. 1, pp. 84 – 90, Jan. 2000.

[24] J. J. Shynk, "Frequency domain and multirate adaptive filtering," *IEEE Signal Process. Mag.*, vol. 9, no. 1, pp. 14 – 37, Jan. 1992.

[25] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: analysis, experiments, and application to acoustic echo cancellation," *IEEE Trans. Signal Process.*, vol. 40, no. 8, pp. 1862 – 1875, Aug. 1992.

[26] J. D. Gordy and R. A. Goubran, "Statistical analysis of doubletalk detection for calibration and performance evaluation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1035 – 1043, Mar. 2007.

[27] J. D. Gordy and R. A. Goubran, "Online doubletalk detector calibration for acoustic echo cancellation in videoconferencing systems," in *Proc. IEEE ICME*, July 2006, vol. 3, pp. 1957 – 1960.

[28] J. D. Gordy and R. A. Goubran, "On the perceptual performance limitations of echo cancellers in wideband telephony," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 33 – 42, Jan. 2006.

[29] J. D. Gordy and R. A. Goubran, "A perceptual performance measure for adaptive echo cancellers in packet-based telephony," in *Proc. IEEE ICME*, Jul. 2005, vol. 1, pp. 431 – 434.

[30] J. D. Gordy and R. A. Goubran, "Fast system identification using affine projection and a critically sampled subband adaptive filter," *IEEE Trans. Instrum. Meas.*, vol. 55, no. 4, pp. 1242 – 1249, Aug. 2006.

[31] J. D. Gordy and R. A. Goubran, "A subband doubletalk detector for echo cancellation in hands-free environments," in *Proc. IEEE VTC*, Sep. 2005, vol. 3, pp. 1397 – 1401.

[32] J. D. Gordy and R. A. Goubran, "A combined LPC-based speech coder and filtered-X LMS algorithm for acoustic echo cancellation," in *Proc. IEEE ICASSP*, May 2004, vol. 4, pp. 125 – 128.

[33] J. D. Gordy and R. A. Goubran, "A low-complexity doubletalk detector for echo cancellers in packet-based telephony," in *Proc. IEEE WASPAA*, Oct. 2005, pp. 74 – 77.

[34] J. D. Gordy and R. A. Goubran, "Postfiltering for suppression of residual echo from vocoder distortion in packet-based telephony," in *Proc. IEEE ICME*, July 2006, vol. 3, pp. 1953 – 1956.

[35] J. D. Gordy and R. A. Goubran, "Reduced-complexity mixing of compressed speech signals for VoIP and cellular telephony," *IEEE Trans. Audio, Speech, Lang. Process.*, 10 pages, submitted Oct. 16, 2006.

[36] F. Kuch and W. Kellermann, "Nonlinear line echo cancellation using a simplified second order Volterra filter," in *Proc. IEEE ICASSP*, May 2002, vol. 2, pp. 1117 – 1120.

[37] T. N. Yensen, R. A. Goubran and I. Lambadaris, "Synthetic stereo acoustic echo cancellation structure for multiple participant VoIP conferences," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 2, pp. 168 – 174, Feb. 2001.

[38] P. Eneroth *et al.*, "A real-time implementation of a stereophonic acoustic echo canceller," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 513 – 523, Jul. 2001.

[39] R. Appel and J. G. Beerends, "On the quality of hearing one's own voice," *J. Audio Eng. Soc.*, vol. 50, pp. 237 – 248, Apr. 2002.

[40] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 217 – 231, Mar. 2001.

[41] International Telecommunication Union, *ITU-T P.800: Methods for subjective determination of transmission quality*, ITU 1996.

[42] International Telecommunication Union, *ITU-T P.862: Perceptual evaluation of speech quality (PESQ)*, ITU 2001.

[43] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 1989.

[44] S. Haykin, *Adaptive Filter Theory*, 3rd ed. Upper Saddle River, NJ: Prentice-Hall, 1996.

[45] K. Mayyas and T. Aboulnasr, "Reduced-complexity transform-domain adaptive algorithm with selective coefficient update," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 51, no. 3, pp. 136 – 142, Mar. 2004.

[46] T. Gänsler, "A robust frequency-domain echo canceller," in *Proc. IEEE ICASSP*, Apr. 1997, vol. 3, pp. 2317 – 2320.

[47] K. Eneman and M. Moonen, "Iterated partitioned block frequency-domain adaptive filtering for acoustic echo cancellation," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 2, pp. 143 – 158, Mar. 2003.

[48] J. J. Shynk, "Adaptive IIR filtering," *IEEE Signal Process. Mag.*, vol. 6, no. 2, pp. 4 – 21, Apr. 1989.

[49] P. P. Vaidyanathan, "Multirate digital filters, filter banks, polyphase networks, and applications: a tutorial," *Proc. IEEE*, vol. 78, no. 1, pp. 56 – 93, Jan. 1990.

[50] T. P. Barnwell, III and M. J. T. Smith, "Filter banks for analysis-reconstruction systems: a tutorial," in *Proc. IEEE ISCAS*, May 1990, vol. 3, pp. 1999 – 2003.

[51] S. S. Pradham and V. U. Reddy, "A new approach to subband adaptive filtering," *IEEE Trans. Signal Process.*, vol. 47, no. 3, pp. 655 – 664, Mar. 1999.

[52] F. Amano *et al.*, "A multirate acoustic echo canceller structure," *IEEE Trans. Commun.*, vol. 43, no. 7, pp. 2172 – 2176, Jul. 1995.

[53] N. J. Bershad and J. C. M. Bermudez, "New insights on the transient and steady-state behavior of the quantized LMS algorithm," *IEEE Trans. Signal Process.*, vol. 44, no. 10, pp. 2623 – 2625, Oct. 1996.

[54] D. R. Morgan, "Slow asymptotic convergence of LMS acoustic echo cancellers," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 2, pp. 126 – 136, Mar. 1995.

[55] S. M. Kuo and D. R. Morgan, "Active noise control: a tutorial review," *Proc. IEEE*, vol. 87, no. 6, pp. 943 – 973, Jun. 1999.

[56] V. Myllyla and G. Schmidt, "Pseudo-optimal regularization for affine projection algorithms," in *Proc. IEEE ICASSP*, May 2002, vol. 2, pp. 1917 – 1920.

[57] H. Rey *et al.*, "Variable explicit regularization in affine projection algorithm: robustness issues and optimal choice," *IEEE Trans. Signal Process.*, vol. 55, no. 5, pp. 2096 – 2109, May 2007.

[58] K. Ozeki and T. Umeda, "An adaptive filtering algorithm using orthogonal projection to an affine subspace and its properties," *Electronics and Communications in Japan.*, vol. 67-A, no. 5, pp. 19 – 27, 1984.

[59] M. Rupp, "A family of adaptive filter algorithms with decorrelating properties," *IEEE Trans. Signal Process.*, vol. 46, no. 3, pp. 771 – 775, Mar. 1998.

[60] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[61] S. L. Gay and S. Tavathia, "The fast affine projection algorithm," in *Proc. IEEE ICASSP*, May 1995, vol. 5, no. 3023 – 3026.

[62] F. Albu *et al.*, "The Gauss-Seidel fast affine projection algorithm," in *Proc. IEEE WSPS*, Oct. 2002, pp. 109 – 114.

[63] H. Ding, "Fast affine projection adaptation algorithms featuring stable symmetric positive-definite linear system solvers," in *Proc. IEEE WASPAA*, Oct. 2005, pp. 166 – 169.

[64] X. Lu and B. Champagne, "Acoustic echo cancellation with post-filtering in subband," in *Proc. IEEE WASPAA*, Oct. 2003, pp. 29 – 32.

[65] P. Sristi, W.-S. Lu, and A. Antoniou, "A new variable-step-size LMS algorithm and its application in subband adaptive filtering for echo cancellation," in *Proc. IEEE ISCAS*, May 2001, vol. 2, pp. 721 – 724.

[66] D. R. Morgan and J. C. Thi, "A delayless subband adaptive filter architecture," *IEEE Trans. Signal Process.*, vol. 43, no. 8, pp. 1819 – 1830, Aug. 1995.

[67] S. Ohno and H. Sakai, "On delayless subband adaptive filtering by subband/fullband transforms," *IEEE Signal Process. Lett.*, vol. 6, no. 9, pp. 236 – 239, Sep. 1999.

[68] M. R. Petraglia, R. G. Alves, and P. S. R. Diniz, "New structures for adaptive filtering in subbands with critical sampling," *IEEE Trans. Signal Process.*, vol. 48, no. 12, pp. 3316 – 3327, Dec. 2000.

[69] H. R. Abutalebi *et al.*, "Affine projection algorithm for oversampled subband adaptive filters," in *Proc. IEEE ICASSP*, Apr. 2003, vol. 6, pp. 209 – 212.

[70] J. H. Cho, D. R. Morgan, and J. Benesty, "An objective technique for evaluating doubletalk detectors in acoustic echo cancellers," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 718 – 724, Nov. 1999.

[71] S. M. Kuo and W. S. Gan, "Analysis of nonlinear residual echo suppressors for telecommunications," in *Proc. IEEE ISCAS*, May 2005, vol. 6, pp. 5994 – 5997.

[72] S. J. Park, C. Lee, and D. H. Youn, "A residual echo cancellation scheme for hands-free telephony," *IEEE Signal Process. Lett.*, vol. 9, no. 12, pp. 397 – 399, Dec. 2002.

[73] L. Ding, M. S. El-Hennawey, and R. A. Goubran, "Measurement of the effects of temporal clipping on speech quality," in *Proc. IEEE IMTC*, May 2005, vol. 2, pp. 1135 – 1138.

[74] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526 – 1555, Oct. 1992.

[75] V. Turbin, A. Gilloire, and P. Scalart, "Comparison of three post-filtering algorithms for residual acoustic echo reduction," in *Proc. IEEE ICASSP*, Apr. 1997, vol. 1, pp. 307 – 310.

[76] S. Gustafsson *et al.*, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 245 – 256, Jul. 2002.

[77] M. Maresca, N. Zingirian, and P. Baglietto, "Internet protocol support for telephony," *Proc. IEEE*, vol. 92. no. 9, pp. 1463 – 1477, Apr. 2004.

[78] Internet Engineering Task Force, *SIP: Session Initiation Protocol*, RFC 3261, Jun. 2002.

[79] Internet Engineering Task Force, *RTP: A Transport Protocol for Real-Time Applications*, RFC 1889, Jan. 1996.

[80] R. Ramjee *et al.*, "Adaptive playout mechanisms for packetized audio applications in wide-area networks," in *Proc. IEEE Infocom*, Jun. 1994, vol. 2. pp. 680 – 688.

[81] A. S. Spanias, "Speech coding: a tutorial review," *Proc. IEEE*, vol. 82, no. 10, pp. 1541 – 1582, Oct. 1994.

[82] M. Schroeder and B. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates", in *Proc. IEEE ICASSP*, Apr. 1985, vol. 10, pp. 937 – 940.

[83] L. Ding and R. A. Goubran, "Speech quality prediction in VoIP using the extended E-model," in *Proc. IEEE GLOBECOM*, Dec. 2003, vol. 7, pp. 3974 – 3978.

[84] P. L. De Leon and D. M. Etter, "Experimental results with increased bandwidth analysis filters in oversampled, subband acoustic echo cancellers," *IEEE Signal Process. Lett.*, vol. 2, no. 1, pp. 1 – 3, Jan. 1995.

[85] T. Jia *et al.*, "Subband doubletalk detector for acoustic echo cancellation systems," in *Proc. IEEE ICASSP*, May 2003, vol. 5, pp. 604 – 607.

[86] M. R. Petraglia, R. G. Alves, and M. N. S. Swamy, "A new open loop delayless subband adaptive filter structure," in *Proc. IEEE ICASSP*, May 2002, vol. 2, pp. 1345 – 1348.

[87] S. G. Sankaran and A. A. Beex, "Convergence behavior of affine projection algorithms," *IEEE Trans. Signal Process.*, vol. 48, no. 4, pp. 1086 – 1096, Apr. 2000.

[88] S. K. Mitra, *Digital Signal Processing: A Computer-Based Approach*, 2nd ed. Boston: McGraw Hill, 2001.

[89] J. Lariviere and R. Goubran, "Noise-reduced GMDF for acoustic echo cancellation and speech recognition in mobile environments," in *Proc. IEEE VTC*, Sep. 2000, vol. 6, pp. 2969 – 2972.

[90] A. Sugiyama, J. Berclaz, and M. Sato, "Noise-robust double-talk detection based on normalized cross correlation and a noise offset," in *Proc. IEEE ICASSP*, Mar. 2005, vol. 3, pp. 153 – 156.

[91] M. Mboup, M. Bonnet, and N. Bershad, "LMS coupled adaptive prediction and system identification: a statistical model and transient mean analysis," *IEEE Trans. Signal Process.*, vol. 42, no. 10, pp. 2607 – 2614, Oct. 1994.

[92] J. Garofolo *et al.*, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM.* Gaithersburg, MD: NIST, 1990.

[93] International Telecommunication Union, *ITU-T G.729 Annex A: Reduced complexity 8 kbit/s CS-ACELP speech codec*, ITU 1996.

[94] P. Ahgren and A. Jakobsson, "A study of double-talk detection performance in the presence of acoustic echo path changes," in *Proc. IEEE ICASSP*, Mar. 2005, vol. 3, pp. 141 – 144.

[95] J. C. Jenq and S. F. Hseih, "Decision of double-talk and time-variant echo path for acoustic echo cancellation," *IEEE Signal Process. Lett.*, vol. 10, no. 11, pp. 317 – 319, Nov. 2003.

[96] N. J. Bershad and J.-Y. Tourneret, "Echo cancellation – a likelihood ratio test for double-talk versus channel change," *IEEE Signal Process.*, vol. 54, no. 12, pp. 4572 – 4581, Dec. 2006.

[97] C. Carlemalm, F. Gustafsson, and B. Wahlberg, "On the problem of detection and discrimination of double talk and change in the echo path," in *Proc. IEEE ICASSP*, May 1996, vol. 5, pp. 2742 – 2745.

[98] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*. Hoboken, NJ: Wiley, 2004.

[99] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Process. Lett.*, vol. 10, no. 7, pp. 204 – 207, Jul. 2003.

[100] Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed. Burr Ridge, IL: McGraw-Hill, 2002.

[101] Rose and M. D. Smith, *Mathematical Statistics with Mathematica.* New York: Springer, 2002.

[102] International Telecommunication Union, *ITU-T G.722.2: Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)*, ITU 2001.

[103] I. Goetz, "VoIP interconnect and quality impacts on 3G core network evolution," in *Proc. IEE Int. Conf. on 3G Mobile Comm. Tech.*, pp. 534 – 536.

[104] X. Lu and B. Champagne, "A centralized acoustic echo canceller exploiting masking properties of the human ear," in *Proc. IEEE ICASSP*, May 2003, vol. 5, pp. 377 – 380.

[105] G. Enzner, H. Kruger, and P. Vary, "On the problem of acoustic echo control in cellular networks," in *Proc. IEEE IWAENC*, Sep. 2005, pp. 213 – 216.

[106] International Organization for Standardization, *Information technology – coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part 3: Audio*, ISO/IEC 11172-3, 1993.

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

[108] International Organization for Standardization, *Acoustics – Normal equal-loudness-level contours*, ISO 226:2003, 2003.

[109] G. Enzner, R. Martin and P. Vary, "Unbiased residual echo power estimation for hands-free telephony," in *Proc. IEEE ICASSP*, May 2002, vol. 2, pp. 1893 – 1896.

[110] M. E. Knappe and R. A. Goubran, "Steady-state performance limitations of full-band acoustic echo cancellers," in *Proc. IEEE ICASSP*, Apr. 1994, vol. 2, pp. 73 – 76.

[111] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, no. 4, pp. 451 – 513, Apr. 2000.

[112] A. Sugiyama and F. Landais, "A new adaptive intersubband tap-assignment algorithm for subband adaptive filters," in Proc. IEEE ICASSP, May 1995, vol. 5, pp. 3051 – 3054.

[113] M. Vukadinovic and T. Aboulnasr, "A study of adaptive intersubband tap assignment algorithms from a psychoacoustic point of view," in *Proc. IEEE ISCAS*, May 1996, vol. 2, pp. 65 – 68.

[114] R. Frenzel and M. E. Hennecke, "Using prewhitening and step-size control to improve the performance of the LMS algorithm for acoustic echo cancellation," in *Proc. IEEE ISCAS*, May 1992, vol. 4, pp. 1930 – 1932.

[115] A. Mader, H. Puder, and G. U. Schmidt, "Step-size control for acoustic echo cancellation filters – An overview," *Signal Process.*, vol. 80, pp. 1697 – 1719, Sep. 2000.

[116] H.-C. Sin, A. H. Sayed, and W.-J. Song, "Variable step-size NLMS and affine projection algorithms," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 132 – 135, Feb. 2004.

[117] T. Creasy and T. Aboulnasr, "A projection-correlation algorithm for acoustic echo cancellation in the presence of double talk," in *Proc. IEEE ICASSP*, Jun. 2000, vol. 1, pp. 436 – 439.

[118] T. Aboulnasr and K. Mayyas, "Complexity reduction of the NLMS algorithm via selective coefficient update," *IEEE Trans. Signal Process.*, vol. 47, no. 5, pp. 1421 – 1424, May 1999.

[119] P. A. Naylor and W. Sherliker, "A short-sort M-Max NLMS partial-update adaptive filter with applications to echo cancellation," in *Proc. IEEE ICASSP*, May 2003, vol. 5, pp. 373 – 376.

[120] D. L. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancellers," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 5, pp. 508 – 518, Sep. 2000.

[121] J. Benesty and S. L. Gay, "An improved PNLMS algorithm," in *Proc. IEEE ICASSP*, May 2002, vol. 2, pp. 1881 – 1884.

[122] H. Deng and M. Doroslovacki, "Proportionate adaptive algorithms for network echo cancellation," *IEEE Trans. Signal Process.*, vol. 54, no. 5, pp. 1794 – 1803, May 2006.