Acoustic Echo Cancellation for Wideband VoIP

by

Brady Laska

A Thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfilment of
the requirements for the degree of
Master of Applied Science

Ottawa-Carleton Institute for
Electrical and Computer Engineering

Department of Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario, Canada

September 7, 2006

Your file Votre référence
ISBN: 978-0-494-18321-2
Our file Notre référence
ISBN: 978-0-494-18321-2

Abstract

Acoustic echo cancellation is investigated in a wideband voice-over-IP (VoIP) framework. Simulations using fullband and subband adaptive filtering algorithms show that subband adaptive filters are an effective solution. Low residual echo levels observed in oversampled subband echo cancellers are attributed to the adaptive filters exploiting the subband signal correlation to lower the output error. In simulated changing acoustic environments the minimum echo return loss enhancement (ERLE) of the subband Normalised LMS (NLMS) is over 3 dB higher than fullband; the result is verified using experimental data. In the presence of narrowband near-end disturbances with an active doubletalk detector some bands in a subband system can continue to adapt resulting in deeper convergence and higher ERLE during doubletalk. Nonlinear distortion created by wideband VoIP vocoders can degrade the performance of linear echo cancellers by 5 – 10 dB. The fast tracking IP Affine Projection Algorithm (IP-APA) is shown to perform the best in the distorted channel, the differences between subband and fullband structures are not significant. Techniques from Fast Affine Projection algorithms are used to reduce the complexity of the IP Affine Projection Algorithm, making it more applicable for wideband acoustic echo cancellation.

Acknowledgments

First and foremost I would like to thank my supervisor Dr. Rafik Goubran for his guidance, encouragement and endless enthusiasm. I would also like to acknowledge the financial support of the Ontario Graduate Scholarship (OGS) program, the Ontario Centres of Excellence (OCE) Centre for Communications and Information Technology (CITO), and Mitel Networks.

I would like thank the Mitel DSP and acoustics research staff for their help and suggestions, especially Franck Beaucoup for his valuable discussions, and Philippe Moquin for his assistance with the echo path impulse response measurements. I would also like to thank my fellow graduate students in the DSP lab, especially James Gordy for his helpful suggestions and discussion.

I would like to thank my sisters, my brother and my parents for their encouragement. Finally I would like to thank Susan for her love, support, and patience.

Table of Contents

List of Figures

List of Symbols

n	Discrete time index
ω	Discrete frequency $\omega \in [0, 2\pi]$
$(\cdot)^T$	Matrix/vector transpose
$(\cdot)^*$	Complex conjugate
$(\cdot)^H$	Matrix/vector Hermitian transpose
$\mathcal{E}\{\cdot\}$	Mathematical expectation
$\underline{h}(n)$	Echo path
N	Adaptive filter length
$x(n)$	Far-end speech signal
$d(n)$	Echo signal
$y(n)$	Near-end microphone signal
$e(n)$	Error signal
$\hat{(\cdot)}$	Signal/vector estimate
M	Number of sub-bands
m	Subband index
D	Subband decimation factor
$\underline{H}_M$	Analysis filterbank
$\underline{F}_M$	Synthesis filterbank

Chapter 1

Introduction

1.1 Echo Cancellation

Echoes in a telecommunications system occur whenever the signal sent out from the transmit side of the system is coupled back to the receiver side. If the coupling is electrical, occurring due to impedance mismatches in the telephone network, the result is known as a network echo, a discussion of which can be found in [3]. Acoustic echoes on the other hand arise when a hands-free telephone is used, as in desktop teleconferencing and in-car wireless mobile telephones. The speaker signal from the hands-free terminal is acoustically coupled to the microphone, producing the echo. With a traditional telephone handset the loudspeaker is blocked by the ear against the earpiece. This disrupts the acoustic path between the loudspeaker and the microphone, resulting in echo signal attenuation of at least 45 dB [4]. In contrast, with a hands-free terminal the acoustic path is open and the microphone is free to pick up the signal radiated by the loudspeaker and transmit it back to the far end. In echo situations the far-end talker hears a version of their own speech delayed by the echo path length, this disturbs the talker, making natural conversation difficult. In

extreme situations coupling at both ends of the connection results in a positive feedback loop which can leading to howling instability. In order to maintain comfortable and natural speech acoustic echoes must be suppressed or removed.

The amount of required echo attenuation is dependent on the echo level and the length of the delay between the original speech and the echo. Very early reflections are interpreted as spectral distortion or reverberation rather than as distinct echoes. A small amount of reverberation is prefereable to none at all; telephone handsets deliberately allow some of the transmitted signal to pass through to the receiver to make the line feel "alive". It is not until the delay between the original speech and the echoed speech reaches approximately 35 ms that the echo becomes an annoyance [3]. The perceptual level of the echo (how loud the echo sounds to a user) increases when either the echo magnitude or the echo path delay increase. The relationship between echo delay and the level of attenuation required for comfortable speech is approximately exponential: when the one-way transmission delay is 5 ms an echo need only be 20 dB lower than the speech signal for acceptable communications, whereas when the delay is 200 ms an echo signal 50 dB lower than the speech signal is still objectionable [5].

Early approaches to acoustic echo control relied on breaking the echo path by converting the channel from full-duplex to half-duplex using voice activated switches that stop transmission when the near-end speaker is silent. This approach is suboptimal for a number of reasons: degrading the channel to half-duplex reduces the naturalness of the conversation; voice-operated switching results in temporal clipping of speech bursts, so the perceptual quality of the conversation is reduced; and if both users are talking simultaneously there is no echo suppression at either end of the link [6]. A more effective solution is echo cancellation. An echo canceller creates a replica of the echo which is subtracted from the signal sent to the far end. An ideal echo canceller

can completely remove the echo signal without distorting the near-end speech.

The adaptive echo canceller was invented in the 1960s by Logan, Kelly and Sondhi [7] as a method for controlling network echoes, it is also the most frequently used technique for acoustic echo cancellation. Figure 1.1 shows a block diagram of a generic adaptive echo canceller. In the model assumed by the echo canceller, the echo signal picked up by the microphone is composed of a direct path signal and signals reflected from neighbouring objects such as tables, walls, the floor and ceiling. With this view of the echo signal, the impulse response of the echo path can be modelled as a series of delayed impulses, where the magnitudes of the impulses are determined by the reflectivity of the corresponding surface, and the delays are determined by the acoustic path length. The far-end speech signal $x(n)$ is convolved with the echo path impulse response $\underline{h}(n)$, which is assumed to be linear. The input to the near end microphone, $y(n)$, is composed of the echoed far-end speech $d(n) = \underline{h}(n)^T \underline{x}(n)$, the near-end speech $v(n)$, and near-end background noise $b(n)$. The adaptive echo canceller creates an estimate $\hat{d}(n)$ of $d(n)$ by convolving $x(n)$ with the adaptive finite impulse response (FIR) filter tap weight vector $\hat{\underline{h}}(n)$, which is an estimate of $\underline{h}(n)$. The replica of the echoed speech is then subtracted from the near end microphone signal to form the error signal $e(n)$. If $\hat{\underline{h}}(n)$ is a perfect estimate of $\underline{h}(n)$ then the echoed speech is completely removed and the error signal transmitted to the far-end contains only the near-end speech and background noise. In practice, system noise, undermodelling of the echo path impulse response, and non-linear distortion from the speaker and microphone limit the amount of echo cancellation that can be performed therefore $e(n)$ contains some residual echo.

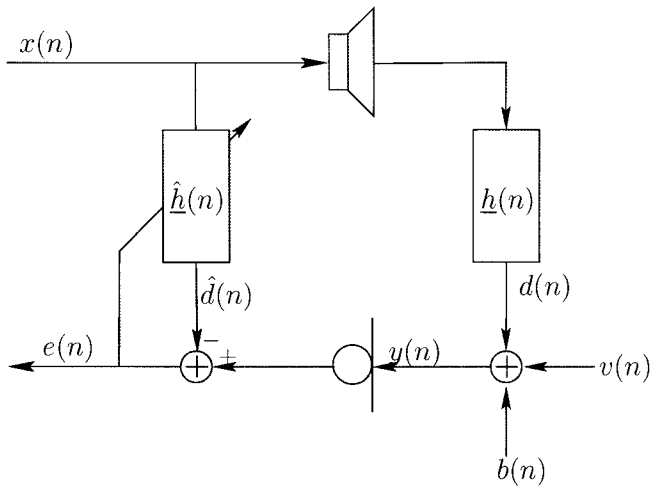


Figure 1.1: Adaptive echo canceller.

1.2 Voice Over Internet Protocol (VoIP)

In traditional wireline telephony speech is sampled at 8 kHz and the samples are directly compressed and transmitted in analogue form over the public switched telephone network (PSTN). In voice over Internet Protocol (VoIP), the sampled speech is divided into frames which are encoded and compressed using sophisticated voice coders (vocoders) which can significantly reduce the bandwidth of the data to be transmitted. The encoded voice signal is assembled into data packets which are transmitted over an Internet Protocol (IP) network such as a local area network (LAN) or a wide area network (WAN) such as the Internet. At the receiver end the speech frames are de-packetised, decoded, and the reconstructed samples are played out at 8 kHz. VoIP calls can be pure VoIP, originating and terminating at IP terminals and travelling over data networks, or they may originate or terminate on the PSTN, in which case a network interface is required to transcode between the VoIP encoded and PSTN representations. In-depth reviews of the engineering issues related to VoIP implementation can be found in [8] and [9].

The dependence of echo delay on the perceptual echo level is an important factor in

VoIP systems. Compared to PSTN conversations, VoIP conversations have additional sources of delay such as packetisation, coding, network, and playback buffer queueing delays [10] resulting in end-to-end delays of up to 200 ms [9]. Echo cancellation is therefore a crucial component of a VoIP system.

1.3 Wideband VoIP

Current phone systems including the PSTN as well as second and third generation (2G and 3G) mobile systems transmit the narrowband 200-3400 Hz frequency range of speech. This frequency limit exists in order to maintain backwards compatibility with early telephone links which were bandwidth restricted due to physical limitations of the transmission medium. With VoIP the speech content is abstracted away from the physical medium. The speech signal no longer travels directly over the PSTN, so backwards compatibility with the existing network is no longer required, and a wider frequency band may be included. The most recent voice coding standard, adopted by both the International Telecommunications Union (ITU) for land line use and the Third Generation Partnership Project (3GPP) for wireless mobile communications, operates at a sampling rate of 16 kHz[1] and increases the transmitted voice spectrum to the 50-7000 Hz frequency range [11], [12]. The extra bandwidth in the 50-200 Hz range contains low frequency background noise, the inclusion of which improves the sense of "presence" between conversation participants, enhancing the comfort and naturalness of the conversation. The high frequency addition from 3400-7000 Hz enhances speaker recognition and improves fricative differentiation and by extension speech intelligibility; certain words, such as "sit" and "fit", are difficult to distinguish at 8 kHz sampling rate, as the energy of the distinguishing sounds lies

[1]For the purposes of this thesis "wideband speech" is speech sampled at 16 kHz and "narrowband speech" is sampled at 8 kHz.

in the 5 – 8 kHz range [13]. Increasing the frequency range of transmitted speech also results in a subjective quality improvements: conversations seem more face-to-face, and communication becomes more transparent [14], [15].

1.4 Problem Statement and Thesis Objectives

The objective of this thesis is to investigate the problem of acoustic echo cancellation in a wideband VoIP environment and to determine which adaptive filtering algorithms and structures offer the best wideband acoustic echo cancellation performance. In order to maintain the perceptual benefits of wideband speech, echoes must be adequately controlled. While there is a large body of research dedicated to acoustic echo cancellation in the PSTN framework (see eg., [4] for a review) very little work has been done to examine the specific issues associated with wideband echo cancellation, or the differences between the wideband and narrowband environments. Acoustic characteristics are not constant across frequencies: walls, floors and ceilings reflect and absorb higher frequency sounds differently than low frequencies. Similarly, the spectral distribution of speech energy is not uniform: it is mostly concentrated in low frequencies, with bursts of high frequency energy arising from plosive or fricative sounds.

The frequency dependence of acoustic environment and speech properties affects the performance of acoustic echo cancellers. The highly coloured wideband speech spectrum slows the convergence of many adaptive echo cancellers. This problem is further hampered by the long adaptive filters required to model wideband environments: an adaptive echo canceller operating at 16 kHz requires twice as many taps to model the same echo path length as an 8 kHz canceller, and long adaptive filters are

slower to converge. This hindered convergence results in high residual echo levels, especially in changing acoustic environments where the adaptive filters must constantly re-converge and adapt to track the non-stationary echo path.

Many have argued that subband structures are required at higher sampling rates to overcome the slow convergence of long adaptive filters for highly coloured inputs and also to manage the computational complexity associated with adapting long filters [16]. Subband adaptive filters may have advantages in other echo cancellation scenarios. Convergence in the presence of high level near-end disturbances and echo path tracking in a non-stationary acoustic environment are two challenging adaptive echo cancellation situations that may benefit from the ability of a subband echo canceller to process different frequency bands independently. This thesis will compare the performance of subband and fullband adaptive filters in diverse wideband echo situations to determine which structure is best suited for wideband echo cancellation.

Since wideband telephony is typically achieved using VoIP, this thesis will also examine concerns specific to echo cancellation in a VoIP network. When a VoIP vocoder is present in an echo path, the echo path becomes non-linear and echo cancellation performance is compromised. Different algorithms and structures will be compared to see how they are affected by this non-linear distortion. Wideband vocoder distortion is frequency dependent, so subband and fullband structures should be affected differently, as should different adaptive algorithms. This thesis will attempt to determine which adaptive filtering algorithms and structures are most effective for acoustic echo cancellation in the presence of wideband vocoder distortion.

1.5 Contributions

- *Analysis of residual echo levels in oversampled subband adaptive filters.* It is observed that under some conditions oversampled subband adaptive filters can produce residual echo levels below those of the linear time invariant Wiener filter. This behaviour is identified as resulting from the adaptive filter exploiting input signal colouration information to achieve output squared error levels below the Wiener MSE. This was previously observed for fullband filters with highly coloured inputs, here it is demonstrated for oversampled subband filters for a broader class of input signals. Bandpass filtering by the analaysis filterbank creates narrowband signals, then these signals are downsampled by a non-critical factor, the resulting subband signal is still highly coloured. It is observed that this colouration created by the oversampling process is sufficient to cause subband filters to exhibit the behaviour, even for white noise inputs. This work is presented in chapter 4.

- *Comparison of fullband and subband adaptive filters in a wideband acoustic echo cancellation framework.* Echo cancellation in the steady state, convergence speed, and tracking ability are compared for subband and fullband adaptive filters. Using simulated data subband echo cancellers are shown to offer better echo path tracking. These results are verified using experimental speech data recorded in a real changing echo environment. Using speech data and measured wideband acoustic echo path impulse responses, subband structures are shown to provide better convergence and echo cancellation when a doubletalk detector is present. These contributions are in chapters 5 and 6.

- *Assessment of the effects of wideband VoIP vocoder distortion on acoustic echo*

cancellation. The impact of vocoder distortion is compared for subband and fullband adaptive echo cancellers. The Improved Proportionate Affine Projection Algorithm (IP-APA) is proposed as an effective algorithm for echo cancellation in the vocoder distorted path. This contribution is in chapter 7.

- *Proposition of complexity reduction and stabilisation enhancements for the IP-APA.* The complexity reductions result from applying techniques from Fast Affine Projection algorithms, and the proposed online regularisation stabilises the algorithm for subband implementations and speech inputs. The modifications are presented in chapter 8.

1.6 Organisation

This thesis is divided into seven parts. Chapter 1 contains an introduction to echo cancellation and wideband VoIP communications. The thesis objectives are presented, and the thesis contributions are summarised.

Chapter 2 is a review of literature relevant to the thesis. Adaptive filtering algorithms used in later simulations are described in detail, and their relative advantages and disadvantages are discussed. Different approaches to subband adaptive filtering are reviewed, with special attention paid to the oversampled local-error adaptation structure employed in this thesis. Fullband and subband doubletalk detection are briefly reviewed, and the normalised cross-correlation doubletalk detection algorithm is explained. An overview of linear predictive voice coding and the G.722.2 vocoder are presented in order to understand the effects of vocoder distortion on acoustic echo cancellation. The review of related background material is provided to give unfamiliar readers the information necessary to interpret the simulation results presented in later chapters.

The simulation framework used in this thesis is described in chapter 3. A description of the apparatus used to measure acoustic impulse responses in real conference rooms is provided, and a number of measured wideband acoustic echo path impulse responses are presented and discussed. The performance metrics used to compare adaptive filtering algorithms are defined, and the parameters used in subsequent simulations are summarised.

In chapter 4 the phenomenon of non-linear, non Wiener behaviour in adaptive filters is examined. The non Wiener effects are observed in subband adaptive filters, and the relationship between the level of subband oversampling and the severity of the non-linear behaviour is explored. The reason for the behaviour is presented, and it is shown that the circumstances under which it occurs are different for fullband and subband adaptive filters. Practical applications of the effect and methods to prevent it are also considered.

In chapter 5 simulation results are presented comparing the performance of adaptive filtering algorithms and structures when the echo path is time-varying. Using measured echo path impulse and magnitude responses, it is shown that echo path changes do not affect all frequencies equally. Tracking and reconvergence ability is compared for fullband and subband echo cancellers in wideband and narrowband situations, using simulated echo path changes and experimental data recorded in a real changing echo environment.

In chapter 6 simulation results are presented that compare the convergence abilities of fullband and subband NLMS adaptive filters when there are near-end disturbances and a doubletalk detector is present. Double-talk detectors halt adaptation in the presence of high-level near-end disturbances, slowing the rate of convergence; when the magnitude of the disturbance varies with frequency subband and fullband adaptive filters are affected differently. Simulation results are used to compare the

differences in fullband and subband convergence for wideband and narrowband echo environments.

In chapter 7 the impact of wideband vocoder distortion on adaptive echo cancellers is studied. The non-linear distortion imposed by a vocoder depends on the content of the speech, and is not constant for all frequencies. How this distortion affects the echo cancellation performance of different adaptive echo cancellation structures is examined, and simulation results are presented comparing the performance of fullband and subband algorithms for undistorted and vocoder distorted echo paths.

In chapter 8 a stabilised, reduced complexity version of the improved proportionate affine projection algorithm (IP-APA) is proposed. The modified IP-APA uses techniques from the Gauss-Seidel fast affine projection algorithm to reduce the complexity, and an online regularisation to improve the stability and make the algorithm it more robust to near-end disturbances. Simulation results are presented that confirm the algorithm's fast convergence for speech inputs and good tracking ability for changing echo environments.

Chapter 9 summarises the findings of the thesis and proposes future extensions to the presented work.

Chapter 2

Background Review

This chapter provides an overview of previous work and relevant literature in the field of study of this thesis. A review of adaptive filtering algorithms is provided, including a detailed description of the algorithms used in the subsequent simulations. Subband adaptive filtering is also discussed with a focus placed on applications to echo cancellation. The subject of doubletalk detection is reviewed, and existing work related to subband doubletalk detection is discussed. Finally, the issue of vocoder distortion in echo cancellation is considered, providing a brief overview of linear predictive coding and a description of the G.722.2 vocoder.

2.1 Adaptive Filtering Algorithms

Considering the echo canceller of figure 1.1, the acoustic echo cancellation problem is formulated as follows.

$$\underline{x}(n) = [x(n), x(n-1), \ldots, x(n-N-1)]^T$$

$$\underline{h}(n) = [h_0, h_1, \ldots, h_L]^T$$

$$\underline{\hat{h}}(n) = [\hat{h}_0(n), \hat{h}_1(n), \ldots, \hat{h}_{N-1}(n)]^T$$

$$d(n) = \underline{h}^T(n)\underline{x}_L(n) \tag{2.1}$$

$$y(n) = d(n) + v(n) + b(n) \tag{2.2}$$

$$\hat{d}(n) = \hat{\underline{h}}^T(n)\underline{x}(n) \tag{2.3}$$

$$\begin{aligned} e(n) &= y(n) - \hat{d}(n) \\ &= (\underline{h}^T(n) - \begin{bmatrix} \hat{\underline{h}}^T(n) \\ \mathbf{0}_{L-N} \end{bmatrix})\underline{x}_L(n) + v(n) + b(n) \end{aligned} \tag{2.4}$$

Where n is the discrete-time index, $\underline{h}(n)$ is the length-L linear echo path impulse response vector, $\hat{\underline{h}}(n)$ is the length-N adaptive filter tap weight vector $\mathbf{0}_{L-N}$ is a zero-vector of length $L-N$ and $\underline{x}_L(n)$ is a vector of the previous L input samples. If $L = N$ it is possible for $\hat{\underline{h}}(n) = \underline{h}(n)$, resulting in total removal of the echoed signal $d(n)$ from the error signal $e(n)$, which is transmitted to the far end listener. However, in practice L is typically much larger than N. In wideband acoustic echo cancellation the room echo path impulse response can be thousands of samples long but the adaptive filter only models the first few hundred samples containing the majority of the impulse response energy. The unmodeled tail of the echo path contributes to the uncancelled residual echo.

It can be shown [17] that the linear filter which minimises the mean-square error (MSE), between the output and the desired signal, $\mathcal{E}\{|d(n) - \hat{d}(n)|^2\} = \mathcal{E}\{e^2(n)\}$, is given by:

$$\hat{\underline{h}}_{opt} = \mathbf{R}_{xx}^{-1}\underline{r}_{xd} \tag{2.5}$$

where $\mathbf{R}_{xx} = E\{\underline{x}(n)\underline{x}^H(n)\}$ is the correlation matrix of the input signal, and $\underline{r}_{xd} = \mathcal{E}\{\underline{x}(n)d^*(n)\}$ is the cross-correlation between the input and the desired signal. The solution in (2.5) is known as the Wiener solution. If the entire data sequence is available, the correlation matrix and cross-correlation vector can be calculated, and

the linear MSE optimal Wiener solution of equation (2.5) can be computed directly. In real-time echo cancellation, the the data is arriving sequentially, and the filter tap weights must be computed online. A number of algorithms have been derived to adapt the modelling filter coefficients so that the tap weight vector converges to the Wiener solution, and there is generally a tradeoff between algorithm complexity and speed of convergence. In real situations the echo path $\underline{h}(n)$ is time varying; acoustic echo paths can change rapidly in response to people moving or objects being moved in the near-end room, disrupting the echo reflecting paths. Adaptive acoustic echo cancellation therefore requires the use of algorithms that can converge quickly, but also rapidly track changes [16]. Since wideband acoustic impulse responses can be hundreds of milliseconds long, requiring thousands of taps to fully model, the algorithms must also have low computational complexity in order to be implemented on low-cost processors.

2.1.1 Recursive Least Squares Algorithm

The recursive least squares algorithm is an adaptive algorithm that recursively computes the least square estimate of the tap-weight vector. Beginning with an initial estimate, the weight vector is updated with the new sample data, using the update equation [17]:

$$\hat{\underline{h}}(n) = \hat{\underline{h}}(n-1) + \mathbf{R}_{xx}^{-1}\underline{x}(n)e^*(n)$$

The estimate of the correlation matrix inverse $\mathbf{R}_{xx}^{-1}$ is computed recursively to avoid the computationally expensive matrix inversion. The reference data vector $\underline{x}(n)$ is whitened by the inverse correlation matrix, so all of the frequencies of the tap weight vector are excited, leading to very fast convergence. The use of all previous inputs to recursively estimate the correlation matrix accelerates the convergence in a stationary

environment, but reduces the ability of the adaptive filter to track changes in $\underline{h}(n)$. To compensate for this, an exponential forgetting factor is used so that the estimation is based on a finite window of previous samples. Unfortunately small forgetting factors that enable the fast tracking required for acoustic echo cancellation can lead to instability for speech inputs [4]. In general, the tracking performance of RLS is not as robust as algorithms such as LMS [17]. The computational complexity of RLS is another disadvantage of the algorithm. The coefficient update of a length N RLS adaptive filter with real coefficients requires $C_{RLS} = 4N + 3N^2$ multiplications per sample period [18], so the computational burden grows quadratically as the filter length increases. The complexity and tracking performance of RLS combine to limit its applicability in acoustic echo cancellation applications, consequently it is not used in the simulations that follow.

2.1.2 Least Mean Square Algorithms

Each possible tap-weight vector has a unique MSE value associated with it, which is the expected value of the squared output error averaged over all possible input sequences. Since the MSE is a quadratic cost function, the set of possible MSEs forms a convex surface in the tap weight vector space, with a unique minimum corresponding to the Wiener solution. The convex shape of the error surface permits a gradient descent approach to finding the optimum tap weight vector. In a gradient search an estimate of the local shape of the error surface is used to adapt the tap weight vector down the error surface towards the minimum MSE solution.

The least-mean square (LMS) algorithm, first introduced by Widrow and Hoff [19], belongs to the family of stochastic gradient algorithms wherein the gradient estimate is based only on the current data vector and desired signal sample; this is equivalent to using the instantaneous squared error as the cost function. The LMS tap weight

update equation is given by

$$\underline{\hat{h}}(n+1) = \underline{\hat{h}}(n) + 2\mu\underline{x}(n)e^*(n), \tag{2.6}$$

where μ is a parameter which controls the size of the "step" taken by the weight vector at each iteration. The convergence time, stability, and steady state error of the LMS algorithm are dependent on the step-size parameter. Larger values of μ lead to faster convergence but, due to the noisy nature of the gradient estimate, larger steady-state error. Also, since the tap weight vector affects the output error, which in turn controls the adaptation, there is feedback in the adaptation process, so the algorithm can become unstable if μ is too large. A small μ effectively lowpass filters the fluctuations in the gradient estimate, keeping the algorithm stable and yielding less steady-state error, but slower convergence. In the absence of near-end disturbances, the LMS-adapted coefficient vector $\underline{\hat{h}}(n)$ converges to the Wiener solution $\underline{\hat{h}}_{opt}$ in the mean-square sense.

The magnitude of the tap weight vector adjustment in LMS update equation in (2.6) is directly proportional to the magnitude of the input vector $\underline{x}(n)$. When the input power is high, the gradient noise is magnified, resulting in slow convergence. Furthermore, the power of the input signal is generally not known a-priori, which makes it difficult to choose a step-size μ that will ensure stability. A solution to both of these problems is provided by the normalised LMS (NLMS) algorithm, which uses a time-varying step-size $\mu(n)$. The NLMS $\mu(n)$ is chosen based on the principle of minimum disturbance as the step-size that modifies $\underline{\hat{h}}(n)$ as little as possible while ensuring the a-posteriori error is zero, ie., $d(n) - \underline{\hat{h}}^H(n+1)\underline{x}(n) = 0$. It can be shown that the value of $\mu(n)$ that satisfies this principle is [17]:

$$\mu(n) = \frac{1}{2\underline{x}^H(n)\underline{x}(n)}.$$

This time varying step-size leads to the NLMS algorithm described in Algorithm 1,

where δ is a small positive value to avoid division by zero and μ is a constant step-size parameter chosen to control steady-state error; convergence occurs for NLMS if and only if $0 < \mu < 2$ [20]. The computational requirement for NLMS is $C_{NLMS} = 3 + 2N$ which, in contrast to RLS, grows linearly with the filter length. The NLMS algorithm is one of the most frequently used adaptive filtering algorithm for acoustic echo cancellation due to its robustness, ease of implementation, and low complexity.

Algorithm 1 (Normalised Least Mean Square (NLMS) Algorithm).

Initialisation

$\underline{x}(0) = \hat{\underline{h}}(0) = \underline{0}_N$

Adaptation

FOR $n \geq 0$

$$e(n) = y(n) - \hat{\underline{h}}^H(n)\underline{x}(n)$$

$$\hat{\underline{h}}(n+1) = \hat{\underline{h}}(n) + \mu \frac{1}{\underline{x}^H(n)\underline{x}(n) + \delta}\underline{x}(n)e^*(n)$$

ENDFOR

Notation

μ: step-size control parameter.

δ: small positive constant to prevent numerical instability.

2.1.3 Affine Projection Algorithm

Despite the advantages of the NLMS algorithm, it is well known to exhibit slow convergence for coloured input signals such as speech [20], [21]. The affine projection algorithm (APA) [22] is a generalisation of NLMS, developed to overcome this problem. Rather than using a single input signal vector APA uses an input signal matrix,

$\mathbf{X}(n)$, composed of the last P input vectors $\mathbf{X}(n) = [\underline{x}(n), \underline{x}(n-1), \ldots, \underline{x}(n-(P-1))]$. APA can be thought of as a tradeoff between RLS and NLMS, both in terms of complexity and convergence speed. While RLS uses a whitened gain vector composed of the input signal vector whitened by the full inverse correlation matrix, APA of order P uses a partially-whitened gain vector formed by multiplying the input signal matrix with a whitening matrix that is a stochastic estimate of the inverse $P \times P$ correlation matrix. Consequently, APA of order P can decorrelate and auto-regressive process of order P, thereby achieving faster convergence than NLMS [23]. The basic regularised APA is presented in Algorithm 2, where $\delta\mathbf{I}$ is a regularisation matrix, analogous to the parameter δ in NLMS, used to prevent numerical instability in the matrix inversion.

Algorithm 2 (Affine Projection Algorithm (APA)).

Initialisation

$\hat{\underline{h}}(0) = \underline{0}_N$

$\mathbf{X} = \mathbf{0}_{N \times P}$

Adaptation

FOR $n \geq 0$

$$\underline{e}(n) = \underline{y}(n) - \mathbf{X}^{\mathbf{T}}(n)\hat{\underline{h}}^*(n)$$

$$\hat{\underline{h}}(n+1) = \hat{\underline{h}}(n) + \mu\mathbf{X}(n)\left[\mathbf{X}^{\mathbf{H}}(n)\mathbf{X}(n) + \delta\mathbf{I}\right]^{-1}\underline{e}^*(n)$$

ENDFOR

Notation

$\delta\mathbf{I}$: diagonal regularisation matrix, to stabilise matrix inverse.

2.1.4 Fast Affine Projection Algorithms

The complexity of the order-P APA as presented in Algorithm 2 is $C_{APA} = 2NP + K_{inv}P^2$ multiplies per update, where K_{inv} is a constant representing the complexity of the matrix inverse [24]. While the complexity for a given value of P grows linearly with increasing filter length, it is still considerably higher than NLMS. The Fast Affine Projection (FAP) algorithm [24] uses simplifying approximations to reduce the complexity to $C_{FAP} = 2N + 20P$, which makes it comparable to NLMS for acoustic echo applications where $N >> P$. The first approximation involves the computation of the error vector $\underline{e}(n)$, which is reduced from the matrix multiplication of full APA to:

$$\underline{e}(n) \approx \begin{bmatrix} e(n) \\ \\ (1-\mu)\overline{\underline{e}}(n-1) \end{bmatrix} \tag{2.7}$$

where $\overline{\underline{e}}(n-1)$ consists of the upper $P-1$ elements of $\underline{e}(n-1)$; this approximation is valid for values of μ close to 1. Further complexity reduction is achieved by using a fast recursive approach to compute the correlation matrix inverse $\left[\mathbf{X}^{\mathbf{H}}(n)\mathbf{X}(n) + \delta\mathbf{I}\right]^{-1}$. FAP also uses an alternative coefficient vector that maintains the fidelity of the error signal $e(n)$, but can be updated more efficiently than the traditional APA tap weight vector.

While it achieves the goal of computational complexity reduction, fast recursive estimation of the correlation matrix inverse causes FAP to suffer from numerical instability problems. To resolve this issue, the conjugate-gradient fast affine projection (CG-FAP) algorithm [25] was developed. In CG-FAP the matrix inverse is reduced to a series of P linear equations which enables the inverse to be computed using the conjugate gradient method. The Gauss-Seidel fast affine projection (GS-FAP) algorithm [26] uses the same principles as CG-FAP, except that it employs the lower

complexity Gauss-Seidel method, described in [27], to compute the matrix inverse. Both algorithms are have a higher asymptotic computational complexity than FAP, CG-FAP requires $2N+2P^2+9P+1$ real multiplications per iteration, while GS-FAP requires $2N + P^2 + 4P - 1$, however, in [25] it is shown that for projection orders $P \leq 8$, CG-FAP is less computationally intensive than FAP. The GS-FAP algorithm is presented in Algorithm 3. The full APA, as presented in Algorithm 2 is used for the simulations that follow, unless otherwise stated.

Algorithm 3 (Gauss-Seidel Fast Affine Projection Algorithm (GS–FAP)).

Initialisation

$\hat{\underline{h}}(0) = \underline{x}(0) = \underline{0}_N \underline{\eta}(0) = \underline{r}_{xx}(0) = \underline{0}_P$

$\mathbf{R}(0) = \delta\mathbf{I} \qquad \underline{b} = [1\ \underline{0}_{P-1}^T]^T$

$\underline{p}(0) = \underline{b}/\delta$

Adaptation

FOR $n \geq 0$

$\underline{r}_{xx}(n) = \underline{r}_{xx}(n-1) + x(n)\underline{\xi}^*(n) - x(n-N)\underline{\xi}^*(n-N)$

Update $\mathbf{R}(n)$ using $\underline{r}_{xx}$

Solve $\mathbf{R}(n)\underline{p}(n) = \underline{b}$ using one GS iteration

$\hat{\underline{h}}(n) = \hat{\underline{h}}(n-1) + \mu\eta_{P-1}(n-1)\underline{x}(n-P)$

$e(n) = y(n) - \hat{\underline{h}}^H(n)\underline{x}(n) - \mu\overline{\underline{\eta}}^H(n)\tilde{\underline{r}}(n)$

$\underline{\epsilon}(n) = e(n)\underline{p}(n)$

$\underline{\eta}(n) = [0\ \overline{\underline{\eta}}^T(n-1)]^T + \underline{\epsilon}(n)$

ENDFOR

Notation

$\delta\mathbf{I}$: diagonal matrix, to initialise $\mathbf{R}(n)$.

$\underline{\xi}(n)$: uppermost P elements of $\underline{x}(n)$.

$\eta_{P-1}(n)$: lowermost element (scalar) of the vector $\underline{\eta}(n)$.

$\overline{\underline{\eta}}(n)$: uppermost $P-1$ elements of $\underline{\eta}(n)$.

$\underline{r}(n)$: left column of the autocorrelation matrix $\mathbf{R}(n)$.

$\tilde{\underline{r}}(n)$: lowermost $P-1$ elements of $\underline{r}(n)$.

2.1.5 Individual Step-Size Algorithms

The NLMS algorithm does not make any assumptions about the nature of the system being modelled, making it a suitable, but not necessarily the most efficient, algorithm for all types of system identification problems. Since acoustic echo path impulse responses tend to have similar time domain characteristics, algorithms have been developed to exploit a-priori knowledge about the echo path shape in order to increase convergence or tracking speed. The Exponential Step-Size NLMS (ES-NLMS) algorithm [28] was derived based on extensive measurements of office room acoustic echo path impulse responses. It was observed that the variance (energy) of acoustic echo path samples decreases in an approximately exponential fashion. To take advantage of this observation the ES-NLMS algorithm uses individual step-sizes for each tap, rather than a single μ for the entire adaptive filter, and the step sizes decrease exponentially with increasing tap index. A downside of the ES-NLMS algorithm is that the step-size profile is fixed, so the reverberation characteristics of the room in which the echo canceller is operating must be known.

By using individual step-sizes that are chosen adaptively, the shape of the echo path can be exploited without a-priori knowledge of the characteristics of the acoustic environment. An example of an individual step-size algorithm that adaptively modifies the step-sizes is the Proportionate NLMS (PNLMS) algorithm [29]. In PNLMS the available adaptation gain is distributed in proportion to the tap energy. This is achieved by using a time-varying $N \times N$ step-size matrix with diagonal elements that are proportional to the absolute value of the corresponding adaptive filter tap weight. While PNLMS offers fast convergence for sparse echo paths (close to a delta function), it is slower than NLMS for dispersive echo paths, making it unsuitable for acoustic echo applications. In [30] this poor performance was explained by the fact that during initial adaptation the tap estimates are inaccurate and distributing the

adaptation energy based on those estimates results in poor convergence. To reduce the effect of inaccurate tap estimates, the improved PNLMS (IP-NLMS) algorithm in [30] changes the energy distribution rule used by PNLMS by adding a fixed element to the step-size matrix. This fixed component smooths the energy distribution thereby improving the convergence when the coefficient estimate is not accurate. The IP-NLMS algorithm is presented as Algorithm 4, where α is a parameter that controls the ratio of fixed to proportionate step-size. For $\alpha = -1$ the algorithm reduces to standard NLMS, and for $\alpha = 1$ it behaves like PNLMS. According to [30], good choices of α are 0 or -0.5, and a value of $\alpha = -0.5$ is used for the simulations in the sequel.

Individual step-size algorithms can also be combined with the APA to further improve convergence speed for coloured inputs [31], producing algorithms that exploit both the input signal statistics and the impulse response shape to achieve fast convergence and good tracking for white or coloured inputs. An example of such an algorithm, the Improved Proportionate Affine Projection Algorithm (IP-APA), is listed in Algorithm 5. The convergence and tracking benefits of using proportionate step-sizes comes at the cost of additional complexity. For IP-NLMS and IP-APA, the computation of the step-size matrix requires N multiplies to calculate $\|\underline{\hat{h}}(n)\|_1$ and an additional N multiplies to compute the individual step-sizes. Furthermore, the calculation of the power normalisation for IP-NLMS requires an additional N multiplications compared to NLMS, and the calculation of the whitening matrix in IP-APA requires an additional PN multiplications compared to APA.

Algorithm 4 (Improved Proportionate NLMS (IP-NLMS)).

Initialisation

$\underline{x}(0) = \hat{\underline{h}}(0) = \underline{0}_N$

$\mathbf{A}(0) = \mu\mathbf{I}$

Adaptation

FOR $n \geq 0$

$$e(n) = y(n) - \hat{\underline{h}}^H(n)\underline{x}(n)$$

$$\mathbf{A}(n) = \mathrm{diag}\{\mathrm{a_0(n)}, \ldots, \mathrm{a_{N-1}(n)}\}$$

where

$$a_l(n) = \frac{\kappa_l(n)}{\|\underline{\kappa}\|_1}$$

$$= \frac{1-\alpha}{2N} + (1+\alpha)\frac{|\hat{h}_l(n)|}{2\|\hat{\underline{h}}(n)\|_1 + \epsilon}, l \in \{0, 1, \ldots, N-1\}$$

$$\hat{\underline{h}}(n+1) = \hat{\underline{h}}(n) + \mu\frac{\mathbf{A}(n)\underline{x}(n)}{\underline{x}^H(n)\mathbf{A}(n)\underline{x}(n) + \delta}e^*(n)$$

ENDFOR

Notation

ϵ: small positive constant to avoid numerical instability.

α: parameter controlling ratio of fixed to adaptive step-size.

Algorithm 5 (Improved Proportionate APA (IP-APA)).

Initialisation

$\underline{\hat{h}}(0) = \underline{0}_N$

$\mathbf{X} = \mathbf{0}_{N \times P}$

$\mathbf{A}(0) = \mu \mathbf{I}$

Adaptation

FOR $n \geq 0$

$$\underline{e}(n) = y(n) - \mathbf{X}^{\mathbf{T}}(n)\underline{\hat{h}}^{*}(n)$$

$$\mathbf{A}(n) = \mathrm{diag}\{\mathrm{a_0(n)}, \ldots, \mathrm{a_{N-1}(n)}\}$$

where

$$a_l(n) = \frac{1-\alpha}{2N} + (1+\alpha)\frac{|h_l h(n)|}{2\|\underline{\hat{h}}(n)\|_1 + \epsilon}, l \in \{0, 1, \ldots, N-1\}$$

$$\underline{\hat{h}}(n+1) = \underline{\hat{h}}(n) + \mu \mathbf{A}(n-1)\mathbf{X}(n)[\mathbf{X}^H(n)\mathbf{A}(n-1)\mathbf{X}(n) + \delta \mathbf{I}]^{-1}\underline{e}^{*}(n)$$

ENDFOR

Notation

$\delta \mathbf{I}$: diagonal regularisation matrix, to stabilise matrix inverse.

ϵ: small positive constant to avoid numerical instability.

α: parameter controlling ratio of fixed to adaptive step-size.

2.2 Subband Echo Cancellation

Subband adaptive filtering was proposed in [32] and [33] as a method to overcome the problems associated with wideband acoustic echo cancellation, namely the high computational complexity associated with modelling a wideband echo path and the slow convergence of LMS adaptive filters with large numbers of taps and for coloured inputs such as speech [16]. In addition to reducing the complexity and speeding

convergence of adaptive filters, subband processing offers flexibility to the designer of an acoustic echo cancellation system; different parameters can be chosen for each subband.

The subband approach makes echo cancellation more computationally feasible in several ways. First, the bandlimited subband signals can be downsampled, thereby allowing the filter adaptation to occur at a lower rate than the system sampling rate. Even for modest downsampling factors, the reduction in processing rate can be enough to overcome the additional computational cost associated with the analysis and synthesis filtering required to decompose and reconstruct the subband signals. Second, if the subband filters are adapted independently, the adaptation can be performed in parallel on a multi-processor system using lower cost DSPs. This is the approach taken in [34] and [35] where subband systems were implemented on general purpose DSPs, the first using an 8-band system and operating at a system sampling rate of 8 kHz, the second using a 64-band system for wideband echo cancellation at 16 kHz sampling rate.

In addition to making the problem more tractable, the subband configuration may also improve the performance of an echo canceller, especially in the wideband case. The adaptive algorithm step-size can be chosen separately based on the energy in each subband, leading to faster convergence and better tracking in the bands where there is more signal energy. More importantly however, for highly correlated input signals, such as speech, the subband signals will be more spectrally flat than the fullband signal. For algorithms such as NLMS for which the convergence speed depends on the spectral flatness of the input, subbanding can result in faster convergence [36].

An example of the convergence improvement offered by subband adaptive filters can be seen in figure 2.1, where the system distance convergence of two-times oversampled subband NLMS is compared to fullband NLMS and fullband APA with $P = 3$.

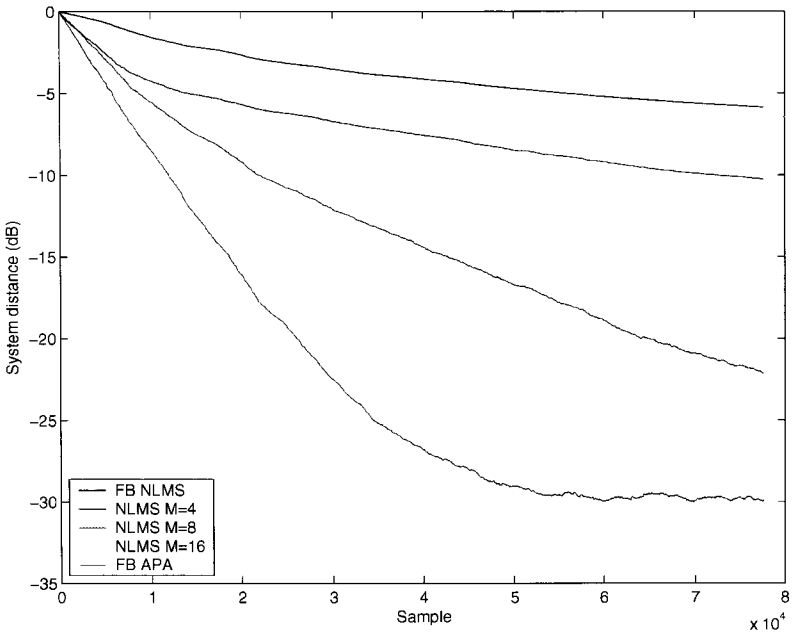


Figure 2.1: System distance convergence of fullband and subband adaptive algorithms for coloured input.

The highly coloured input is formed by filtering white noise through an all-pole filter obtained from 16^{th} order linear prediction of a voiced speech segment (see section 2.4.1). The convergence of the fullband NLMS is very slow as a result of the severe spectral coloration. As the number of subbands is increased, individual bands become more spectrally flat, and the rate of convergence of the overall system increases, such that the initial convergence of the $M = 16$ subband system approaches that of APA. The issue of convergence speed for correlated inputs is magnified in a wideband setting, as the spectral coloration of wideband speech signals is greater than that of narrowband, and the higher sampling rate necessitates more adaptive filter taps to model the same echo path length.

The flexibility that subband adaptive filtering provides arises in the way it allows for different processing in each subband. In [1] it is suggested that the number of taps in each subband can be adjusted so that fewer taps are assigned in the bands with little

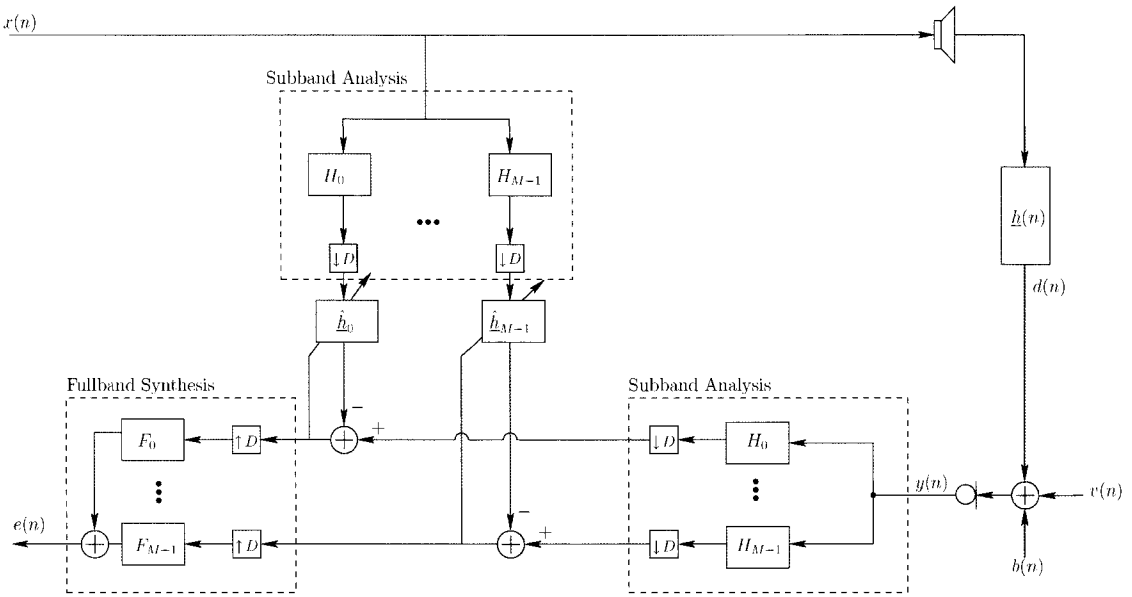


Figure 2.2: Subband echo canceller after [1]

echo energy or where the echo path has a shorter reverberation time, thereby reducing the complexity and improving the convergence speed without degrading the overall performance. A similar conclusion is reached in [37] where, based on listening tests, it is shown that higher frequency echoes are less perceptually significant, so subband structures can allocate fewer taps to those bands without audible degradation of the echo cancellation performance.

2.2.1 Subband Adaptive Filter Structures

In a subband echo cancellation system, as shown in figure 2.2 the the input signal $x(n)$ and the reference signal $y(n)$ are split into M subbands by a bank of analysis filters $\underline{H}_m$, $m \in \{0, \ldots, M-1\}$, and downsampled by a factor $D \leq M$. The subband adaptive filters $\underline{\hat{h}}_m(n)$ are adapted using the local error signals $e_m(n)$, and the fullband error signal $e(n)$ is obtained by upsampling the subband error signals by D, and filtering through the bank of synthesis filters $\underline{F}_m$. With the local error structure, in

order for the fullband error to be driven to zero the subband errors must also be zero. When this occurs, the system satisfies the condition [34]:

$$\sum_{m=0}^{M-1} \underline{\hat{h}}_m(z^D)\underline{H}_m(z) = \sum_{m=0}^{M-1} \underline{H}_m(z)\underline{h}(z). \tag{2.8}$$

An alternative approach to subband adaptive filtering involves synthesising the fullband echo replica $\hat{d}(n)$, computing the fullband error signal $e(n) = y(n) - \hat{d}(n)$, splitting the fullband error into subbands, and using the subbanded global error to adapt the subband adaptive filters. In this configuration, a zero fullband error signal does not imply zero subband errors, instead the approach leads to synthesis-dependent solutions that satisfy the weaker condition [34]:

$$\sum_{m=0}^{M-1} \underline{F}_m(z)\underline{\hat{h}}_m(z^D)\underline{H}_m(z) = \sum_{m=0}^{M-1} \underline{F}_m(z)\underline{H}_m(z)\underline{h}(z). \tag{2.9}$$

A primary advantage of the global error approach is that it is capable of driving the fullband error to zero without requiring the subband error to be zero, consequently it is capable of cancelling modest reconstruction errors in the filterbanks. The global error method has not been examined as much as the local error method because of two main drawbacks: error components outside of the frequency range of a given subband act as noise in the subband signal, slowing the adaptation; and the subband error signals used for adaptation are delayed by the synthesis filterbank, degrading the tracking ability of the adaptive filter and reducing its effectiveness for acoustic echo cancellation [16]. All of the simulations in this thesis use the synthesis-independent configuration of figure 2.2.

The early analyses of subband echo cancellation systems in [34] and [38] demonstrated that for a critically downsampled subband system (ie., where $D = M$) using non-ideal filterbanks, the output error signal contains significant aliasing distortion in the transition band region. The reason for this is as follows. In an analysis filterbank

where only adjacent filters overlap, the filters have a magnitude of 0.5 or -3 dB at the crossover point, in order to ensure that the filterbank is power complimentary and the overall filterbank frequency response is flat. Since the crossover point of a critically sampled filterbank is the Nyquist frequency, and the filters are non-ideal, all of the spectral content in the transition band is aliased back into the baseband. If no processing is performed on the subband signals it is be possible to compensate for the aliasing with an appropriately designed synthesis filterbank, however in subband echo cancellation processing is performed as the echo signal is removed at the subband level. The aliasing in a critically sampled subband acts as high level noise, disrupting the adaptation and resulting in poor echo cancellation in the transition band regions.

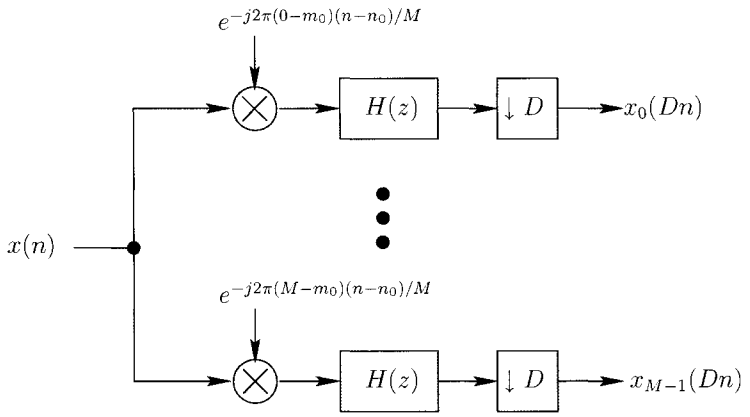
One option to deal with the aliasing problem is to use adaptive cross-filters as described in [38]. The cross-filters compensate for the non-ideality of the filterbank by including some energy from adjacent bands in the subband error signal. Even though the cross-filters can be shorter than the main filters, they still present an added computational burden. The most common approach to dealing with subband signal aliasing is to avoid aliasing distortion by using non-critical downsampling (ie., $D < M$), this approach was first suggested in [1]. In an M band filterbank, where the subband signals are downsampled by a factor $D < M$, the crossover point between adjacent filters is π/M while the Nyquist frequency of the downsampled signal is π/D, allowing the transition band of the analysis and synthesis filters to be $(\pi/D - \pi/M)$ wide without introducing aliasing from the transition band. As an additional advantage, this wide transition band permits the use of shorter analysis and synthesis filterbanks, reducing the signal path delay. This is especially important in a VoIP setting where any delay above the significant network, buffering and queueing delays is undesirable. Naturally the lower downsampling factor results in an increase in computational complexity; the choice of decimation factor is therefore a tradeoff between

aliasing in the subband signals, signal path delay and computational complexity.

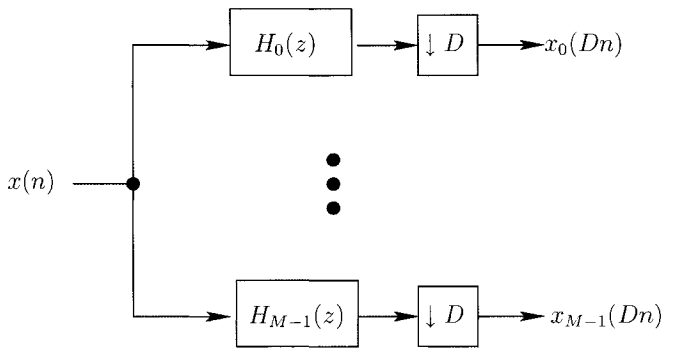
2.2.2 Filterbank Design

The topic of filterbank design is extensive and complex, and has been treated in detail in previous works such as [39] and [2], and will not be discussed here. However, the structure of the filterbank is an important design decision, and will be briefly addressed.

A common structure for filterbanks is the uniform modulated filterbank, where the input signal is split into a set of subband signals each covering an equal fraction of the frequency spectrum. There are two complimentary ways of visualising a uniform modulated filterbank. The first way, depicted in figure 2.3(a) is to see the subbanding process as consisting of distinct modulation and filtering steps. For each branch of the analysis portion of the uniform modulated filterbank the input signal is frequency translated so that the centre of the desired frequency band is at $\omega = 0$, the modulated signal is then lowpass filtered and downsampled to create a subband signal at a reduced sampling rate. On the synthesis side, the fullband signal is reconstructed by upsampling the subband signals, passing them through an anti-imaging lowpass filter, de-modulating them from the baseband to form a bandpass signal, and summing the bandpass signals. An equivalent way of viewing the process, depicted in figure 2.3(b) is to combine the filtering and modulation steps into a bank of modulated bandpass filters $H_m(z),\ m \in \{0, 1, \ldots, M-1\}$, evenly spaced across the frequency spectrum. The bandpass filters are derived by convolving a lowpass prototype filter with a modulation kernel. The uniform modulated structure is desirable as the design of the entire bank of filters reduces to designing the lowpass prototype. The choice of modulation kernel affects the frequency positioning of the filters and determines whether the resulting subband signals are real or complex. A popular choice is the Discrete



(a) Modulation and lowpass filtering interpretation.



(b) Bandpass filterbank interpretation.

Figure 2.3: Alternate interpretations of a uniform GDFT analysis filterbank, after [2].

Fourier Transform (DFT) kernel W_M^{-mn}, where $W_M = e^{j(2\pi m/M)}$, which produces a bank of complex filters centred at $\omega_m = 2\pi m/M$, such that the $m = 0$ channel centred is at $\omega = 0$. The Generalised DFT (GDFT) extends DFT-modulated filterbanks to allow the $m = 0$ band centre to be other locations than the origin. The GDFT kernel is given by $W_M^{-(m+m_0)(n+n_0)}$ where n_0 and m_0 are the time and frequency offsets controlling the frequency bin centres and phase offsets of the modulated bandpass filters [39]. The structure used in this thesis is a uniform GDFT modulated filterbank, implemented as in figure 2.3(b).

The analysis and synthesis filters used in this thesis are near perfect reconstruction GDFT modulated filters designed using a MATLAB program based on the work in [40]

and available at [41]. The program employs an iterative least squares approach to design a prototype filter that minimises an error function consisting of a weighted sum of the analysis-synthesis reconstruction error and the stopband energy. In this way a tradeoff can be obtained between perfect reconstruction and in-band aliasing. The prototype filter has real valued coefficients and exhibits linear phase. The GDFT frequency offset is $m_0 = 1/2$, which places the bin centres at $w_m = 2\pi m/M + \pi/M$. This is known as odd-channel stacking and it allows the real frequency range $\omega = [0, \pi]$ to be covered with a bank of $M/2$ evenly spaced filters. The time offset of $n_0 = (L_p - 1)/2$, where L_p is the length of the prototype filter, ensures that the overall analysis-synthesis system is linear phase [40]. Figure 2.4 presents the magnitude response of an example GDFT modulated oversampled filterbank for the case of $M = 8$ and $D = 4$. The prototype filter has $L_p = 64$ coefficients and was designed to have a stopband attenuation of 90 dB as in [40]. It should be noted that fast versions of oversampled GDFT filterbanks exist that employ a polyphase factorisation and the Fast Fourier Transform (FFT) to achieve an overall filterbank complexity of $C_{fast} = \frac{1}{D}(4M \log_2 M + 6M + L_p)$ real multiplications per sample [18]. For simplicity, this thesis uses the direct form implementation with M complex linear phase bandpass filters which requires $C_{direct} = 4M(L_p/2)$ real multiplications per sample.

If complex signal processing is not desirable, there are several possible methods to obtain real-valued subband signals including using a bank of bandpass filters or a single-sideband modulated filterbank. The allowable downsampling rate of a bank of real bandpass filters is limited by the bandpass sampling theorem and is therefore is a function not only of the bandwidth, but also the upper cutoff frequency. A bandpass signal can be represented by uniform sampling at the rate [42]:

$$F' = \frac{2f_u}{\lfloor \frac{f_u}{B} \rfloor}$$

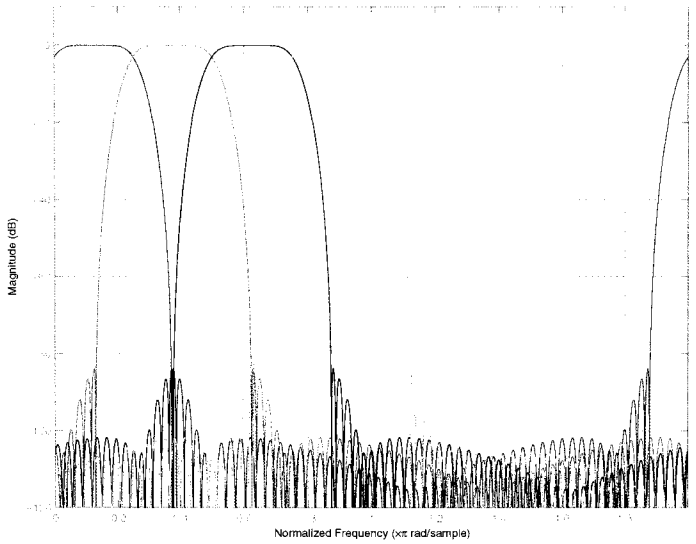


Figure 2.4: GDFT modulated analysis filterbank, $M = 8$, $D = 4$

where f_u is the upper frequency and B is the bandwidth. For integer decimation factors, the band-edge must be at an integer multiple of the sampling rate divided by the B. For oversampled filterbanks, this restriction brings about the need for non-uniform subbands [40].

To obtain real-valued subband signals while avoiding the issues associated with bandpass sampling, the bandpass signals may be modulated to the baseband prior to downsampling. Real modulated subband signals can be constructed from the complex signals produced by GDFT modulated filterbanks by: frequency shifting the complex baseband signal so that the spectrum lies entirely in positive frequencies; and taking the real part of the modulated signal, thereby reflecting the spectrum about $\omega = 0$ and creating a real signal with twice the bandwidth of the original complex signal. The result is known as a single-sideband (SSB) signal [39]. The act of twinning the spectrum and doubling the bandwidth of the subband signal halves the allowable downsampling factor, however real calculations are less computationally intensive than complex calculations. In [43] it is shown that for the LMS algorithm,

the computational complexity of a SSB implementation is higher than a DFT filter-bank version due to the extra computational burden of creating the SSB signal from the complex subband signal. However, it is been argued in [18] that for algorithms such as the affine projection algorithm that have a large overhead calculation, a SSB implementation may be more efficient.

2.2.3 Steady State Performance of Subband Echo Cancellers

The echo cancellation performance of all adaptive filters is limited by background noise, undermodelling of the echo path impulse response [44] and non-linear distortion from sources such as the hands-free terminal loudspeaker and microphone [45]. In the case of subband echo cancellers there are additional limits posed by aliasing distortion [41] and the need to model non-causal subband impulse responses [34]. According to [16] undermodelling of the echo path and aliasing distortion caused by non-ideal analysis filterbanks are the greatest sources of excess MSE, and the MSE performance of the system is dominated by the larger of the two until the MSE is small enough that they both contribute.

The effect of echo path undermodelling can be seen in figure 2.5 where the steady state ERLE for white noise trained echo cancellers is plotted as a function of the number of adaptive filter taps. Since the impulse response used for the simulation has 4000 samples and no noise was added, the ERLE of the fullband system approaches infinity (MSE goes to zero) for adaptive filter lengths greater than 4000. In the subband case, the analysis filters have a stopband attenuation of 90 dB, so the ERLE is initially dominated by undermodelling of the echo path and therefore grows at the same rate as the fullband system. Once the number of taps exceeds approximately 3500, the error incurred by undermodelling of the echo tail becomes small, and the ERLE growth rate slows, eventually reaching a maximum of approximately 43 dB for

4000 taps. This ERLE limit is not caused by aliasing distortion, but by the need to model non-causal taps. There is an early mention of this issue in [34] an in-depth analysis is provided in [46].

The need to model non-causal taps arises when a fullband echo path is approximated by a set of subband filters. Each of the subband magnitude responses is a frequency windowed (bandpass filtered), and frequency expanded (decimated) version of the fullband magnitude response. In the time domain this frequency windowing corresponds to convolution with a modulated two-sided $\mathrm{sinc}(x)$ function, which results in a non-causal subband impulse response. Not modelling the non-causal taps is essentially undermodelling the front of the impulse response rather than the tail, and the effect on ERLE is the same. Some of the non-causal taps can be accounted for by delaying the desired signal relative to the reference signal, effectively increasing the flat delay of the echo path. In VoIP systems where the flat delay is already high, the non-causal taps can be modelled adequately without added delay. Figure 2.6 shows the effect that modelling these non-causal taps has on the achievable ERLE, 4300 taps were used to ensure that the tail remained fully modelled as the desired signal was increasingly delayed. The rate of ERLE increase begins to decay as the limit of approximately 70 dB is reached where the effects of aliasing distortion mix with the effects of undermodelling. In summary, for a noise and distortion-free subband echo canceller, undermodelling of both ends of the impulse response and aliasing distortion combine to limit the achievable ERLE so that for a fully modelled system, a fullband structure will always outperform an equivalent subband structure in the steady state. In practice however, background noise and signal distortion likely the achievable ERLE before these limitations become apparent.

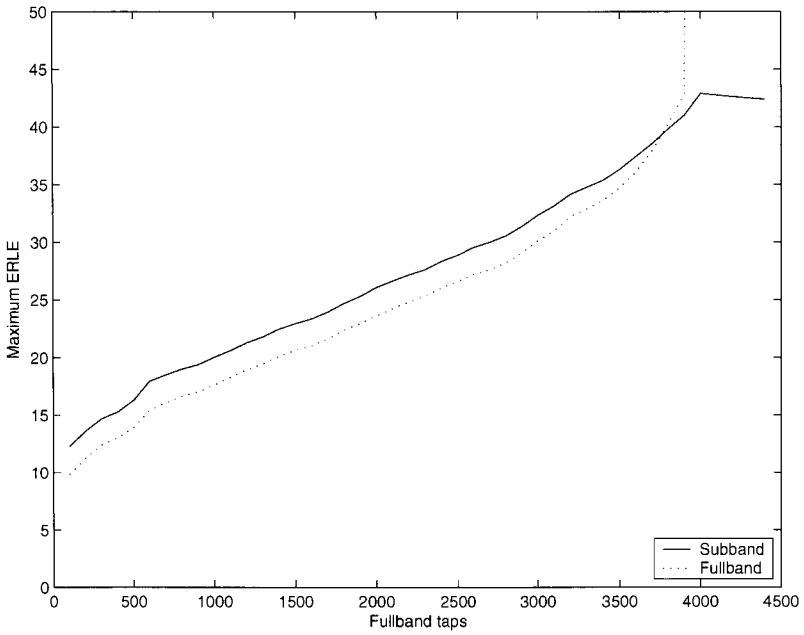


Figure 2.5: Effect of adaptive filter length on achievable ERLE. The solid line is for the subband system with and the dotted line is the fullband structure.

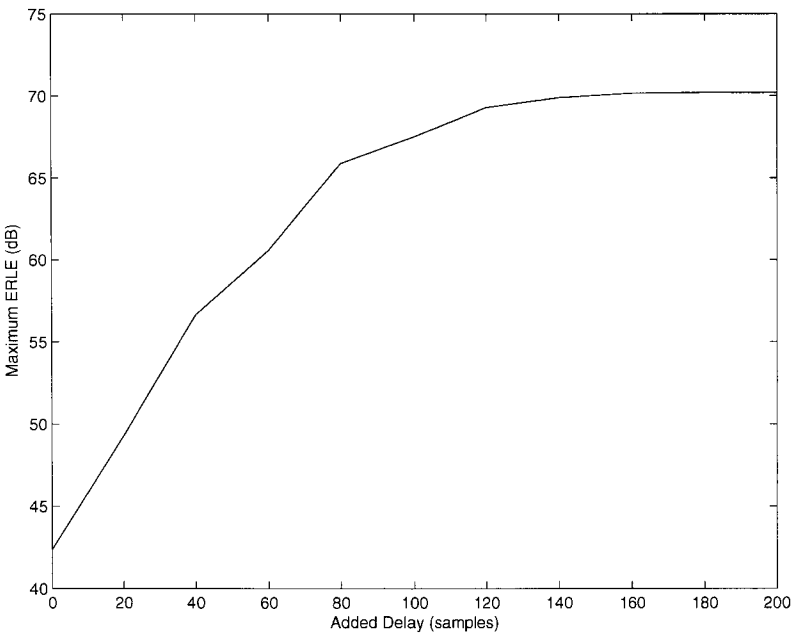


Figure 2.6: Effect of delaying reference signal to model non-causal taps on achieveable ERLE

2.3 Doubletalk Detection

When both the near-end and far-end users of a telephone link are speaking at the same time, the situation is known as doubletalk. During doubletalk, the near end speech acts as a high-level uncorrelated noise source and can result in mis-convergence of the adaptive filter. A doubletalk detector must be used to detect the presence of near-end speech and halt or slow the filter adaptation while the near-end talker is active.

The simplest and earliest doubletalk detection algorithm is known as the Geigel algorithm [47]. The Geigel algorithm compares the levels of the current near-end sample, and a history of previous far-end samples and declares doubletalk when:

$$\xi = \frac{|y(n)|}{\max\{|x(n-1)|, |x(n-2)|, \ldots, |x(n-N)|\}} > T. \tag{2.10}$$

In [47] a threshold value of $T = 0.5$ is used, based on the assumption that the hybrid coupling loss is approximately 6 dB. The Geigel algorithm is frequently used in network echo cancellation, but it does not perform well in acoustic echo situations, where the coupling loss is much more variable.

A more general and robust algorithm that does not assume a fixed coupling loss is the normalised cross-correlation algorithm [48] which uses the decision variable

$$\begin{aligned} \xi &= \sqrt{\underline{r}_{xy}^H (\sigma_y^2 \mathbf{R}_{xx})^{-1} \underline{r}_{xy}} \\ &= \sqrt{\frac{\underline{r}_{xy}^H \underline{h}}{\sigma_y^2}} \end{aligned} \tag{2.11}$$

Where σ_y^2 is the near end signal power and $\underline{r}_{xy} = \mathcal{E}\{\underline{x}(n)y^*(n)\}$ is the correlation between the input signal vector and the near end signal. The algorithm works in the following way: assuming the echo path $\underline{h}$ to be linear, stationary and of finite length

L, when no doubletalk is present the microphone signal $y(n)$ consists only of echo:

$$y(n) = \underline{h}^H \underline{x}_L(n)$$

and therefore

$$\sigma_y^2 = \underline{h}^H \mathbf{R}_{xx} \underline{h} \tag{2.12}$$

Furthermore, since $\underline{h}$ is linear

$$\underline{r}_{xy} = \mathcal{E}\{\underline{x}(n) y^*(n)\} = \mathbf{R}_{xx} \underline{h}$$

Which gives an alternate form of (2.12)

$$\sigma_y^2 = \underline{r}_{xy}^H \mathbf{R}_{xx}^{-1} \underline{r}_{xy} \tag{2.13}$$

Taking the square root of the ratio of (2.13) and (2.12) gives (2.11). In the absence of doubletalk the numerator and denominator of (2.11) will be equal, giving the decision variable a value of 1. When doubletalk is present the denominator will include the power of the near end speech in addition to the echo power, so the decision variable will take on a value less than 1. In real situations local background noise will also corrupt the echo power estimate in the denominator. In that case the threshold value can be adjusted to compensate for near end background noise, or a noise offset can be adaptively calculated and added to the numerator, as in [49]. In practical implementations the approximation $\hat{\underline{h}}(n) \approx \underline{h}$ is used, based on the assumption that the coefficients have converged to a neighbourhood of the Weiner solution. The signal power and cross-correlation vector, σ_y^2 and $\underline{r}_{xy}$ are assumed to be stationary, or at least slowly varying, and are typically estimated using a sliding window average. A hangover period, T_{hold}, is also employed to account for the noisy behaviour of the

doubletalk detection variable [50]; if doubletalk is declared, adaptation is halted for at least T_{hold}.

In [50] numerous methods for detecting doubletalk in acoustic echo cancellers are discussed and compared using an objective measure. It is shown that the normalised cross-correlation approach of [48] provides the most robust and accurate detection, consequently this is the doubletalk detection algorithm employed in this thesis.

2.3.1 Subband Doubletalk Detection

As with adaptive filtering algorithms, all doubletalk detection algorithms can be used in a subband system. Also, just as subband adaptive filters can operate using either global or local error signal, subband doubletalk detectors can use either global or per-subband doubletalk decisions. In [51] two subband doubletalk detection configurations are studied, in both configurations the decision to declare doubletalk is a global decision based on a combination of the local decisions from the subband doubletalk detectors. In the first configuration the decision for all bands is based on the decision of the subband detector found to have the best detection performance, in this case the lowest frequency subband. In the second configuration the decision for all bands is formed by taking a sum of all of the local subband decision variables weighted by the signal power in the corresponding subband. It was demonstrated that, compared to the fullband configuration, the "optimal subband selection" method, where only the decision from the lowest frequency band was used, yielded a better probability of detection for the same probability of false alarm. The weighted sum method also performed better than the fullband scheme, but by a lesser amount. The optimal subband selection method requires an a-priori selection of which subband will be used for the decision, rendering it vulnerable if narrowband noise degrades the SNR in the chosen band.

The authors of [51] also mentioned, but did not investigate, the possibility of having the subband doubletalk detectors independently control their respective subbands. This is the approach taken in [52], where local decision doubletalk detectors are used in a critically decimated cosine modulated subband echo cancellation structure. To compensate for the aliasing created by critical sampling, the subband error and reference signals and the gradient estimates incorporate some adjacent-channel components. The doubletalk detection variable in equation (2.11) is also modified to include adjacent-band terms. It was found that when doubletalk was present during the initial convergence period, the fullband doubletalk detector halted the adaptation more frequently than the subband detectors, resulting in faster convergence of the subband structure, and a 4.2 dB higher ERLE during the convergence period. The faster convergence was attributed to the doubletalk signal energy not being present in all bands during the doubletalk period, so some bands continued to adapt while the global decision structure was halted entirely. This effect would likely be more prominent in a wideband scenario, where the spectral distribution of the doubletalk energy is even less uniform.

2.4 Vocoder Distortion

2.4.1 Linear Predictive Coding

Human speech production can be roughly modelled as an excitation source in the throat driving a time-varying filter representing the vocal tract. Many low bit-rate voice coding systems achieve high quality speech transmission at significantly reduced bit-rates by dividing a speech signal into frames, and fitting each frame of data to this simplified model. By transmitting only the vocal tract model parameters and a

description of the excitation signal, rather than the entire sequence of speech samples, the data rate of the speech signal is greatly reduced.

In linear predictive (LP) coding the vocal tract filter is modelled as an all-pole filter. To regenerate the speech signal from the transmitted parameters, the all-pole analysis filter is driven with an excitation signal. In the z-domain, the speech signal $\hat{s}(n)$, produced by the input excitation signal $x(n)$, is given by [53]:

$$\begin{aligned} \hat{S}(z) &= \frac{g}{1+A(z)}X(z) \\ &= \frac{g}{1+\sum_{k=1}^{p} a_k z^k}X(z), \end{aligned} \tag{2.14}$$

where upper-case letters denotes the z-transform, g is the gain of the vocal tract, the coefficients a_k of the polynomial $A(z)$ are known as the predictor coefficients, and p is the prediction order. The term linear prediction is used because each reconstructed speech sample is predicted from a linear combination of the p previous samples as $\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k)$. The LP coefficients a_k are selected using a least mean squares criterion to minimise $e^2(n) = (s(n) - \hat{s}(n))^2$, the squared error between the reconstructed and original speech sample values.

The LP residual signal given by $r(n) = s(n) - \hat{s}(n)$ is the ideal excitation, as it perfectly reproduces the original speech, however transmitting the entire residual signal requires too much bandwidth for a low-rate system. Code-excited LP (CELP) uses codebooks known to both the transmitter and the receiver; the transmitter only sends the codebook index and an excitation magnitude scaling factor, rather than the entire excitation signal, significantly reducing the amount of data transmitted. In CELP systems the excitation vector is chosen from the codebook to minimise the perceptual difference between the original and coded speech. The objective is to have the reconstructed speech resemble the original to the human ear. This is done using a closed-loop analysis-by-synthesis search procedure whereby an estimate of the original

speech is created using the synthesis filter and the excitation codebook vectors in equation (2.14), the estimate is subtracted from the original speech and the resulting error signal is shaped by a filter which reduces the weight of the quantisation error in areas that are less perceptually relevant. By minimising the MSE of this weighted error signal, an excitation is chosen which sounds the most like the original input signal.

2.4.2 G.722.2 / AMR-WB Codec Overview

The AMR-WB codec is a recently developed low bit-rate voice coder designed for packet network telephony. It has been adopted by both the International Telecommunications Union (ITU) for land line use (as ITU standard G.722.2) and the Third Generation Partnership Project (3GPP) for wireless mobile communications [11], [12]. Unlike previous vocoders, such as those used in second generation wireless mobile systems, AMR-WB operates at a sampling rate of 16 kHz, and is therefore able to offer high quality wideband speech. The AMR-WB coder/decoder (codec) defines nine speech coding modes, allowing a trade off between speech quality and transmission rate. The two lowest bit rate modes are intended only to be used during severe channel conditions, while the upper seven modes offer varying degrees of high quality wideband speech. All of the modes use the same coding model, differing only in the size of the CELP codebook, level of parameter quantisation and the number of parameters transmitted. The AMR-WB coder is based on the algebraic code-excited linear prediction (ACELP) [54] coding model which is a form of CELP using algebraic codes to simplify the excitation search procedure.

The AMR-WB codec operates on 20 ms frames of speech, which corresponds to 320 samples at a sampling rate of 16 kHz. The LP analysis and excitation search are performed at a rate of 12.8 kHz, so the input speech is first downsampled by a

factor of 4/5. The speech signal is also filtered through high-pass and pre-emphasis filters. The high pass filter has a cutoff of 50 Hz to eliminate unwanted low frequency components, and the pre-emphasis filter enhances the high frequency components to reduce the dynamic range of the input signal thereby reducing the effects of fixed-point arithmetic.

After input pre-processing, LP analysis is performed to calculate the prediction coefficients. Since there can be a large change in LP coefficients over the 20 ms frame period, the frames are divided into four 5 ms subframes where the computed LP coefficients are used for the fourth subframe, and the coefficients for the first three subframes are calculated by interpolating between the current coefficients and those of the previous frame.

Once LP analysis is complete the excitation is selected from the algebraic codebook. In addition to the fixed codebook described in section 2.4.1, there is also an "adaptive codebook" consisting of the previously transmitted excitation vectors. This past excitation is delayed, scaled and added to the fixed codebook excitation vector to create the long-term pitch structure of the synthesised speech. The adaptive codebook gain and lag are found using the same analysis-by-synthesis procedure used to find the fixed codebook excitation. When the adaptive codebook parameters have been identified, the adaptive contribution is subtracted from the target speech signal to create a new target signal which is used in the closed loop search for the fixed codebook vector. Once the entire excitation is chosen, the fixed and adaptive code book indices, the excitation gain and index, and the quantised LP coefficients, are encoded and transmitted to the receiver

At the receive end, the decoder decodes the received LP coefficients and generates the excitation based on the fixed and adaptive codebook information. The excitation vector is then filtered through the synthesis filter to create a 12.8 kHz sampling rate

speech signal. The signal is post-processed to reverse the pre-processing performed by the encoder and then upsampled to the original 16 kHz. The spectral gap left by the antialiasing filter is filled by a high-band signal which is generated by filtering random noise through a frequency expanded version of the LP filter.

In addition to employing the basic ACELP coding model, the AMR-WB codec has some additional features to permit further transmission rate reductions. As the "Adaptive Multirate" part of the name implies, it is possible for the encoder to adaptively switch between coding modes at a speech frame boundary, in order to maximise the speech quality under changing channel conditions. The codec can also contains a Voice Activity Detection (VAD) component which attempts to determine the presence of a signalling tone, speech or other information which should be transmitted. Depending on the transmission situation, if the VAD determines that no speech is present the coder may either switch to a lower rate to code the background comfort noise, or simply transmit parameters describing the background noise, which is synthesised at the receiver.

2.4.3 Echo Cancellation in the Presence of Vocoder Distortion

For optimum performance an acoustic echo canceller should be placed as close to the echo source as possible, typically it is integrated within the hands-free terminal itself. However, as discussed in [55], there are cases where a centralised echo canceller is desirable. A centralised echo canceller placed at a network gateway could service the echo cancellation needs of multiple terminals, thereby reducing terminal complexity, and enabling decreases in terminal size, power consumption and/or cost. A centralised or network gateway echo canceller may also be required if the echo

cancellation provided by the terminal is not adequate. With the delay dependence of perceptual echo levels, an acceptable level of residual echo may become objectionable when it is subjected to the delays of a VoIP network. When a centralised echo canceller is used in a VoIP situation, the encoder and decoder become part of the echo path. Since LP coders aim for perceptual fidelity rather than waveform matching, the encoding and decoding process introduces significant non-linear distortion into the echo path, degrading the performance of a linear echo canceller.

An extensive investigation of the effects of vocoder distortion was performed by Huang in [6]. The work focuses on network echo cancellation, and compares the ERLE performance of fullband NLMS, Fast Affine Projection (FAP) and Fast Transversal Filter (FTF) algorithms when the echo path includes vocoder distortion in both the transmit and receive paths. The vocoders tested were the ITU G.729 and G.723.1 vocoders, both of which are based on the linear predictive coding model, and operate at 8 kHz sampling rate. It was demonstrated that while FTF provides the best echo performance when no vocoder is present, FAP performs the best in the non-linear echo path. This was attributed to the FAP's superior tracking ability. The source of the non-linear distortion was also investigated: perceptual quality enhancement post-filtering, model parameter quantisation, and LP filter excitation modelling, were all found to affect the echo cancellation performance, however excitation modelling was identified as having the greatest impact.

Following the work of Huang, Lu studied non-linear post-filtering approaches to reduce the significant residual echo which results from vocoder distortion degrading the performance of a linear acoustic echo canceller. The approach in [56], uses a non-linear pitch analysis based post processor to reduce the residual echo left after APA echo cancellation when the channel includes a G.729 vocoder. The work in [55] uses the adaptive cross spectral algorithm rather than APA, and a psychoacoustic

Wiener-type post-filter, which attempts to lower the residual echo below the audible threshold, rather than the pitch analysis approach. As with [56], the G.729 vocoder is used to produce the non-linear echo path.

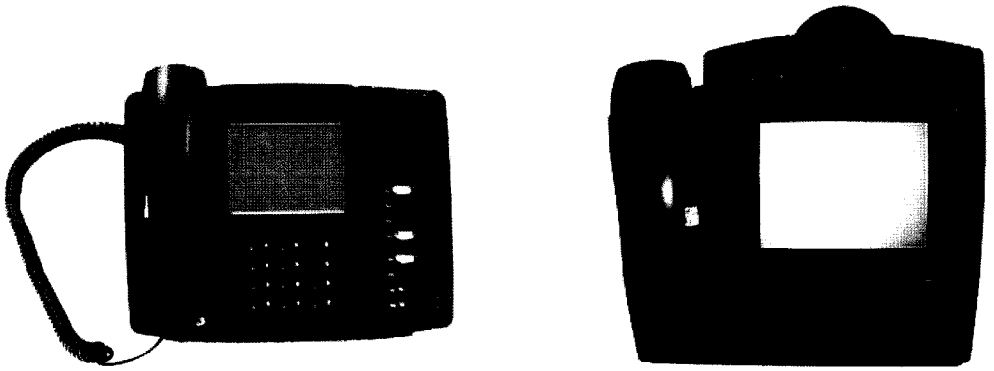
Other than the work of Huang and Lu, there has been little research related to the impact of vocoder distortion on echo cancellation. In particular, the effect of the wideband G.722.2 vocoder has not been investigated. While G.722.2 employs the same coding model as G.729, it operates at a higher sampling rate and uses higher order linear prediction, therefore the distortion it imposes should differ.

Chapter 3

Experimental Setup and Simulation Parameters

3.1 Echo Path Impulse Response Measurement

In order to obtain accurate Loudspeaker-Room-Microphone echo path impulse responses for simulation use, echo path measurements were carried out in real conference rooms using commercially available hands-free terminals: the Mitel 5235 IP Phone and Mitel 5140 IP Appliance, shown in figure 3.1. Signal leads were attached to the terminal allowing the loudspeaker to be driven externally and the microphone signal to be measured directly without any processing from the terminal. The setup used to measure the echo path impulse responses is depicted in figure 3.2. The analog to digital conversion (ADC) and digital to analog conversion (DAC) tasks were handled by a PreSonus Firepod digital sound recording interface. The Firepod has integrated pre-amps with individual gain adjustment which allowed the high level reference and low-level microphone signals to be captured with high dynamic range; a 5 kHz tone was used as a calibration signal to measure the relative gains of the channels. The Firepod was connected via an IEEE 1394 (Firewire) link to a Pentium



(a) Mitel 5235 IP Phone (b) Mitel 5140 IP Appliance

Figure 3.1: Hands-free terminals used for echo path impulse response measurements.

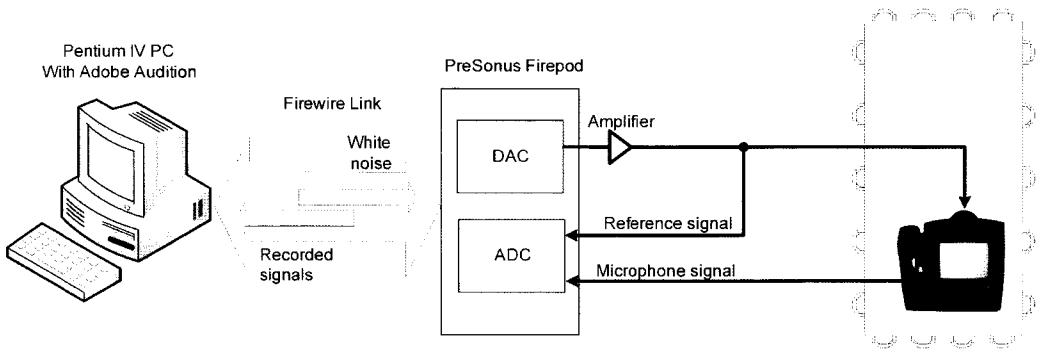


Figure 3.2: Echo path impulse response measurement setup.

IV PC running Windows XP. Thirty seconds of Gaussian white noise excitation was generated on the PC, converted to analog by the Firepod, amplified by an external audio amplifier and used to drive the speaker of the hands-free terminal. The echo signal was captured using the stock microphone in the hands-free terminal, digitised by the Firepod at 48 kHz and 24 bits/sample, and transferred to the PC via the firewire link. The signal used to drive the speaker was also captured, to serve as a delay and distortion compensated reference signal. The playback and simultaneous recording were managed by Adobe Audition version 1.0 recording software running on the PC.

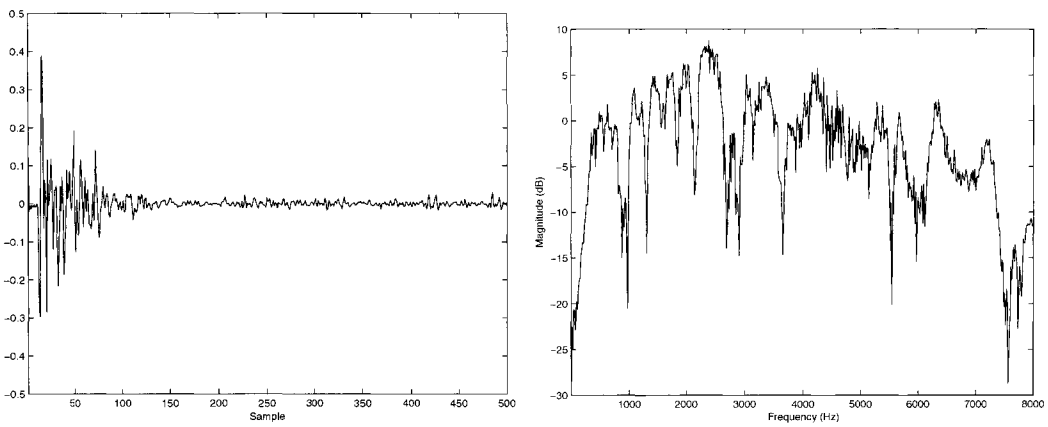
To extract the impulse responses the recorded microphone and speaker signals were

used as the echo and reference inputs to a 12000 tap NLMS adaptive filter running at 48 kHz. A small step size was used to ensure a small steady-state misadjustment. The 48 kHz sampling rate allowed for easy downsampling to 16 kHz wideband and 8 kHz narrowband echo path impulse responses. The 16 kHz impulse response was obtained using the MATLAB command `h = 3*decimate(h,3,256,'FIR')`, which uses a 256 tap FIR anti-aliasing pre-filter before decimating the impulse response by a factor of 3. The scaling factor of 3 compensates for the magnitude scaling introduced by downsampling. It should be noted that decimating the original data sequences prior to NLMS deconvolution produced the same 16 kHz impulse responses.

The measurements were carried out in one large and one small conference room (Carleton University rooms ME4359 and ME4439). The first 500 samples of the normalised impulse responses of the echo paths sampled at 16 kHz, and the corresponding magnitude responses, are presented in figures 3.3 – 3.6. The echo path of figure 3.4 is the path that was used for simulations unless otherwise noted.

For a given terminal, the early samples of impulse responses are similar as they arise from mechanical coupling (vibration) and direct acoustic path coupling. However the later samples, corresponding to the reflections off of the walls, table and ceiling, differ between the rooms. Comparing figures 3.3 and 3.4, which were measured in the large conference room, to figures 3.5 and 3.6, which were measured in the small room, the contributions of the early reflections can be seen more distinctly in the impulse response plots for the small room. This is likely because the walls and ceiling are closer to the phone in the smaller room, and the reflected echoes aren't attenuated before they reach the microphone as much as they are in the large room.

The magnitude responses for a given phone are also similar, especially in the low frequency region. The echo paths of figures 3.5 and 3.3 both possess comb-filter like magnitude responses in the region below 2500 Hz, and the responses of figures 3.6 and



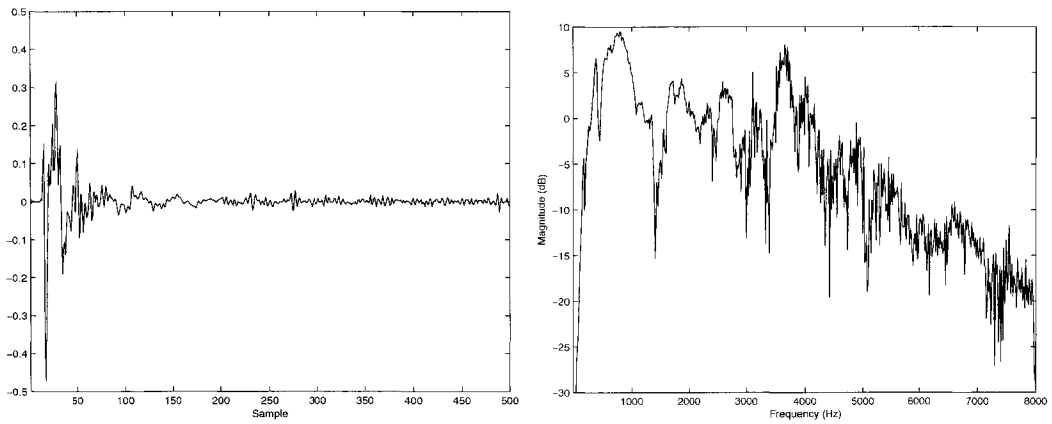
(a) Echo path impulse response (b) Echo path magnitude response

Figure 3.3: Echo path impulse and magnitude response for Mitel 5235 hands-free terminal in a large conference room.

3.4 both have a strong primary peak around 500 Hz and a secondary peak around 2000 Hz. The high frequency regions are more variable for both terminals, but the spectral tilt for all four measured paths is greater in the highband region from 4000 – 8000 Hz than the 0 – 4000 Hz narrowband region. High frequencies are also attenuated more than lower frequencies; in figures 3.5 and 3.6, both measured in the in the small conference room, the magnitude response at 7 kHz is approximately 10-15 dB lower than it is at 4 kHz, depending on the terminal. Additional observations pertaining to the echo path impulse and magnitude responses, and their effect on the simulation and experimental results, are described in the appropriate chapters.

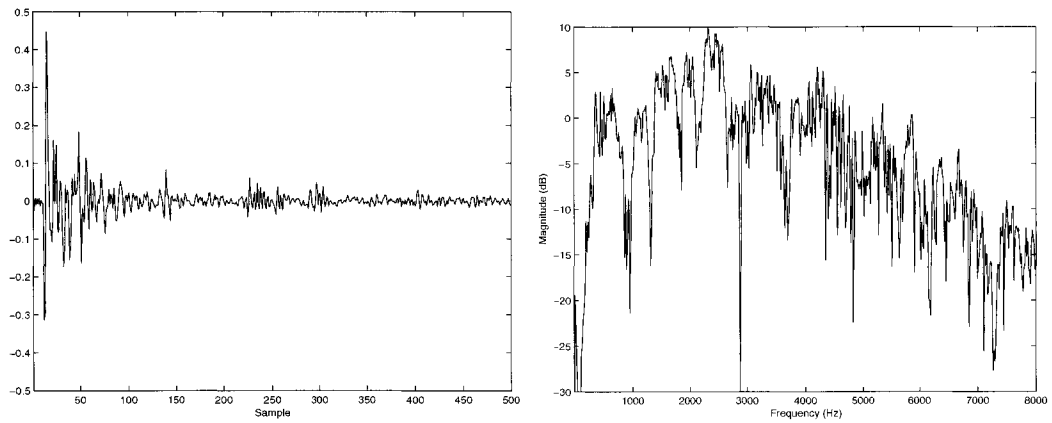
3.2 Performance Metrics

Two typical performance metrics used to compare adaptive echo cancellers are the echo return loss enhancement (ERLE) and the system distance. The ERLE is an estimate of how well the echo signal is being cancelled by the adaptive filter, and is defined as the ratio of the echo power going in to echo canceller to the power of the



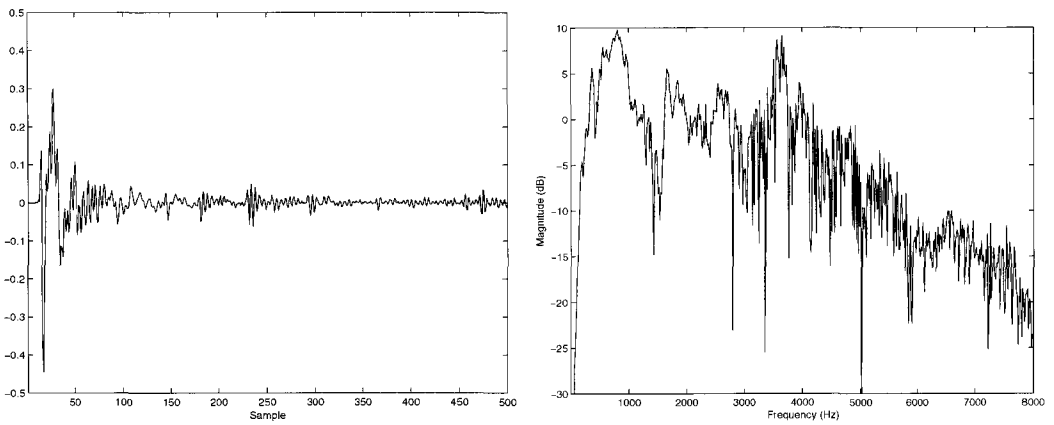
(a) Echo path impulse response (b) Echo path magnitude response

Figure 3.4: Echo path impulse and magnitude response for Mitel 5140 hands-free terminal in a large conference room.



(a) Echo path impulse response (b) Echo path magnitude response

Figure 3.5: Echo path impulse and magnitude response for Mitel 5235 hands-free terminal in a small conference room.



(a) Echo path impulse response (b) Echo path magnitude response

Figure 3.6: Echo path impulse and magnitude response for Mitel 5140 hands-free terminal in a small conference room.

residual echo leaving the echo canceller:

$$ERLE(n) = 10\log_{10}\frac{\mathcal{E}\{d^2(n)\}}{\mathcal{E}\{e^2(n)\}}\ \text{dB}$$

In practice, the expectations are estimated using a sliding window:

$$ERLE(n) \approx 10\log_{10}\frac{\sum_{k=0}^{N_w} y^2(n-k)}{\sum_{k=0}^{N_w} e^2(n-k)}\ \text{dB} \tag{3.1}$$

Where $y(n)$ is the microphone signal, $e(n)$ is the residual echo, and N_w is the size of the averaging window. A window size of $N_w = 500$ was chosen for this thesis as it is long enough to smooth the ERLE estimate, but short enough that the power of a 16 kHz speech signal does not fluctuate greatly during the window period.

System distance provides an estimate of how deeply the system has converged by measuring the norm of the difference between the adaptive filter tap weight vector and the time-invariant Wiener solution. For a fullband filter, the system distance is given by:

$$\Delta(n) = 10\log_{10}\left(\frac{\|\underline{h} - \underline{\hat{h}}(n)\|^2}{\|\underline{h}\|^2}\right)\ \text{dB} \tag{3.2}$$

Where $\underline{h}$ is the Wiener solution and $\hat{\underline{h}}(n)$ is the adaptive filter tap weight vector at time n. As discussed in section 2.2 a subband adaptive filter does not converge exactly to the Wiener solution, rather a set of fully-converged local error adapted subband filters satisfies the condition:

$$\sum_{m=0}^{M-1} \hat{\underline{h}}_m(z^D)\underline{H}_m(z) = \sum_{m=0}^{M-1} \underline{H}_m(z)\underline{h}(z).$$

The subband system distance is therefore calculated as:

$$\tilde{\underline{h}} = \sum_{m=0}^{M-1} \underline{h} * \underline{H}_m$$

$$\tilde{\hat{\underline{h}}}(n) = \sum_{m=0}^{M-1} \underline{H}_m * \hat{\underline{h}}_m(n/D)$$

$$\Delta_{SB}(n) = 10\log_{10}\left(\frac{\|\tilde{\underline{h}} - \tilde{\hat{\underline{h}}}(n)\|^2}{\|\tilde{\underline{h}}\|^2}\right) \text{ dB}$$

Since the goal of the echo canceller is to remove the echo rather than match the Wiener solution, ERLE will be the more frequently used performance metric. For white inputs ERLE and system distance generally oppose, a large ERLE indicates a small system distance. This is not the case for narrowband inputs where the adaptive filter will rapidly converge in the frequency region where there is power, resulting in a large ERLE, but the tap weight vector will still be far from the Wiener solution, resulting in a large system distance. If the filter has converged for a narrowband input, and is then driven by frequencies beyond the original input, the ERLE performance will be poor while the system distance will improve, as the filter converges in the new frequency region. System distance will be used to compare convergence speed of different adaptive filtering structures and algorithms, and to show how deeply the adaptive filter converges.

3.3 Simulation Parameters

All of the simulations were carried out in MATLAB version 6.5.1 for Windows XP, running on a Pentium IV 2.6 GHz PC with 1 GB of RAM. The speech signal inputs are from the TIMIT database [57], and the white Gaussian inputs were generated using MATLAB's `randn` command. No noise was added to the measurement signals, instead the ERLE was limited by undermodelling. The number of fullband taps was selected to be 2400, which, based on figure 2.5 yields a maximum ERLE of approximately 30 dB. The NLMS step-size was chosen to be $\mu = 1.0$, which gives the fastest convergence for NLMS [20]. The affine projection order was selected to be $P = 3$, as the performance gains were minimal for higher orders, especially in the vocoder distorted channel (this is consistent with the findings of [56]). The number of subbands and decimation factor were chosen to be $M = 8$, and $D = 4$, as the combination offered reduced complexity compared to the fullband system, while maintaining low delay due to the short analysis and synthesis filters enabled by two-times oversampling. Structures with $M = 4$ and $M = 16$ were also tested, and offered comparable performance. The G.722.2 encoder and decoder programs used in chapter 7 are the 3GPP floating point reference implementation of the AMR-WB codec [58]. Table 3.1 summarises the main adaptive filtering algorithm parameters that were used for the simulations, unless otherwise noted.

Parameter	**Notation**	**Value**
Number of subbands	M	8
Decimation factor	D	4
Number of fullband adaptive filter taps	N	2400 taps
Number of subband adaptive filter taps	N/D	600 taps/subband
NLMS step-size	μ	1.0
Affine projection order	P	3
IP-NLMS proportioning parameter	α	-0.5

Table 3.1: Default simulation parameters.

Chapter 4

Output Error Levels in Oversampled Subband Adaptive Filters

Under certain conditions, the mean square error performance of adaptive filters can exceed that of the linear time-invariant Wiener filter. The causes of the phenomenon are presented and the conditions under which it occurs are discussed, and shown to hold true for lowpass, non-Markov correlated signals. In this chapter observations of the effect in subband adaptive filters are presented and the practical implications are discussed.

4.1 Output Error Performance Bounds for Adaptive Filters

Contrary to the simplified description presented in section 2.1, the NLMS algorithm does not adapt the tap weight vector to the Wiener solution, but rather it adapts to minimise the output squared error. If the statistics of the input and reference signals obey certain conditions, known as the "independence assumptions", the Wiener filter is the solution that minimises the output squared error, and the two descriptions

can be considered equivalent. In certain cases where those conditions are not met, NLMS adaptive filters have been shown to out-perform the time-invariant Wiener filter. While the Wiener filter is the MSE-optimal linear time invariant (LTI) filter, the NLMS estimator is non-linear and may exploit signal correlation information not used by the Wiener filter to achieve a lower MSE. This effect was demonstrated in [59] for the case of adaptive equalisation, in [60] for adaptive linear prediction and in [61] for adaptive noise cancellation. In [62] near-end distortion was observed in a subband APA echo canceller, and was attributed to the colouration of the highly oversampled subband signals. An in-depth review of the topic of non-Wiener effects in adaptive filtering can be found in [63], and summary of the explanation presented in [63] and [61] will be presented here.

Incorporating the adaptation history, the tap weight vector of an NLMS adaptive filter (the echo path estimate) at time n can be written as [63]:

$$\underline{\hat{h}}(n) = \underline{\hat{h}}(0) + \mu \sum_{i=0}^{n-1} \frac{e^*(i)}{\underline{x}^H(i)\underline{x}(i)} \underline{x}(i). \tag{4.1}$$

This yields the estimated echo replica:

$$\begin{aligned} \hat{d}(n) &= \underline{\hat{h}}^H(n)\underline{x}(n) \\ &= \underline{\hat{h}}^H(0)\underline{x}(n) + \mu \sum_{i=0}^{n-1} \frac{e(i)}{\underline{x}^H(i)\underline{x}(i)} \underline{x}^H(i)\underline{x}(n). \end{aligned} \tag{4.2}$$

Since $e(n) = d(n) - \hat{d}(n)$, the estimate in (4.2) is a function of all past input and reference signal samples and the mean-square optimal estimator, which upper-bounds the NLMS estimator, is given by:

$$\hat{d}_{opt}(n) = \mathcal{E}\{d(n)|x(n), x(n-1), \ldots, x(0), d(n-1), d(n-1), \ldots, d(0)\}. \tag{4.3}$$

In order to investigate the convergence of (N)LMS, most authors impose restrictions on the statistics of the input and desired signals to simplify the estimator in (4.3).

There are a variety slightly differing "independence assumptions", but those used in [61] are:

1. the current data vector, $[d(n),\ \underline{x}^T(n)]^T$, consisting of the concatenation of the current desired signal and current input vector, and all previous data vectors $\{[d(n-1), \underline{x}^T(n-1)]^T, \ldots, [d(0), \underline{x}^T(0)]^T\}$ are independent of one another.

2. the current desired signal $d(n)$ is dependent only on the current input vector $\underline{x}(n)$.

3. the desired signal $d(n)$ and the input vector $\underline{x}(n)$ are mutually Gaussian.

The first two conditions reduce the terms in the expectation of (4.3) so that it simplifies to

$$\hat{d}_{ind}(n) = \mathcal{E}\{d(n)|\underline{x}(n)\},$$

and the third condition restricts the optimal estimator to be linear, so that the optimal estimate can be written as

$$\hat{d}_{ind,lin}(n) = \underline{\hat{h}}^H \underline{x}(n).$$

Under these conditions the MSE-optimal estimator, and the upper bound for the NLMS estimator, is the Wiener solution.

The independence assumptions are invoked to simplify analysis of the NLMS estimator in order to predict the average convergence of the adaptive filter, but in some cases they do not apply. Assumption 1 is especially restrictive, as it precludes correlation in the input data sequence. In [61], rather than investigate the *average* performance the authors wished instead to *upper-bound* the performance of the NLMS estimator. To achieve this goal the first two assumptions were dropped, allowing for correlated inputs and resulting in an optimal estimator that is a function of all past data vectors, while the third independence assumption was maintained to produce a

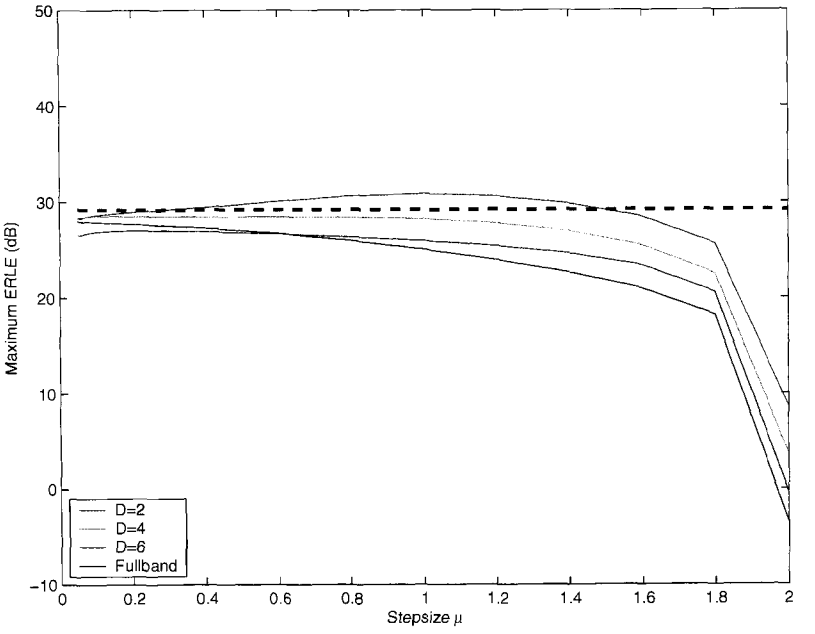
linear estimator. Under these more relaxed conditions, the optimal linear estimator is a time-invariant two-channel infinite horizon Wiener filter which uses all previous input and desired signal samples to estimate the current desired signal sample. In [63] it is shown that since the correlation sequence of most signals decays rapidly, the optimal estimator can be well approximated by a two-channel finite horizon Wiener filter, which incorporates a limited history of reference and desired signal samples. It is also shown that the two-channel Wiener filter can be represented by an equivalently optimal one-channel, time-varying Wiener filter which uses only the reference signal $x(n)$ as input.

Using this time-varying Wiener filter model, the superior MSE performance of adaptive filters was explained. When the input processes to the NLMS filter are sufficiently narrowband, the short term correlation will be high. If the adaptation step-size is large enough, the NLMS adaptive filter can approximate the time-varying Wiener filter and exploit the correlation to achieve MSE performance above that of the one-channel LTI Wiener filter. If the input signals are not coloured, no information about the current desired signal can be gleaned from previous desired signal samples or the distant past of the input signal, and the traditional Wiener filter is optimal. If the step-size is not large enough, the NLMS filter cannot track the time-varying Wiener filter, so it converges towards the optimal time-invariant Wiener solution. Clearly the NLMS estimator does not and can not make use of all of the information in (4.3) as most of it is embedded in the tap weight vector, consequently its MSE performance can only approach that of the two-channel Wiener filter when the input signals exhibit high short-term auto and cross-correlations. On the other hand, the more general affine projection algorithm makes explicit use of the previous P input vectors and desired signal samples, so it may be better able to exploit the coloration of both the input and the reference signals.

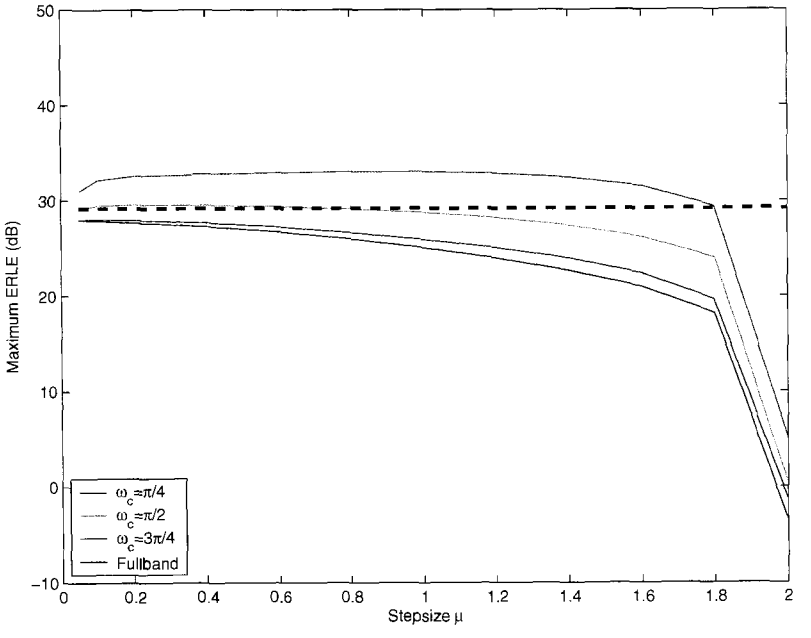
4.2 Observations in Oversampled Subband Acoustic Echo Cancellers

In oversampled subband adaptive filters, the analysis filterbank combined with non-critical downsampling results in subband reference and desired signals that are severely coloured. In highly oversampled subband adaptive filters, this subband signal colouration can lead to non-Wiener effects and near-end signal cancellation. A simulation was performed to compare the steady state ERLE levels of fullband NLMS and APA adaptive filters to that of 8-channel subband NLMS and APA systems with decimation factors of 2, 4 and 6. The objective was to determine whether or not an oversampled subband adaptive filtering system with a large step size and high oversampling factor can be made to exhibit non-Wiener behaviour when the equivalent fullband system does not. The input to the fullband and subband adaptive filters was 5×10^5 samples of white Gaussian noise signal, and the echo signal was formed by convolving the input with a measured echo path impulse response. The filters were allowed to adapt to the sequence, and steady-state ERLE was calculated from the last 1000 samples and plotted as a function of step-size. To verify that the effect was not caused by some other phenomenon in the subband configuration, the experiment was repeated using fullband filters with varying degrees of input signal coloration. Fullband coloured inputs mimicing those of the oversampled subband signals were created from the WGN data sequence by lowpass filtering with decimated (frequency expanded) versions of the subband analysis prototype filters. The results are presented in figure 4.1 for the NLMS case and figure 4.2 for APA with order $P = 3$; the ERLE that would be achieved by the Wiener filter is indicated by the dashed line.

There are several notable points from the ERLE results. Clearly the more highly oversampled the subband structure, or equivalently the more narrowband the input to

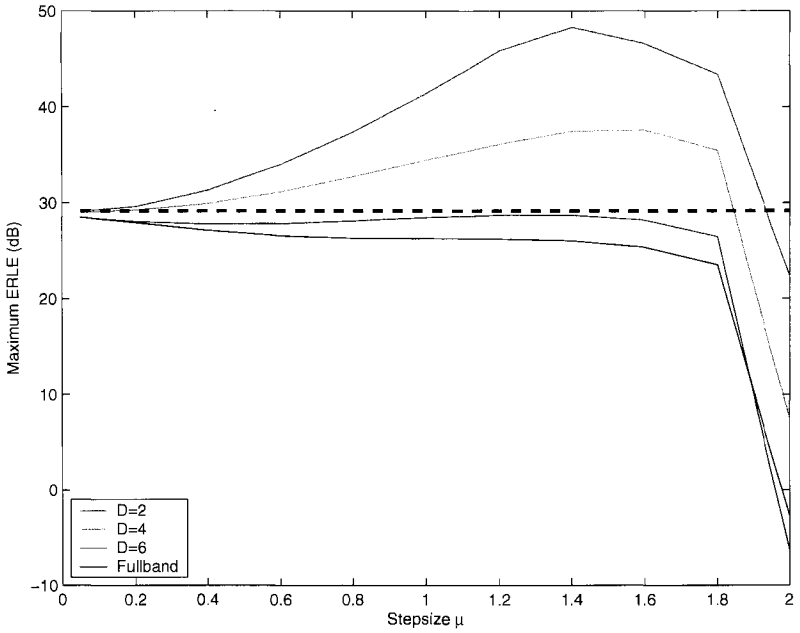


(a) Subband NLMS with varying degrees of oversampling.

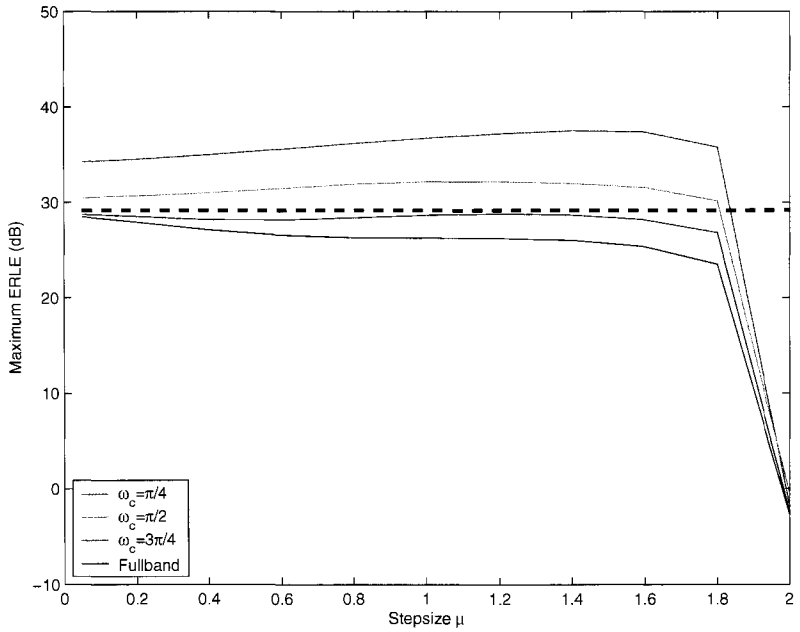


(b) Fullband NLMS with varying degrees of coloration.

Figure 4.1: Maximum ERLE as a function of step-size for oversampled subband and fullband adaptive filters using the normalised LMS algorithm. The heavy dashed line indicates the Wiener filter ERLE.



(a) Subband APA with varying degrees of oversampling.



(b) Fullband APA with varying degrees of input signal coloration.

Figure 4.2: Maximum ERLE as a function of step-size for oversampled subband and fullband adaptive filters using the affine projection algorithm. The heavy dashed line indicates the Wiener filter ERLE.

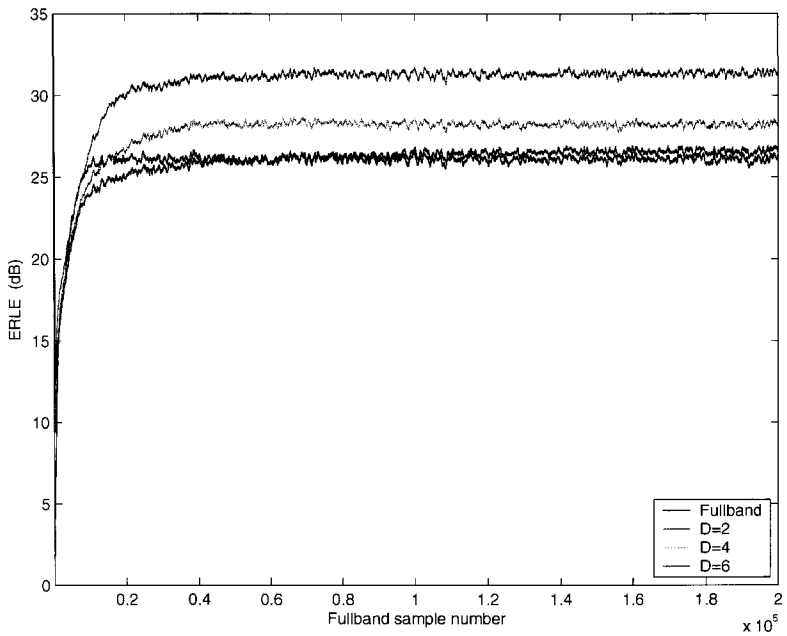
the fullband structure, the higher the steady state ERLE for large step-sizes. This is in accordance with the hypothesis that the narrowband/highly coloured nature of the input signals can be exploited to achieve a higher ERLE. Also of note is the fact that, of the subband NLMS configurations, only the four-times oversampled (decimation factor of 2) system achieves an ERLE that is higher than the Wiener level. While the others seem to be exploiting the subband signal correlation to achieve an ERLE that is greater than the fullband structure, the signals are not correlated enough to reach a level that is in excess of the linear time-invariant Wiener filter. For subband APA both the four-times, and 2-times oversampled configurations achieved ERLEs in excess of the Wiener level, and the APA structures achieved a higher maximum ERLE than NLMS, indicating that it is better at exploiting the colouration. The discrepency between the fullband and subband APA results may be a result of the better tracking ability of subband filters (see chapter 5). Finally, the ERLE improvement begins to decrease once the step-size reaches a critical level, approximately 1 for NLMS and 1.5 for APA. This is likely because the excess MSE caused by gradient noise overrides the benefit gained by tracking the time-varying Wiener filter. What is not clear from the figures is what impact the ERLE gain has on the adaptive filter tap weight vector.

When the NLMS filter is tracking the time-varying Wiener filter, the adaptive filter tap weights fluctuate around some neighbourhood of the LTI Wiener solution, rather than converging to it. In order to show that this results in a tradeoff between ERLE performance and system distance, the performance of the fullband and oversampled subband systems for these metrics was compared for the case of large and small step-sizes. As with the previous experiment the trials were run for subband structures with varying degrees of oversampling and for fullband structures with varying degrees of input signal coloration. Figure 4.3 compares the ERLE, averaged over 20 independent trials, and figure 4.4 compares the system distance performance

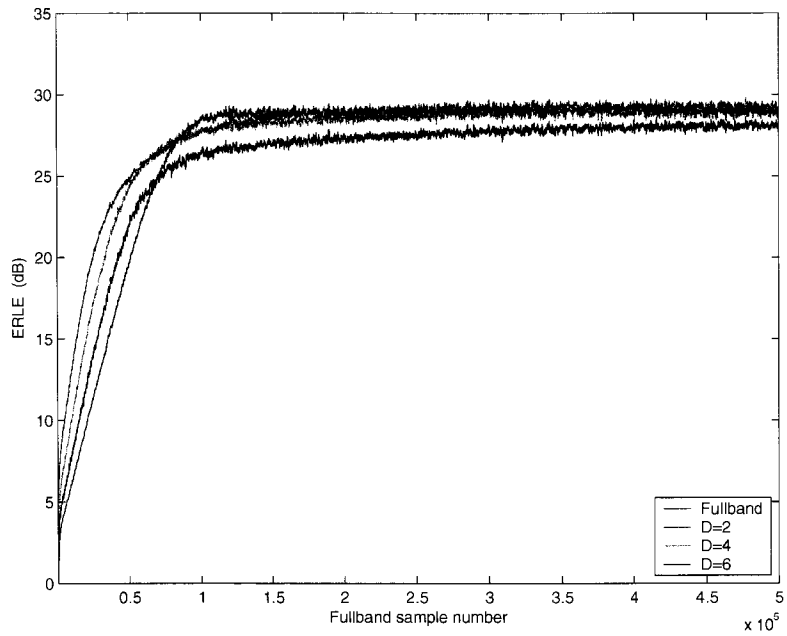
of the subband systems with step-sizes of $\mu = 1.0$ and $\mu = 0.1$. Figures 4.5 and 4.6 present the same information for the fullband systems with coloured inputs. Note that the $\mu = 1.0$ trials were run for 2×10^5 samples, while the $\mu = 0.1$ trials were run for 5×10^5 samples, to account for the slow convergence in the latter case.

As expected the ERLE gain of the oversampled structures, which can be seen in figure 4.3(a) for $\mu = 1.0$ is no longer present in figure 4.3(b) when $\mu = 0.1$. Similarly, comparing figure 4.4(a) to 4.4(b), it can be seen that decreasing the step-size results in slower, but deeper convergence to the LTI Wiener solution. Both of these observations are consistent and indicate that for small step-sizes the NLMS tap weight vector will converge to the LTI Wiener solution, whereas for large step-sizes it will track the time-varying optimal Wiener solution [63]. The slow asymptotic convergence of the subband structure system distance is well known [64], and results from slowly converging modes in the transition band region, where there is very little adaptation energy.

Comparing figures 4.5(a) and 4.5(b), it can be seen that, as with the oversampled subband systems, reducing the step-size for the fullband systems with coloured inputs removes most of the ERLE advantage except for the most coloured input. Figures 4.6(a) and 4.6(b) show that unlike the subband systems, lowering the step-size did not significantly deepen the convergence. Contrary to the subband situation where the signal coloration is introduced by oversampling and the input signal is white, the coloration in this case is present in the original input signal. As a result, regardless of the step-size, the filter can not converge in the frequency region corresponding to the stopband of the lowpass filter, as there is no energy to drive the adaptation, so the final system distance remains the same. It is also interesting to note that the initial system distance convergence rate is not lower for the coloured inputs than it is for the white noise. This seems to be a contradiction to the common guideline that LMS

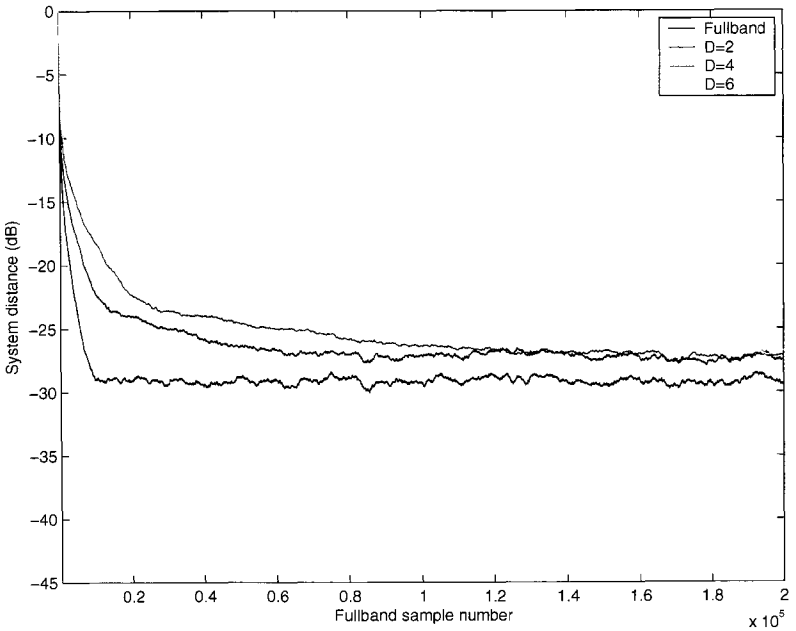


(a) $\mu = 1.0$

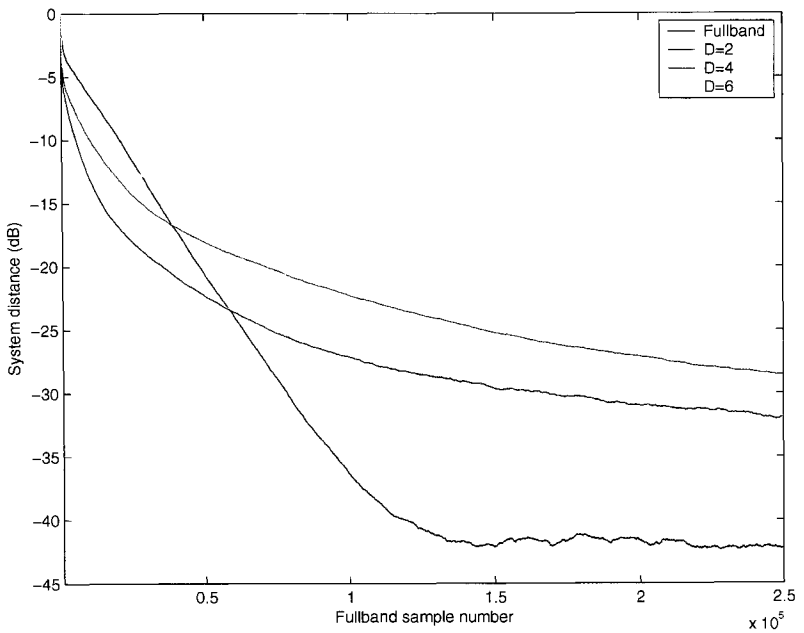


(b) $\mu = 0.1$

Figure 4.3: Comparison of ERLE convergence using two different step-sizes for subband adaptive filters with varying degrees of oversampling, average over 20 trials.

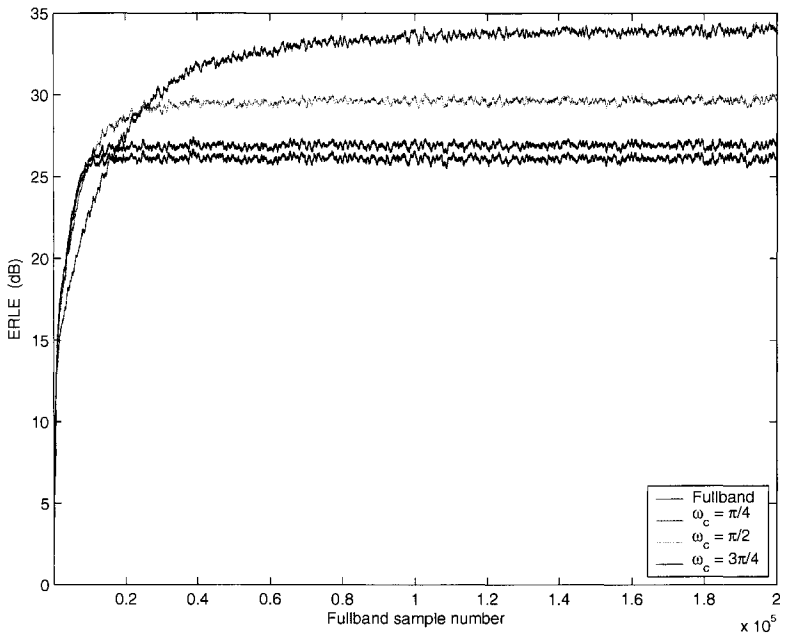


(a) $\mu = 1.0$

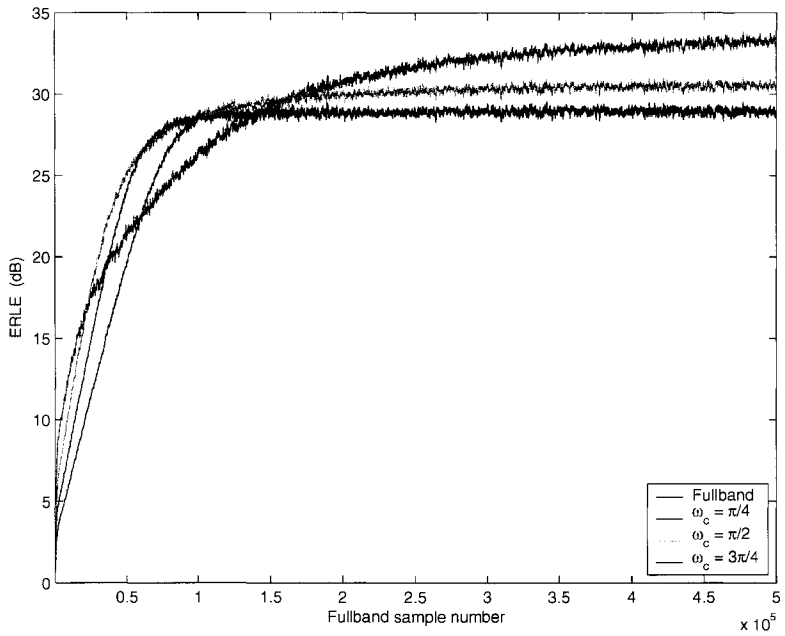


(b) $\mu = 0.1$

Figure 4.4: Comparison of system distance convergence using two different step-sizes for subband adaptive filters with varying degrees of oversampling, average over 20 trials.

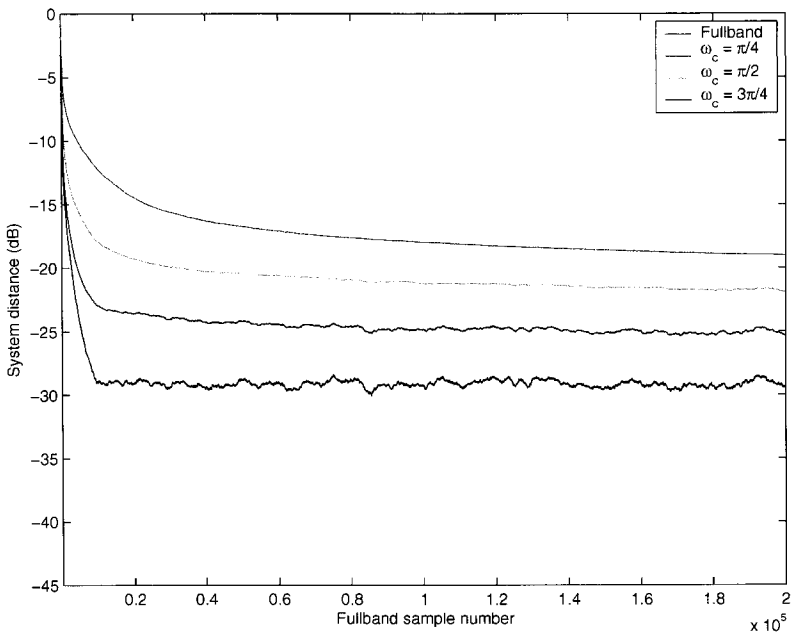


(a) $\mu = 1.0$

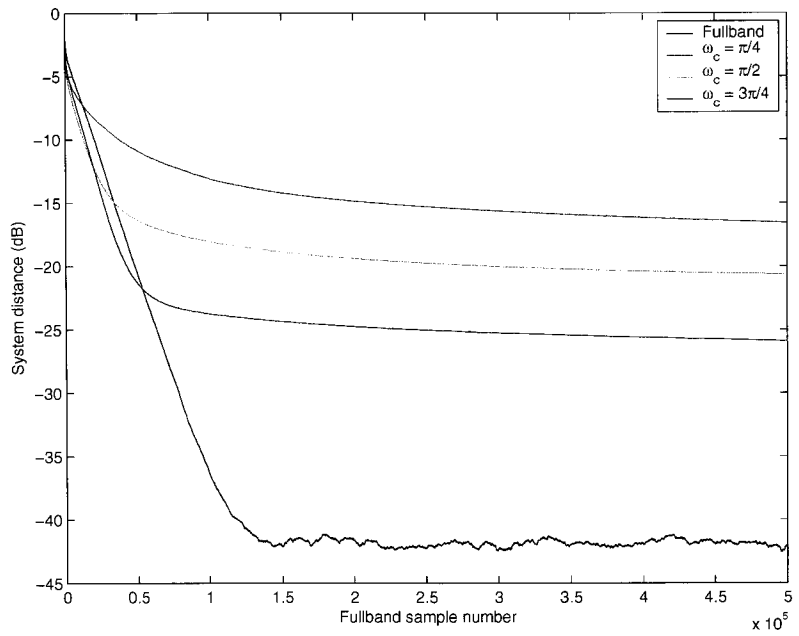


(b) $\mu = 0.1$

Figure 4.5: Comparison of ERLE convergence for two different step-sizes using fullband adaptive filters with coloured inputs, average over 20 trials.



(a) $\mu = 1.0$



(b) $\mu = 0.1$

Figure 4.6: Comparison of system distance convergence for two different step-sizes using fullband adaptive filters with coloured inputs, average over 20 trials.

convergence speed is governed by the ratio of the largest to the smallest eigenvalue of the autocorrelation matrix, and is therefore slower for correlated inputs [4]. While it is true that the asymptotic convergence is fastest for the white noise case, the initial convergence shown in the figures is consistent with the theory presented by Slock in [20] and Morgan in [21]. In [20] it is demonstrated that for any given point in time, there exists an autocorrelation matrix eigenvalue distribution for which the MSE is lower than the white noise MSE at that same time. For input signals with high eigenvalue spreads (ie., highly coloured signals), the adaptive filter will experience a period of rapid initial convergence followed by a very slow asymptotic convergence, which is what is observed in figures 4.6 (a) and (b). Unlike Markov (autoregressive) process correlated signals, the lowpass filtered input signals are spectrally flat in the active region so the adaptation energy is equally distributed amongst all frequencies that are present. As a result the initial convergence rate in the active region is faster than the white noise case for the same signal power. After the filter has converged in the active region, there is very little energy left to drive adaptation, so the filters do not converge in the stopband region and the steady state system distance is large.

4.3 Summary and Practical Applications

In this chapter, it was demonstrated that when an adaptive filter with a relatively large step-size receives correlated inputs, the filter can exploit the correlation to achieve output error levels below those of the linear time-invariant Wiener filter. Since the APA explicitly uses past data, it is better than the NLMS algorithm at exploiting the correlation in both the desired and reference signals to reduce the output error; most of the signal correlation information is embedded in the NLMS tap weight vector. It was also shown that the colouration inherent in oversampled

subband signals is sufficient to cause oversampled subband adaptive filters to produce output error levels below the time invariant Wiener filters, even if the original fullband signal is uncorrelated.

While it is evident that adaptive filters can be made to exhibit this non-Wiener behaviour, it is not entirely clear whether or not the effect is desirable. The effects of the phenomenon may depend on the target application, and whether it is being exhibited in a subband or fullband system. In an echo cancellation application the objective is to minimise the residual echo and achieving below Wiener output error may be desirable. However, if the target application is system identification the low output error, which comes at the cost of higher system distance, reduces system modelling accuracy. In a fullband configuration, the ERLE gain in the frequency region where the filter is excited is obtained by varying the tap weights about the Wiener solution. These fluctuations sacrifice the modelling in the region where there is less energy in order to improve the performance in the active region. If a fullband filter exhibiting non-Wiener effects filter were to receive echo energy in the previously non-active frequency region, the cancellation would be very poor until the filter converged to the new signal. In contrast, the low energy regions in the subband correspond to the analysis filterbank stopband, so there is no risk that these modes will be excited. The authors of [46] claim that oversampled subband systems may treat the stopband regions of the subband signals as "don't care" regions, offering extra degrees of freedom to achieve a lower MSE. While their discussion was in the context of modelling non-causal taps, the principle may apply in this case as well.

In wideband speech, the voiced speech segments consist of strong narrowband frequency components, so it is possible that non-Wiener effects could occur in a fullband structure. In both the subband and fullband configurations the phenomenon relies upon the time-varying nature of the adaptive filter, therefore the tap weights must

keep changing for the benefit to be realized. If adaptation halts, as in a doubletalk situation, the weights will be frozen in a sub-optimal state in a neighbourhood of the Wiener filter weights, resulting in a somewhat higher residual echo during doubletalk periods. However, according to [37] the desired echo return loss during doubletalk is about 5 to 10 dB lower than the singletalk case, on account of temporal and frequency masking of the echo signal by near end speech, so the temporarily higher residual echo during doubletalk may not be objectionable. On the other hand, in [62] a subband APA filter is allowed to adapt during doubletalk, while the adaptation remained stable, some of the near-end signal was cancelled resulting in undesirable audible distortion. As this demonstrates, despite its potential benefits there are cases where the phenomenon should be avoided, so a system designer should be aware of it, especially when employing oversampled subband adaptive filters.

Chapter 5

Effect of Changing Echo Path on Echo Cancellation

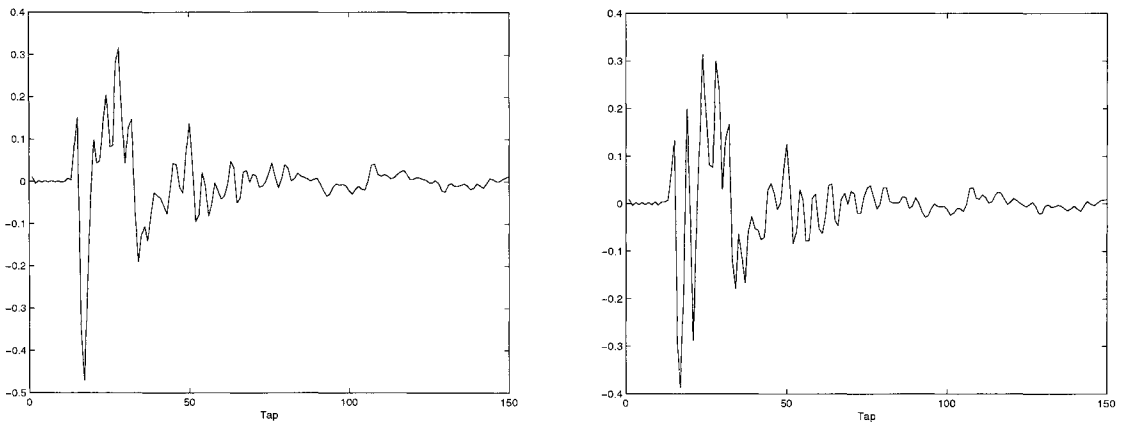
Many comparisons of echo cancellation algorithms focus on the speed of initial convergence, while less attention is given to tracking ability. Acoustic echo paths change rapidly and frequently in response to changes in the near-end room. A person moving, an obstructing object being placed near the speaker or a door opening all affect surfaces off of which echoes reflect, thereby changing the echo path. Acoustic echo cancellers must respond quickly to track these changes in order to maintain good echo cancellation performance.

In this chapter, it is demonstrated that echo path changes do not impact all regions of the echo path impulse and magnitude responses equally. With that observation in mind, fullband and subband versions of NLMS, IP-NLMS and IP-APA are compared to determine which structure and algorithm performs the best in simulated and real changing echo environments. NLMS is used as a benchmark to evaluate the fast-tracking IP-NLMS, and IP-APA is also considered to determine the impact of the memory of the algorithm on its tracking ability. The 16 kHz wideband and 8 kHz narrowband cases are both considered to investigate the impact of signal bandwidth

on the tracking ability of adaptive filters.

5.1 Changing Echo Path Impulse Responses

Depending on the nature of the echo path change, some of the components of the impulse response may remain the same despite the disruption. For a given hands-free terminal, the distance between the speaker and the microphone is fixed. Consequently the delay associated with the direct path and the early reflections off of the table will remain relatively constant, and the location of the samples in the impulse response corresponding to those reflections will remain fixed, though the magnitudes may vary. This effect can be observed in figure 5.1, which shows the first 150 samples of two echo path impulse responses measured in the same room using the same hands-free terminal. The measurement of figure 5.1(a) was made with the hands-free terminal on an empty table, and the impulse response of figure 5.1(b) was measured with a small box placed on the table in front of the hands-free terminal. Since the box does not greatly obstruct the direct path between the speaker and the microphone, the location of the first large echo path component is at sample 17 for both measurements, although the magnitude differs slightly. Since many of the reflecting surfaces (table, ceiling, walls) are consistent between the two measurements, the general shape of the impulse response also remains the same. One exception is the presence of additional early reflections, likely corresponding to reflections off of the box, which can be seen around sample 20. Figure 5.1 also demonstrates the justification for individual step-size algorithms. The impulse response samples with the greatest change in magnitude are the samples with the greatest magnitude. The proportional algorithms are able to exploit the similar overall shape of the changing impulse responses to achieve better tracking than ordinary NLMS; the larger early reflection components have greater



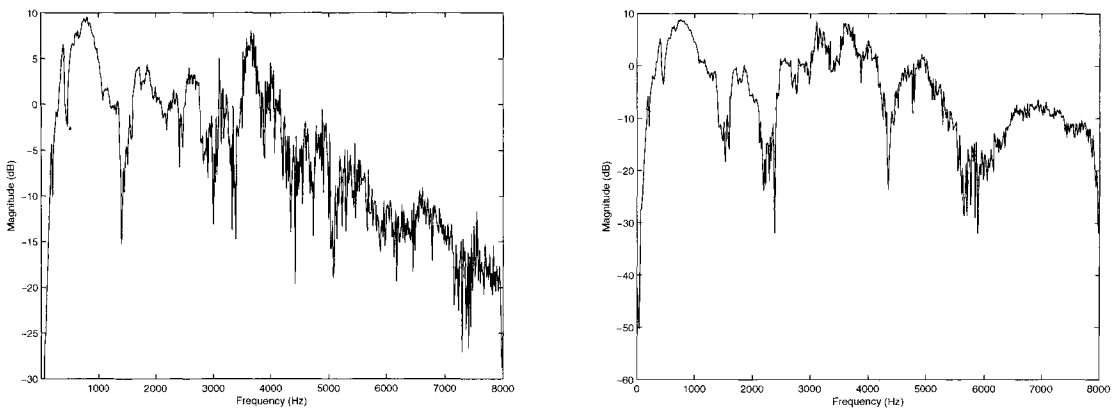
(a) Unobstructed echo path

(b) Obstructed echo path

Figure 5.1: First 150 samples of the impulse responses used for changing echo path simulations.

magnitude fluctuations, so they are allotted more of the adaptation power, producing faster initial convergence and tracking.

Just as echo path changes leave some samples of the impulse response unchanged, the effect on the echo path magnitude response can be limited to certain frequency regions. A moving hand in the echo path or an object being placed in front of the speaker may affect higher frequencies more than lower, because high frequencies are absorbed more easily and because the wavelength of the lower frequencies may be larger than the hand or obstructing object. Figure 5.2 presents the magnitude responses corresponding to the echo path impulse responses of figure 5.1. In contrast to the impulse responses, the overall shape of the magnitude responses differs considerably. The exception is the region below approximately 1.5 kHz, where the magnitude responses are quite similar, indicating that the low frequency echo components are not affected as greatly by the obstruction. This is consistent with the observation from section 3.1 that for a given hands-free terminal, the low frequency region of the magnitude response does not vary as much as the high frequency region. For this type of echo path change a subband adaptive filter might offer better tracking. While



(a) Unobstructed echo path (b) Obstructed echo path

Figure 5.2: Magnitude response of echo paths used for changing path simulations.

a fullband filter would have to adapt all tap weights in response to the disruption, a subband adaptive filter would only be required to adjust the weights of the filters in the affected bands. This is supported by figure 5.3, which presents the real part of the first 100 samples of the subband analysed impulse responses of figures 5.1. The impulse responses of lowest subband, covering the frequency range from 0 – 2 kHz, are almost identical. There are only slight changes in the sample values, and there are no additional reflection components, so the subband adaptive filter would quickly reconverge in that band. For the upper 3 bands, the differences between the unobstructed and the obstructed echo path impulse responses are similar to the fullband case. However, the subband adaptive filters may have a tracking advantage even in the upper bands; each of the subband adaptive filters is shorter than the fullband filter, so they should be faster to respond to and reconverge for all types of echo path change.

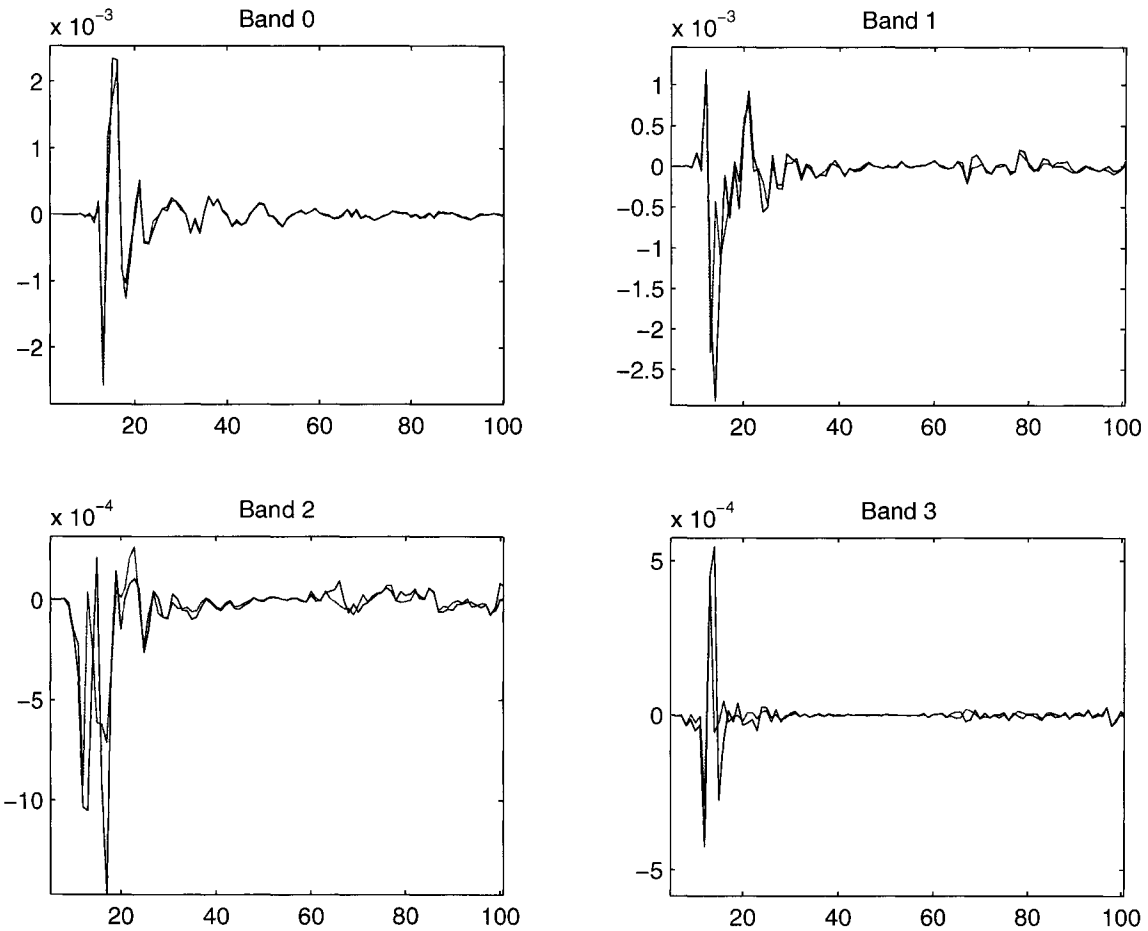


Figure 5.3: First 100 samples of subband unobstructed (blue) and obstructed (red) echo path impulse responses.

5.2 Simulated Echo Environment Results

To examine the impact of a changing echo path on the echo cancellation performance of adaptive filters under controlled conditions, fullband and subband versions of the NLMS, IP-NLMS and IP-APA algorithms were compared using 5 seconds of white gaussian noise excitation with a simulated changing acoustic environment. After 1.25 seconds of initial convergence an abrupt echo path change was simulated by changing the echo path coefficients from those of the small conference room of figure 3.6 to the coefficients of the large conference room of figure 3.4. The filters were allowed to reconverge for 2 seconds, then from $t = 3.0$ to $t = 3.5$ seconds a more realistic gradual change was simulated by linearly interpolating between the two impulse responses of figure 5.1, as in [65]. The ERLE performance, averaged over 20 trials, is shown in figure 5.4 for 8 kHz sampling rate and figure 5.5 for 16 kHz sampling rate. From the figures it is clear that the subband IP-APA consistently offers the highest ERLE performance, followed by the subband IP-NLMS algorithm. The constant gap, of approximately 5 dB, between subband IP-APA and all of the other algorithms, combined with the fact that subband IP-APA has this advantage over subband IP-NLMS while fullband IP-APA performed identically to fullband IP-NLMS, as would be expected for a white input, indicates the presence of the non-Wiener effects discussed in chapter 4.

According to [65], the minimum ERLE during the echo path change plays an important role in a user's subjective evaluation of echo cancellation. From the figures it is clear that the subband structures outperform their fullband counterparts in terms of minimum ERLE for both abrupt and gradual echo path changes. The average and standard deviation, taken over 20 trials, of the minimum ERLE for all of the algorithms is summarised in table 5.1 for 8 kHz sampling rate, and table 5.2 for 16

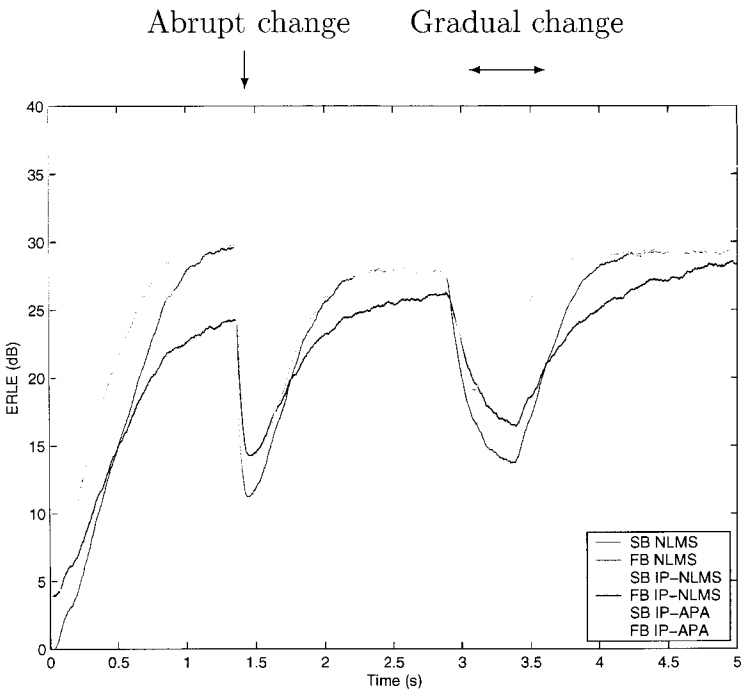


Figure 5.4: Reconvergence and tracking ERLE performance of fullband and subband algorithms, 8 kHz sampling rate.

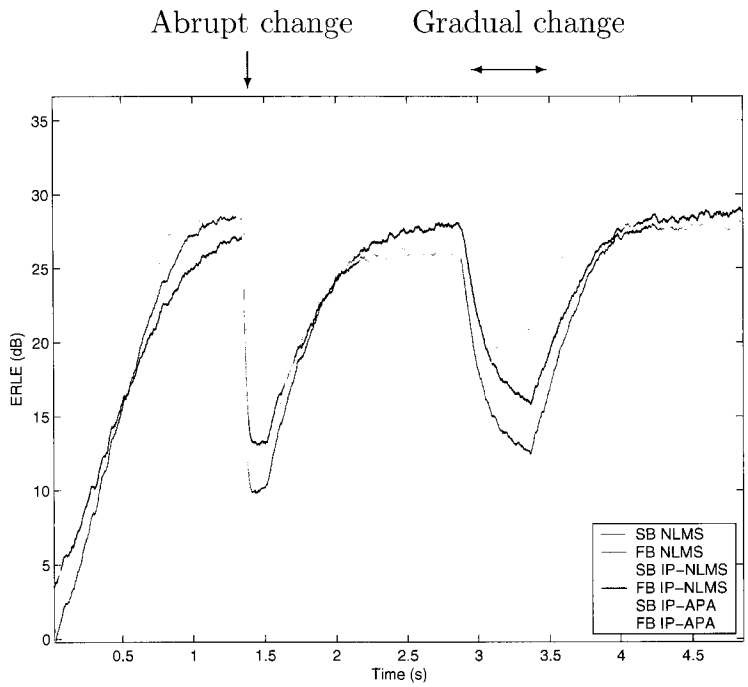


Figure 5.5: Reconvergence and tracking ERLE performance of fullband and subband algorithms, 16 kHz sampling rate.

	Minimum ERLE (dB)	
Algorithm	**Abrupt Change**	**Gradual Change**
Fullband NLMS	10.81 ± 0.61	13.06 ± 0.58
Subband NLMS	13.71 ± 0.50	15.84 ± 0.61
Fullband IP-NLMS	11.69 ± 0.63	18.78 ± 0.40
Subband IP-NLMS	14.76 ± 0.57	24.49 ± 0.43
Fullband IP-APA	11.69 ± 0.64	18.80 ± 0.40
Subband IP-APA	18.94 ± 0.68	29.15 ± 0.46

Table 5.1: Minimum ERLE during path changes, 8 kHz sampling rate.

	Minimum ERLE (dB)	
Algorithm	**Abrupt Change**	**Gradual Change**
Fullband NLMS	8.93 ± 0.44	11.69 ± 0.42
Subband NLMS	12.13 ± 0.53	15.05 ± 0.48
Fullband IP-NLMS	10.06 ± 0.39	17.86 ± 0.39
Subband IP-NLMS	13.20 ± 0.43	23.63 ± 0.35
Fullband IP-APA	10.06 ± 0.39	17.88 ± 0.40
Subband IP-APA	17.62 ± 0.38	28.01 ± 0.38

Table 5.2: Minimum ERLE during path changes, 16 kHz sampling rate.

kHz.

The minimum ERLE varies between sampling rates, structures and algorithms, and some trends become apparent when the tables are examined. For both types of echo path change, at both sampling rates, the subband configurations outperform their fullband counterparts. For all algorithms the minimum ERLE for both types of changes was lower for the 16 kHz sampling rate than for 8 kHz, approximately 1.5 dB lower for the abrupt change and between 0.5 and 1 dB dB lower for the gradual

change. The gap between fullband and subband differs significantly between algorithms. The minimum ERLE for subband NLMS is approximately 3 dB higher than that of fullband NLMS for both gradual and abrupt changes at both sampling rates. In contrast, the gap for IP-NLMS is approximately 3 dB for the abrupt change, but almost 6 dB for the gradual change. Similarly, while the minimum ERLE values differ, the gap for IP-APA is approximately 3 dB higher for the gradual change than it is for the abrupt change. The differences between NLMS and the proportionate algorithms are greater for the gradual change than for the abrupt change. For the abrupt echo path change, the minimum ERLE of subband IP-NLMS is about 4 dB higher than fullband NLMS, and subband NLMS even outperforms fullband IP-NLMS and IP-APA. However, for the gradual echo path change the superior tracking ability of the proportional algorithms becomes evident. For the 16 kHz sampling rate simulation the fullband NLMS ERLE is degraded to 11.69 dB, while the subband IP-NLMS ERLE is only degraded to 23.63 dB, a difference of almost 12 dB.

5.3 Experimental Data Results

In addition to the simulated changing echo path, the algorithms were compared using experimental white noise and speech data, recorded in a real changing echo environment using the setup described in section 3.1 with the Mitel 5140 hands-free terminal in the large conference room. A changing echo path was produced by moving a hand in front of speaker at a rate of approximately 1 Hz, and the SSNR for the recording was measured at approximately 35 dB. Figures 5.6 and 5.7 show the ERLE performance of the algorithms for one trial of the white input at 8 kHz and 16 kHz sampling rates respectively. The effect of the hand moving at an approximately constant rate can be seen in the periodic peaks and valleys of the ERLE curve. It is clear from

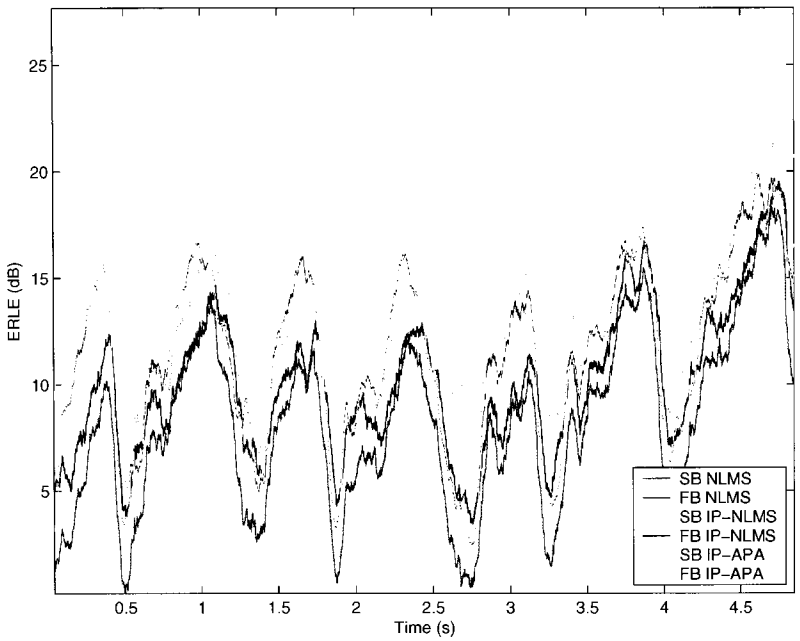


Figure 5.6: ERLE performance comparison of fullband and subband algorithms with white noise input in a real changing acoustic environment, 8 kHz sampling rate.

the figures that, for this trial, the subband IP-APA has a significant advantage over the other algorithms, fullband and subband. The valleys of the subband IP-APA ERLE curve are even higher than some of the peaks of fullband NLMS. While some qualitative observations about the ERLE performance can be made, the continously varying echo path makes quantitative statistical comparisons difficult.

Rather than comparing minimum or mean ERLE, the ERLE values from four trials were aggregated into ERLE histograms, showing the number of ERLE samples that fall in a given range. The width of the histogram is an indication of how variable the ERLE is over the measurement period, which in this case corresponds to how greatly an algorithm is affected by the changing echo path. Compared to an algorithm with poor tracking capabilities, the histogram curve of a fast tracking algorithm would therefore have a peak more to the right of the plot, corresponding to a high modal average ERLE, and would be more narrow, indicating a more consistent ERLE.

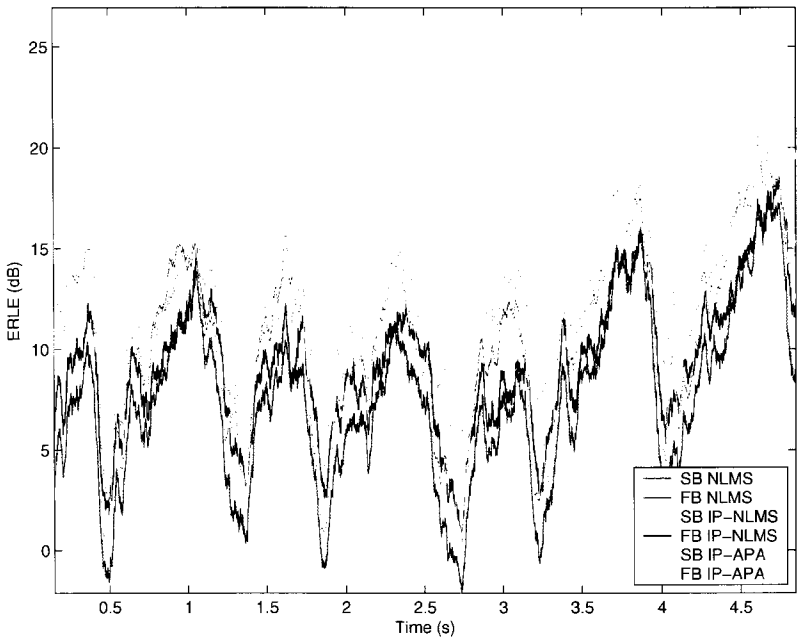


Figure 5.7: ERLE performance comparison of fullband and subband algorithms with white noise input in a real changing acoustic environment, 16 kHz sampling rate.

Figures 5.8 and 5.9 show the ERLE histograms for white noise inputs at 8 kHz and 16 kHz sampling rates respectively. The shape of the histograms is a result of the shape of the ERLE curves: the extreme right of the histogram corresponds to the ERLE of the peaks, the extreme left to the ERLE in the troughs, and the modal peak to the transition region. The differences between the 16 kHz and 8 kHz curves are not significant, the relative spacing of the histograms was the same for both sampling rates, although the 8 kHz histograms were positioned 1 – 2 dB higher than the 16 kHz curves. Examining the figures, parallels between the results from the recorded data and those of the simulated changing environment become apparent. The subband IP-APA stands out from the rest of the algorithms with a modal peak around 20 dB, almost 10 dB greater than subband IP-NLMS. In contrast the histograms of fullband IP-NLMS and IP-APA are almost indistinguishable. In general the curves for the subband adaptive filters indicate better tracking performance: for NLMS the subband

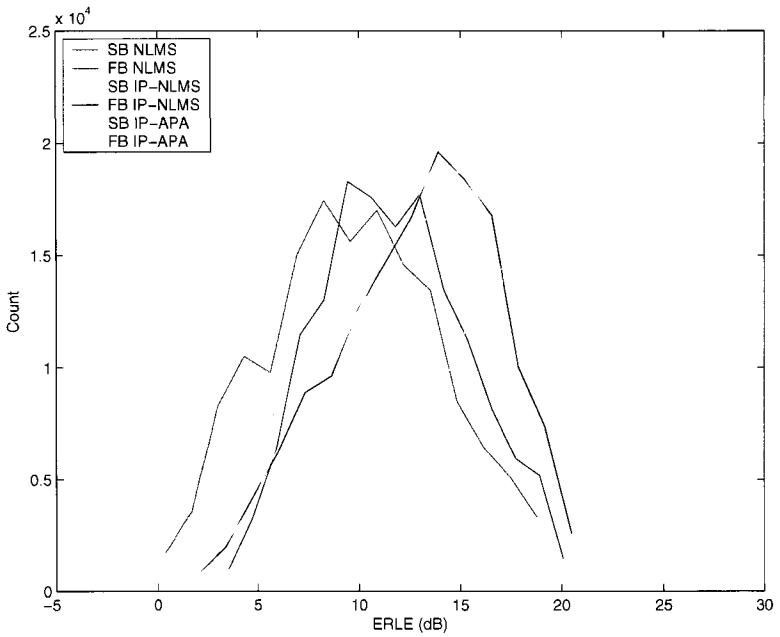


Figure 5.8: ERLE histogram for four trials of white noise input in a real changing acoustic environment, 8 kHz sampling rate.

curve is more narrow, and has a peak that is 3 – 5 dB higher than the fullband; for IP-NLMS the modal average of the subband version almost coincides with that of fullband IP-NLMS and IP-APA, but the subband histogram is more narrow and has less weight in the low ERLE regions, indicating better overall tracking performance.

The algorithms were also compared using speech data recorded in a real changing echo environment, a total of four trials were carried out using sentences from two male and two female speakers in the TIMIT database. Figures 5.10 and 5.11 show the ERLE results for one trial of female speech input. In contrast to the white input case the content of the 16 kHz sampling rate speech signals is different from that of the 8 kHz signals, consequently there there are differences in algorithm performance between the two sampling rates. For the speech used in this trial, no single algorithm clearly outpaces the rest for 16 kHz sampling rate: the fullband IP-APA has faster initial convergence, but the subband IP-APA has a consistently higher ERLE after

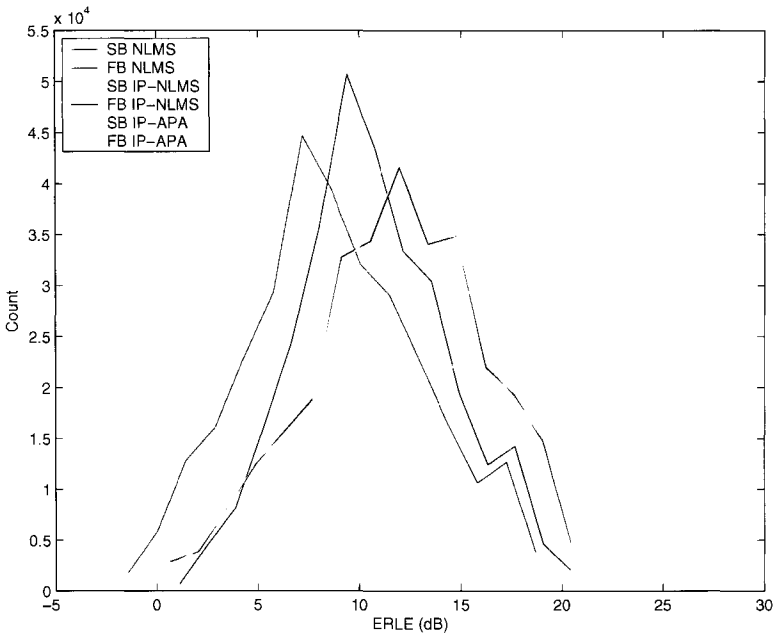


Figure 5.9: ERLE histogram for four trials of white noise input in a real changing acoustic environment, 16 kHz sampling rate.

approximately $t = 1$ second. For 8 kHz sampling rate the subband IP-APA has the highest ERLE for almost the entire segment.

As with the white noise date, the ERLE samples from the four trials were aggregated in to histograms, shown in figures 5.12 and 5.13, for 8 kHz and 16 kHz sampling rates. The variable nature of speech power and SNR leads to a wider ERLE histogram that is roughly bimodal, with the higher ERLE peak corresponding to the speech periods, and the other to the ERLE during silence. The histograms support the observation from the single trial that the performance gap between the subband and fullband versions of the algorithms is much smaller for speech input than for white noise. The histograms of subband and fullband IP-NLMS are very close, as are those of subband and fullband NLMS. Fullband and subband NLMS have the same maximum ERLE, but fullband NLMS has a lower minimum ERLE, indicating that it is affected more by the changing echo path. Similarly, the minimum ERLE

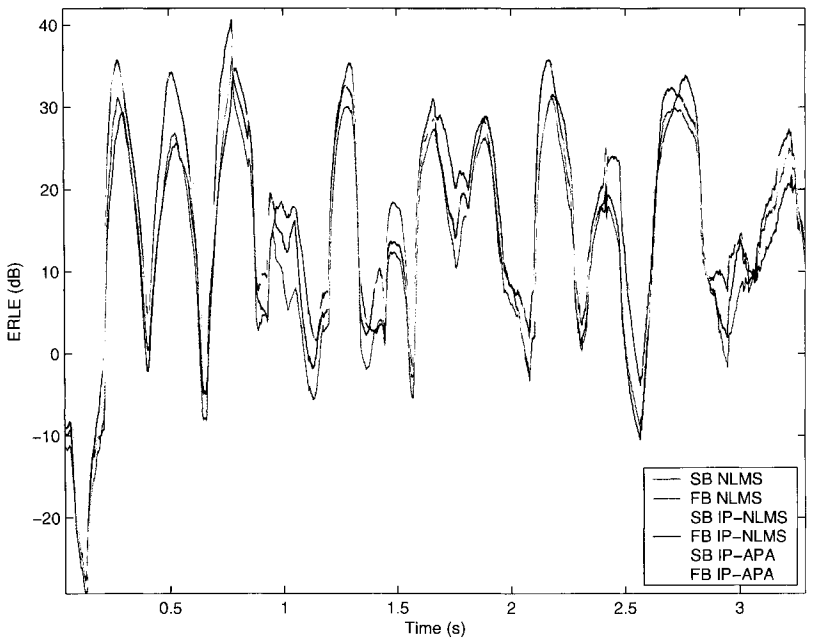


Figure 5.10: ERLE performance comparison of fullband and subband algorithms with speech input in a real changing acoustic environment, 8 kHz sampling rate.

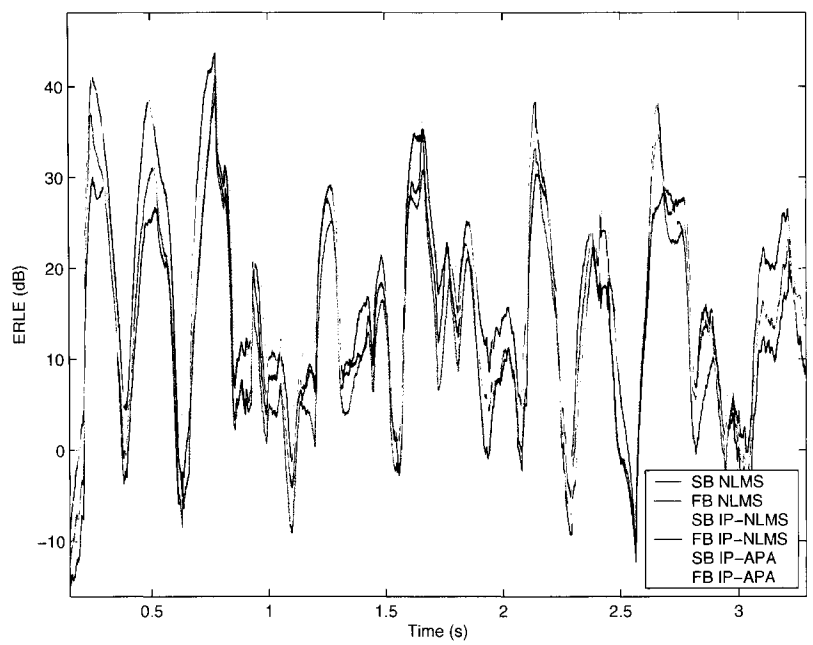


Figure 5.11: ERLE performance comparison of fullband and subband algorithms with speech input in a real changing acoustic environment, 16 kHz sampling rate.

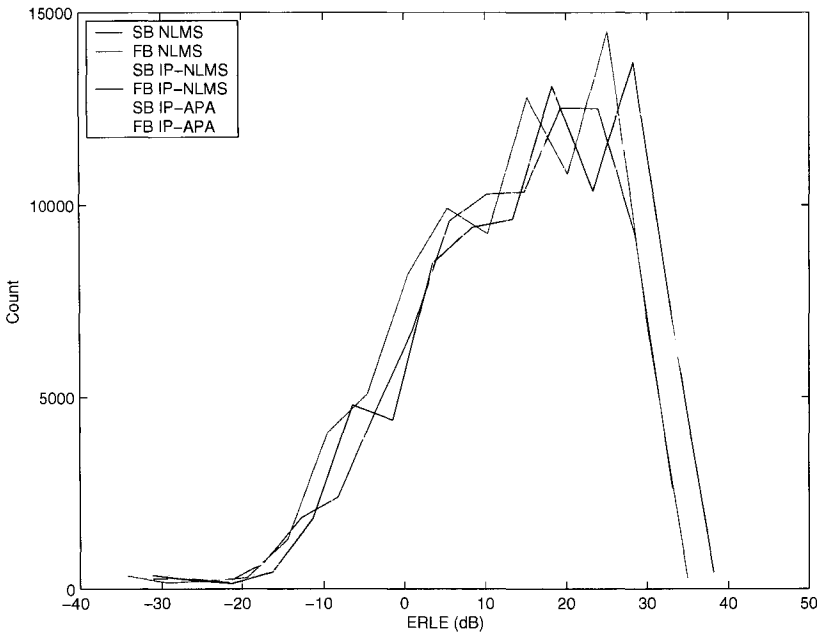


Figure 5.12: ERLE histogram for four trials of speech input in a real changing acoustic environment, 8 kHz sampling rate.

is lower for fullband IP-NLMS than it is for subband IP-NLMS, while the maximum ERLE is the same. Somewhat surprisingly, at 8 kHz sampling rate the modal average ERLE of fullband IP-NLMS is higher than subband IP-NLMS, and even exceeds fullband IP-APA. This may be due to the fact that the narrowband speech signal is less correlated than the wideband, so the subband structures and the APA have less of an advantage. At 16 kHz sampling rate, the fullband and subband IP-APA perform the best. While the maximum ERLE achieved by the fullband version is higher, the modal peaks of the subband version correspond to higher ERLE values, and the overall histogram is weighted more towards higher ERLE. For the 8 kHz sampling rate data, subband IP-APA outperforms the other algorithms by approximately 5 dB in terms of minimum, maximum and modal ERLE.

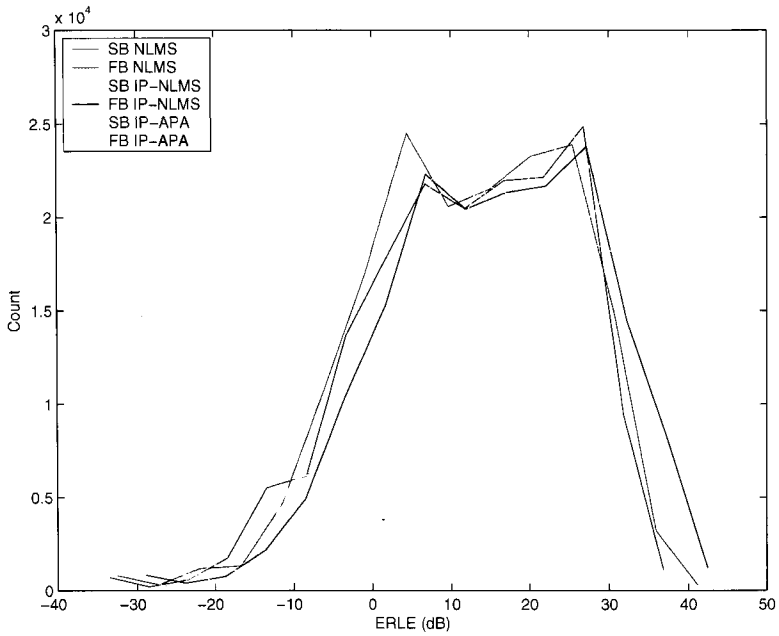


Figure 5.13: ERLE histogram for four trials of speech input in a real changing acoustic environment, 16 kHz sampling rate.

5.4 Summary

In this chapter the tracking performance of subband and fullband adaptive filtering algorithms was investigated. By examining impulse responses recorded with the same hands-free terminal under different conditions it was demonstrated that not all frequency regions are affected equally by echo path changes. For a given hands-free telephone terminal the magnitude of the low frequency response does not change significantly between rooms or when a small obstructing object is placed in front of the speaker, the effect on the high frequency region is greater. It was postulated that subband adaptive filters would be able to exploit this frequency disparity and offer better tracking than fullband filters, especially in the wideband case. This hypothesis was supported by simulation results from adaptive filters operating in a synthetic changing echo environment with white noise excitation; the subband adaptive filters all outperformed their fullband counterparts for both abrupt and gradual changes.

The hypothesis was also verified using experimental data recorded in real changing acoustic environment. When experimental white noise data was used, the advantage of the subband systems persisted, and they yielded a higher modal average ERLE. For speech recorded in a changing echo environment, the performance gap between the fullband and subband algorithms narrowed, such that IP-APA was the only algorithm for which the subband version clearly outperformed the fullband. Among the algorithms tested, subband IP-APA consistently offered the best tracking performance, for real and synthetic echo environments, and the the individual step-size algorithms offered superior reconvergence and tracking capabilities compared to traditional NLMS. The advantages of IP-APA were more evident for wideband speech input than for narrowband.

Chapter 6

Effect of Doubletalk Detectors on Echo Canceller Convergence

Doubletalk detectors are important components of an echo cancellation system; if adaptation is not halted during periods of near-end speech, the adaptive filter will quickly diverge resulting in poor echo cancellation performance. To prevent divergence doubletalk detectors are typically configured to have a very low probability of missing a doubletalk period, unfortunately this low miss probability results in a correspondingly high probability of a declaring doubletalk when there is none. If adaptation is repeatedly halted due to these false alarms, the convergence and tracking abilities of the adaptive filter are degraded. This chapter compares the convergence performance of subband and fullband echo cancellers employing normalised cross-correlation based doubletalk detectors. In the first simulation the echo cancellers are adapted in the presence of continuous near-end speech, in the second case there is high-level coloured background noise and a short period of near-end speech.

6.1 Convergence in the Presence of Doubletalk

The decision variable used in the cross-correlation based doubletalk detector, presented in equation (6.1), uses power comparison to determine when doubletalk is present.

$$\xi(n) = \sqrt{\frac{\underline{\hat{r}}_{xy}^{H}(n)\underline{\hat{h}}(n)}{\sigma_y^2(n)}} \tag{6.1}$$

The power predicted by convolving the cross-correlation vector with the echo path estimate is compared to the power of the microphone signal, therefore contributions of all frequencies are summed together with equal weighting to determine the final decision metric score. With a fullband doubletalk detector, adaptation of the entire filter is halted even if the doubletalk signal is narrowband, as with a voiced phoneme in wideband speech. In [52] it is demonstrated that if the doubletalk decision is made on a per-subband basis rather than globally, some bands are able to adapt during the doubletalk period, leading to better overall convergence for a subband structure. While the speech samples used in [52] were sampled at 8 kHz, this effect should be more prominent for 16 kHz sampled speech, as the signal power is even less uniformly distributed across the frequency spectrum and a segment of speech occupies a smaller instantaneous portion of the total bandwidth.

Simulations were carried out to compare the impact of doubletalk on subband and fullband adaptive filters. As the objective was to compare subband and fullband implementations, not to determine which algorithm was best for convergence in doubletalk conditions, only the NLMS algorithm was considered. The system distance convergence of fullband and subband NLMS adaptive filters with normalised cross-correlation based doubletalk detectors was compared for speech inputs at 8 kHz and 16 kHz sampling rates. The same number of subbands ($M = 8$) and decimation rate ($D = 4$) were used at both sampling rates, and the 8 kHz fullband and subband

filters used half the taps of their 16 kHz counterparts: 1200 taps and 300 taps per subband respectively. The filters were allowed to adapt for 0.5 seconds before the doubletalk detector was enabled. The doubletalk detector was calibrated, using the methods described in [50], to achieve a probability of doubletalk miss of $P_m = 0.05$, and the doubletalk hangover period was set to $T_{hold} = 30$ ms as in [50]. For both calibration and simulation, background noise was added to the microphone signal with a segmental SNR of 35 dB. The background noise, recorded in a conference room, is from the heating, ventilation and air conditioning (HVAC) system and is lowpass in nature, as can be seen in the power spectral density plot in figure 6.1. As in [52] the doubletalk signal was filtered to create a more narrowband signal; in this case the filter was a two pole IIR filter designed to mimic the long-term average male and female speech power spectrum [66]. The doubletalk signal was scaled to yield a ratio of near-end to far-end speech power (NFR) of 0 dB. The 16 kHz sampling rate versions of the speech signals used in the simulation are shown in figure 6.2. Both signals are sentences from the TIMIT database, the far-end signal is spoken by a female speaker, and the near-end (doubletalk) signal is spoken by a male. The 8 kHz sampling rate versions of the signals were obtained by downsampling the 16 kHz signals.

The system distance plots for both the 8 kHz and 16 kHz setups are shown in figure 6.3. The plots are normalised so that the system distance at $t = 0.5$ seconds, when the doubletalk detectors were activated, is 0 dB. As expected the fullband system distance is constant during the doubletalk period, as the fullband filters are completely halted. In contrast, the subband system distance continues to decrease, albeit at a much slower rate, indicating that some of the bands are not halted by the doubletalk. After 3 seconds of adaptation with the doubletalk detector active, the system distance for the subband structure was 0.96 dB less than that of the fullband structure for 8 kHz sampling rate, and 1.59 dB less at the 16 kHz sampling rate.

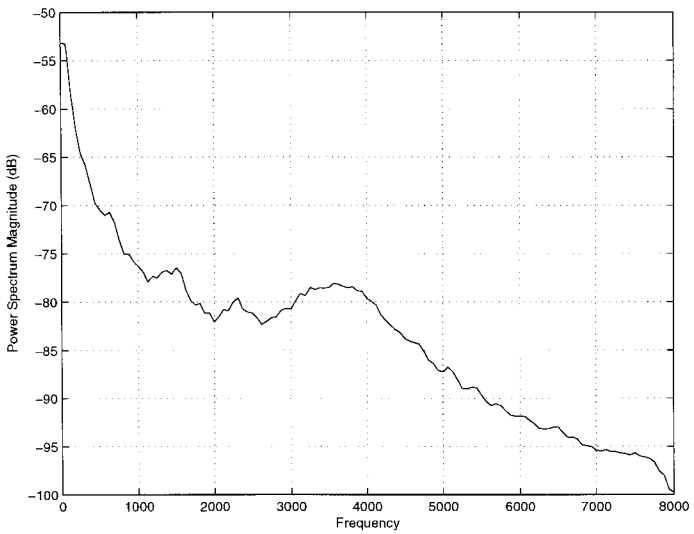


Figure 6.1: Power spectral density of background noise recorded in a conference room, shown for 16 kHz sampling rate.

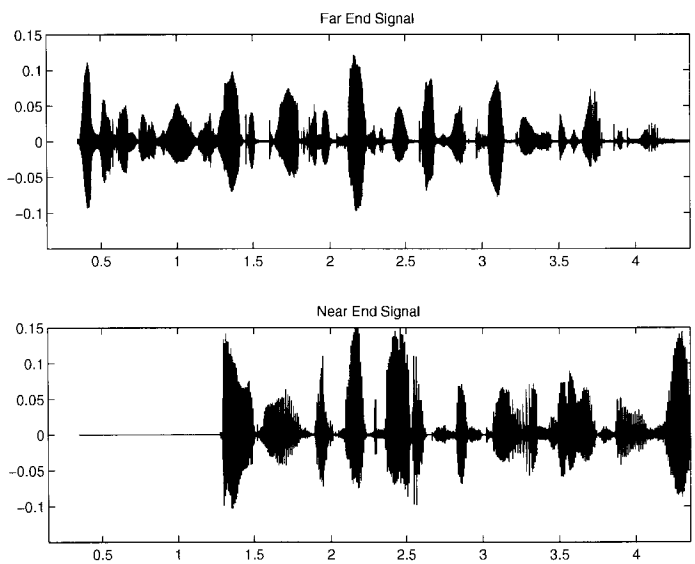
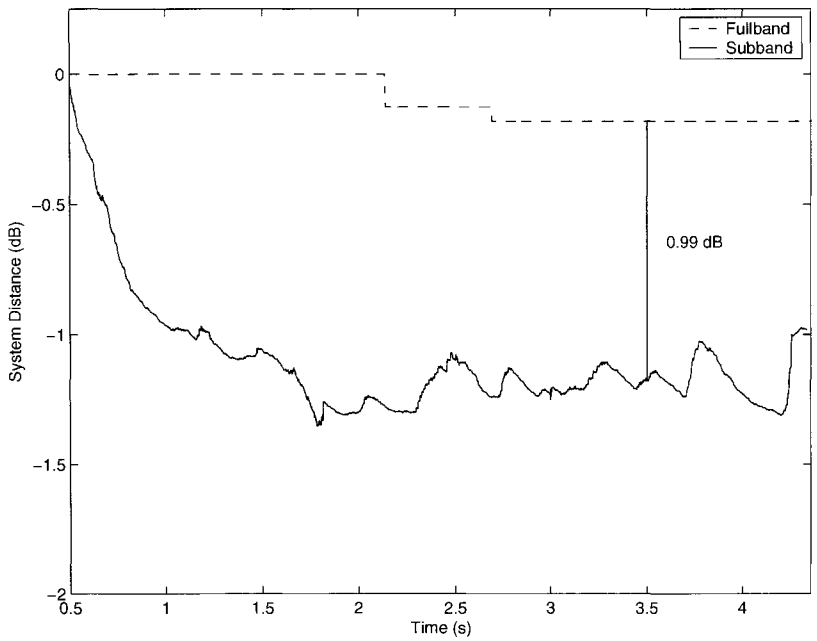


Figure 6.2: Far-end (top) and near-end (bottom) speech signals used to investigate convergence under doubletalk conditions.

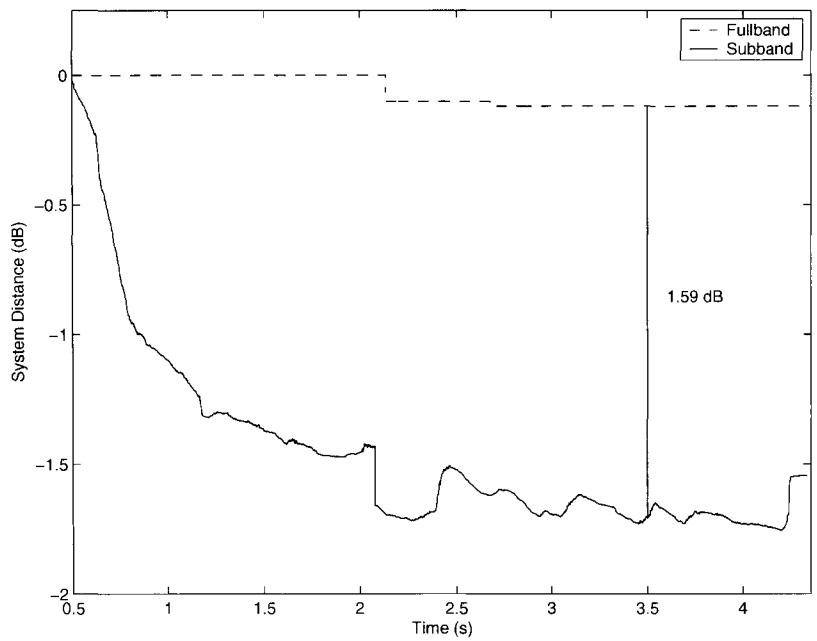
The convergence improvement of the subband configuration is also reflected in the ERLE, shown in figure 6.4. The average ERLE of the subband structure during the active periods is 0.6 dB higher than the fullband at 8 kHz, and 0.9 dB higher at 16 kHz. That the system distance and ERLE advantages of the subband system are greater at 16 kHz than 8 kHz sampling rate is possibly due to the doubletalk signal being relatively more narrowband at the 16 kHz than 8 kHz, so more bands covering a larger portion of the normalised frequency range can continue to adapt during the doubletalk period.

6.2 Convergence in the Presence of Narrowband Background Noise

Since the doubletalk decision variable uses the power of the microphone signal in the denominator, background noise in the near end room picked up by the microphone is included in the calculation and can affect its value. This problem is investigated in [49], where it is shown that local background noise lowering the value of the doubletalk decision variable can cause a false declaration of doubletalk. While stationary noise can be compensated for by lowering the decision threshold, the method proposed in [49] compensates for stationary and non-stationary noise by adaptively estimating the noise power and adding a noise correction term to the numerator of the decision variable in equation (6.1). However, even with noise correction the variable nature of speech power can result in a low instantaneous SNR, which lowers the detection variable and causes doubletalk to be declared. As with near-end speech, if the near-end background noise is narrowband the doubletalk detection variable of a subband doubletalk detector may only drop below the threshold in some bands, allowing the others to continue adapting.

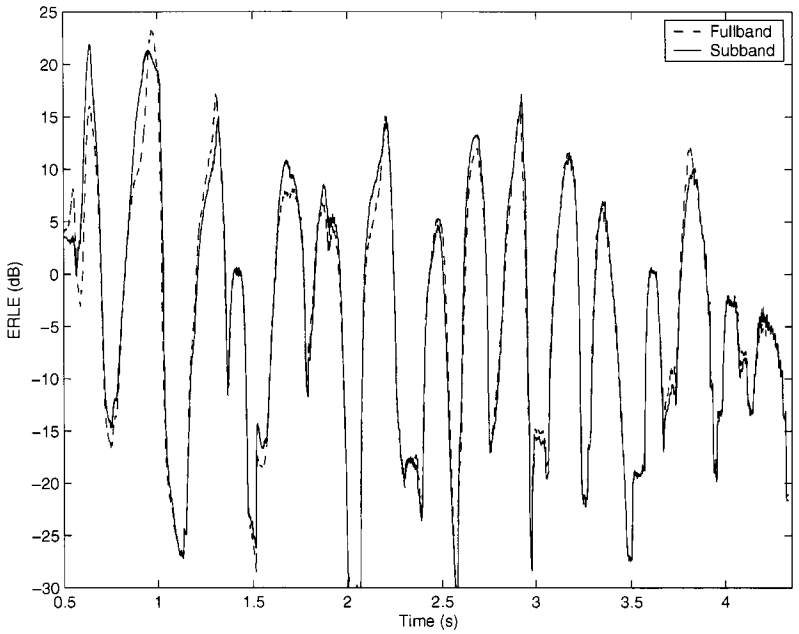


(a) 8 kHz sampling rate

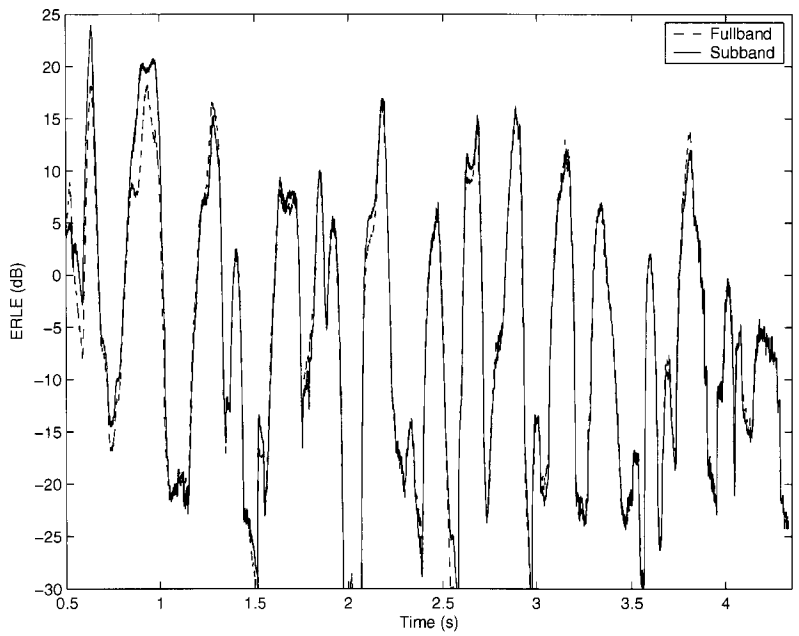


(b) 16 kHz sampling rate

Figure 6.3: Convergence of doubletalk detector controlled fullband (dashed line) and subband (solid line) echo cancellers in the presence of narrowband near-end speech.



(a) 8 kHz sampling rate



(b) 16 kHz sampling rate

Figure 6.4: ERLE for doubletalk detector controlled fullband (dashed line) and subband (solid line) echo cancellers in the presence of narrowband near-end speech.

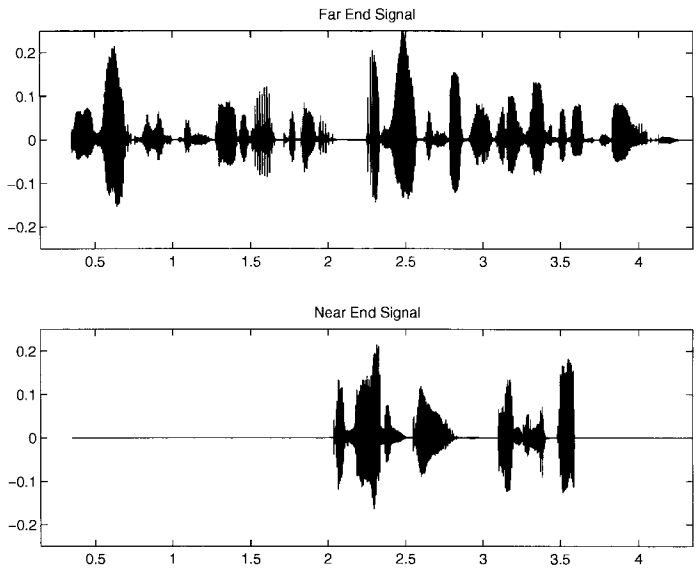


Figure 6.5: Far-end (top) and near-end (bottom) speech signals used to investigate convergence in the presence of narrowband background noise.

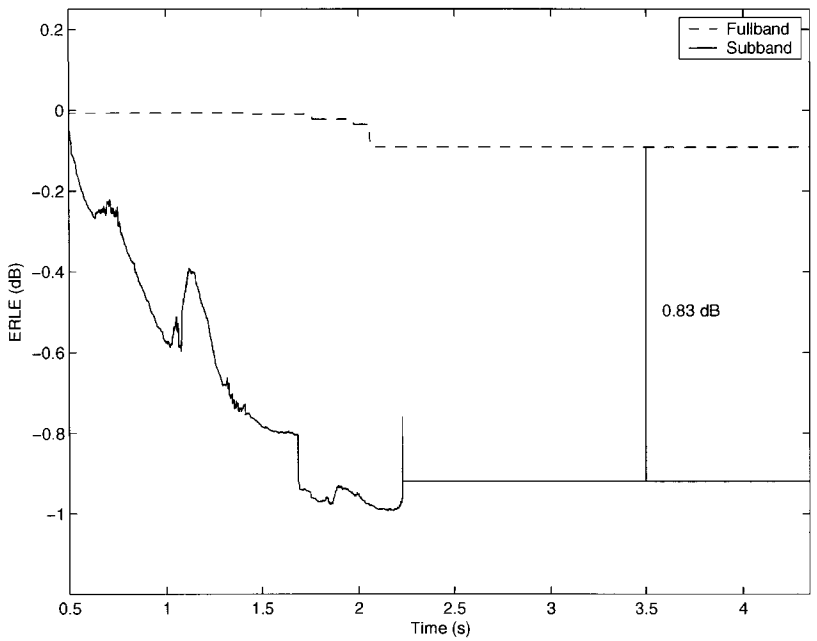
To compare the convergence of fullband and subband NLMS echo cancellers in the presence of high level background noise, the recorded noise signal was scaled so that the segmental SNR was 20 dB, and the doubletalk detectors were then re-calibrated to achieve $P_m = 0.05$ at the increased noise level. The near and and far-end speech segments used for the simulation are both of male speakers from the TIMIT database and are shown in figure 6.5. As with the previous simulations, the filters were allowed to converge for 0.5 seconds before the doubletalk detectors were enabled.

Figure 6.6 shows plots of the system distance convergence of subband and fullband echo cancellers operating at 8 kHz and 16 kHz sampling rates. Once again the plots are normalised so that the system distance at $t = 0.5$ seconds, when the doubletalk detectors were activated, is 0 dB. Observing the regions of the plots where the system distance remains constant, it is evident that the fullband configurations are frequently halted due to background noise causing the doubletalk detection variable to drop below the decision threshold. It should be noted that lowering the fullband thresholds

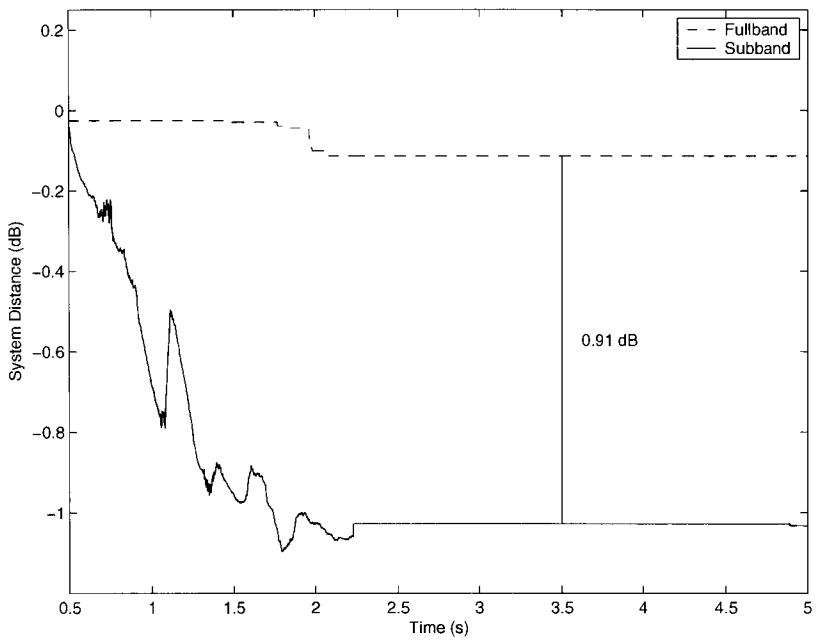
to allow more adaptation during singletalk resulted in the doubletalk period not being detected and the filters diverging. The rate of convergence of the subband structures slowed after the initial 0.5 seconds, indicating that adaptation was halted in some bands. After 3 seconds of adaptation with the doubletalk detector enabled the system distance of the subband structure had converged by 0.83 dB more than the fullband structure at 8 kHz and 0.92 dB more at 16 kHz sampling rate. From the ERLE plots of figure 6.7 it can be determined that the subband systems also had an average ERLE advantage during doubletalk of 0.9 dB at 16k, and 0.7 dB at 8 kHz.

6.3 Summary

These experiments seem to confirm the findings of [52], that there are cases where per-subband doubletalk detectors can offer convergence advantages over fullband doubletalk detectors. When a near-end disturbance is present, either background noise or near-end speech, a fullband doubletalk detector senses the disturbance and halts adaptation for the entire filter. If the disturbance is narrowband, a subband doubletalk detector will only halt adaptation in the bands where the disturbance is high. As the simulations presented in this chapter demonstrate, the ability of subband filters to continue adapting some bands in the presence of doubletalk results in improved convergence and better echo cancellation performance.

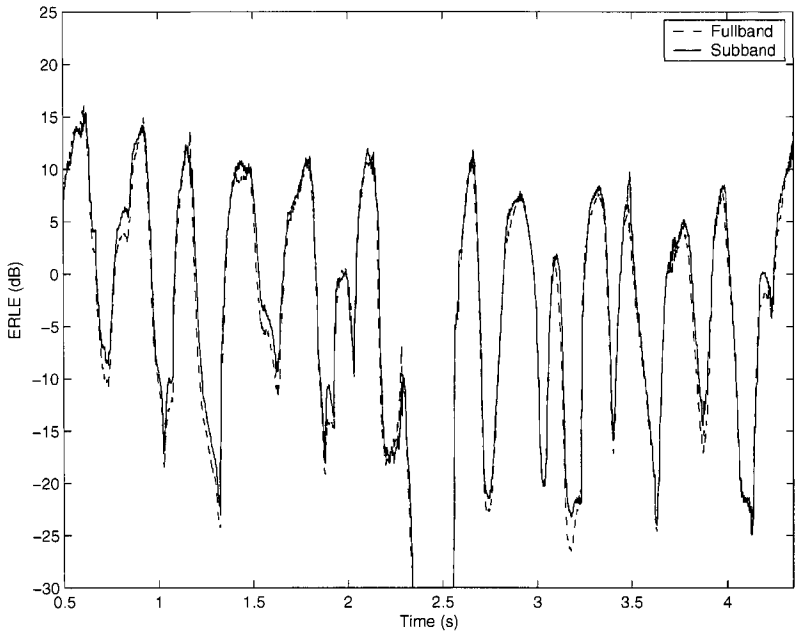


(a) 8 kHz sampling rate

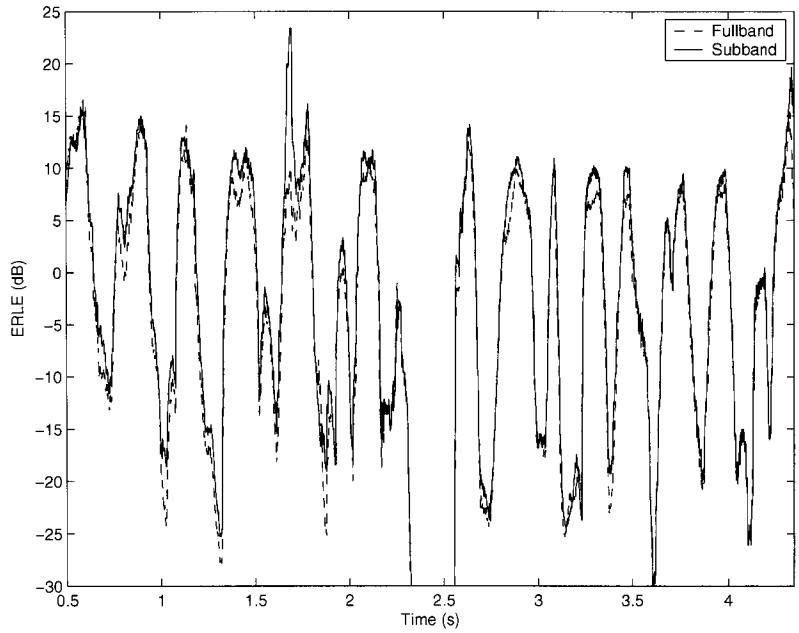


(b) 16 kHz sampling rate

Figure 6.6: Convergence of doubletalk detector controlled fullband (dashed line) and subband (solid line) echo cancellers in the presence of high level narrowband background noise.



(a) 8 kHz sampling rate



(b) 16 kHz sampling rate

Figure 6.7: ERLE for doubletalk detector controlled fullband (dashed line) and subband (solid line) echo cancellers in the presence of high level narrowband background noise. The doubletalk period is from approximately $t = 2$ to $t = 4$.

Chapter 7

Effect of Vocoder Distortion on Echo Cancellation

Effective echo cancellation is difficult if there is a VoIP vocoder in the echo path. A low bitrate vocoder such as G.722.2 is designed to produce an output that closely resembles the input as heard by the human ear, while using a very small number of bits per sample. The reconstructed waveform, while sounding like the original input, is very different from the original. Operating on a block-by-block basis, the linear predictive coder introduces significant non-linear, time-varying distortion.

Figure 7.1 shows a block diagram of an acoustic echo canceller operating in a VoIP network with vocoders in the echo path, as in the case of a centralized echo canceller.

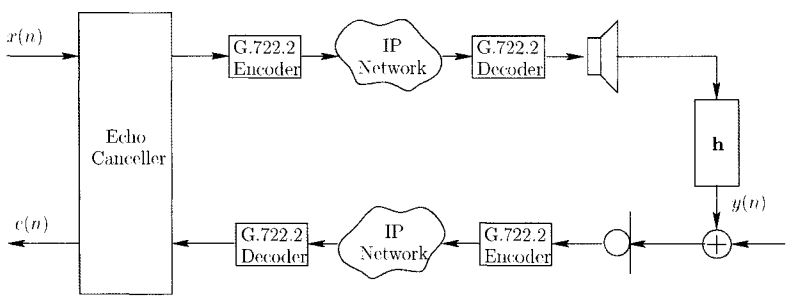
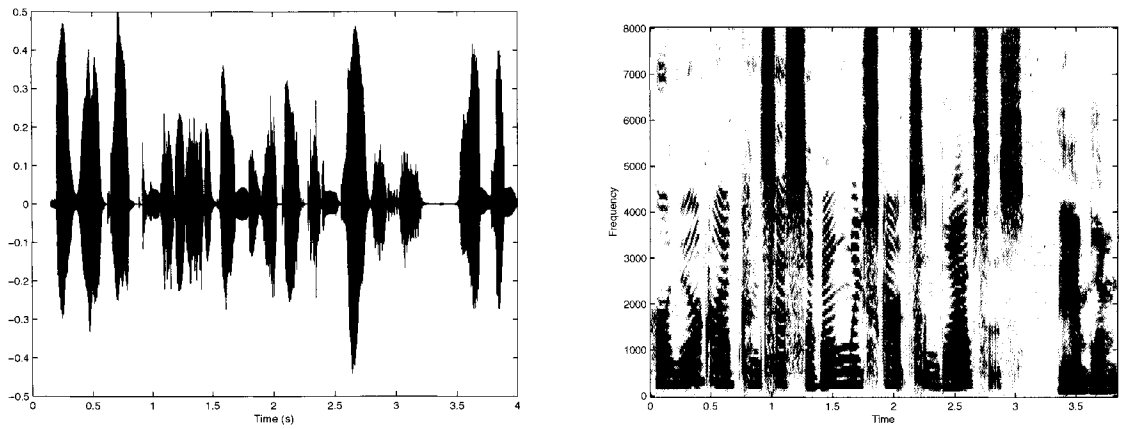


Figure 7.1: Acoustic echo canceller with G.722.2 wideband VoIP vocoders in the echo path.

In this thesis the IP network was not simulated, and the encoder and decoder were directly connected. When the undistorted signal is used as the reference signal for a linear echo canceller, and the echo path producing the desired signal contains vocoder distortion, the adaptive filter will attempt to model the transformation in order to create an accurate echo replica and minimize the output error. There are many steps in the the G.722.2 encoding and decoding process where distortion is introduced and while the linear time-invariant components of the vocoder, such as the fixed high-pass and lowpass filters, can be effectively modelled by an adaptive filter, the time-varying components are more difficult to capture. In [6] it is suggested that the superior performance of FAP in the presence of vocoders be a result of the strong tracking capabilities of the algorithm. Supporting that claim, in [55] it is stated that faster tracking algorithms result in less residual echo power when there is vocoder distortion in the echo path. If faster tracking improves the echo cancellation performance when vocoders are present, then fast tracking algorithms, such as IP-NLMS and IP-APA, should therefore offer better performance than NLMS. Furthermore, subband algorithms should also perform well, in accordance with the tracking performance results presented in chapter 5.

This chapter investigates the effects of G.722.2 vocoder distortion on acoustic echo cancellation. Using fullband NLMS as a representative algorithm the influence of the different coding rates offered by G.722.2 is examined to determine how vocoder distortion impairs echo cancellation performance. Next, the NLMS and IP-APA algorithms are used to demonstrate how fullband and subband adaptive filters are affected differently by the presence of vocoders. Finally, the ERLE performance of fullband and subband implementations of NLMS, APA, IP-NLMS and IP-APA are compared with and without vocoder distortion in the echo path, to determine which algorithm and structure best handles the non-linearity imposed by the vocoder. Since the G.722.2



(a) Speech signal.

(b) Specgtrogram of speech signal.

Figure 7.2: Speech segment used to investigate the effects of vocoder distortion on echo canceller performance.

vocoder is designed to encode speech, it is unable to properly represent white noise, therefore the the distortion imposed on a white noise signal is not representative of the distortion in practice. For this reason real speech excitation is used for all simulations in this chapter.

7.1 Effect of Coding Rate

The G.722.2 vocoder has nine coding rates. The main differences between them are the level of LPC model parameter quantization and the size of the fixed and adaptive codebooks used to generate the excitation sequence. When smaller codebooks are used, the excitation is likely to be further from the ideal excitation, resulting in a more distorted echo signal. In [6] excitation modelling was identified as having the greatest effect on ERLE performance. To examine the impact of the coding rate, the ERLE performance of a fullband NLMS adaptive echo canceller was compared for echo paths containing G.722.2 vocoders operating in different modes. Figure 7.2(a) shows the speech signal used to examine the impact of the coding rate, and figure

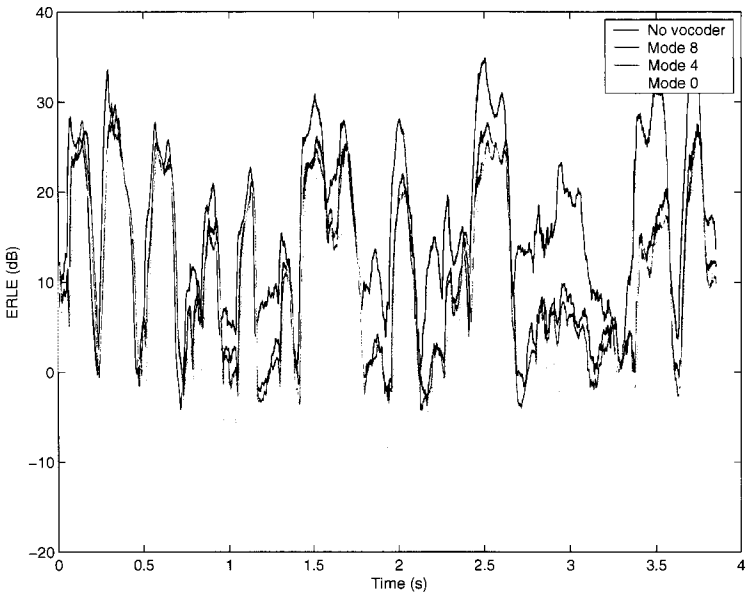


Figure 7.3: ERLE performance of fullband NLMS adaptive filter for different modes of G.722.2 vocoder distortion in the echo path.

7.2(b) shows the corresponding spectrogram. The speech signal contains both female and male speech segments from the TIMIT database. Figure 7.3 shows the resulting ERLE performance, where only the results from modes 0, 4 and 8 are shown for clarity of presentation.

The ERLE results of figure 7.3 can be analyzed by considering the speech input signal of figure 7.2 in the context of the operation of the G.722.2 vocoder. The G.722.2 encoder splits the input signal into two frequency bands which are encoded separately. The LP analysis and excitation search are performed on the lower frequency band, from 50–6400 Hz, and the high band signal, from 6400–7000 Hz, is generated using the low-band LP coefficients and a random excitation. Of the nine coding modes, only mode 8 transmits a gain factor to adjust the scaling of the highband excitation, the other modes infer the highband gain from the lowband signal. As a result the distortion imposed on speech frames with high frequency content is greater than

low frequency content frames. Furthermore, as discussed in [6] and [55], fricative-dominated, or noise-like speech segments cannot be accurately modelled by a LP coder. The ERLE difference between no the echo paths with no vocoder and those with vocoder distortion is therefore greatest when the speech content is non-harmonic and contains high frequency components, such as the periods around $t = 1.25$ and 1.75 seconds and the region from $t = 2.75$ to $t = 3.5$ seconds. These regions are all noise-like, and most contain high frequency content and therefore cannot be well represented by the LP coder, so the ERLE of the echo cancellers in the distorted echo path is several dB lower. In constrast, the region around $t = 0.5$ is composed of low frequency harmonic content, visible as distinct lines on the spectrogram. This signal can be accurately modelled by LP coder, and as a result the ERLE is almost the same for all vocoder modes and approaches the ERLE of the undistorted signal. Considering the different modes, it can be seen that the ERLE with the mode 4 distorted path is generally within 2 dB of the mode 8 distorted path. Furthermore, despite the fact that mode 0 is an emergency mode not intended for continuous use, the ERLE of the mode 0 distorted path is almost as high as the higher rate modes when the content is harmonic. In contrast, for the broad spectrum noise-like region from $t = 2.75$ to $t = 3.5$ seconds, the difference between mode 0 and the other modes is significant.

7.2 Effect of Adaptive Filter Structure and Algorithm

Since some of the distortion imposed by the G.722.2 vocoder is frequency dependent, subband structures are likely affected differently than fullband structures. Figures 7.4 and 7.5 directly compare the ERLE degredation experienced by fullband and

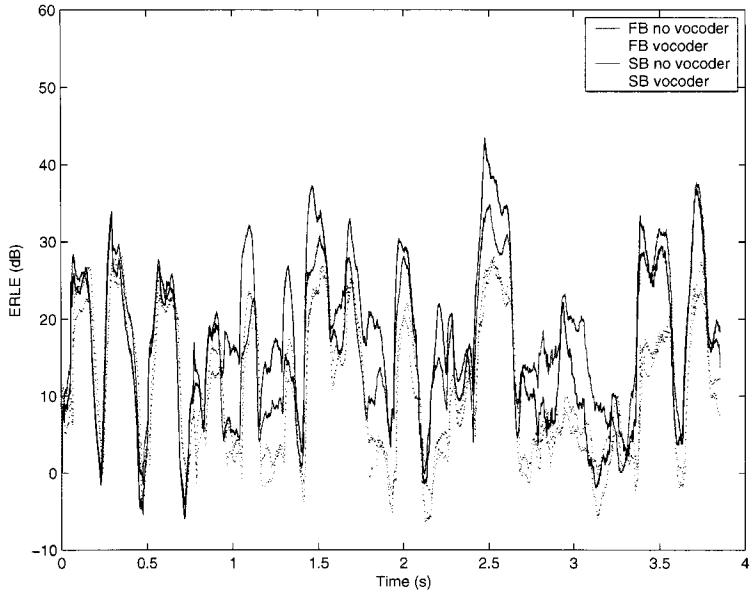


Figure 7.4: Comparison of ERLE degredation caused by G.722.2 vocoder distortion for fullband and subband NLMS.

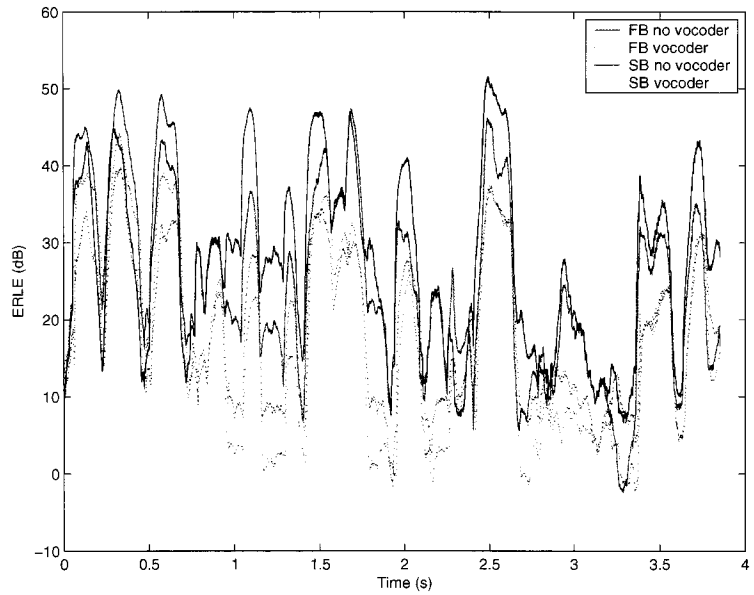


Figure 7.5: Comparison of ERLE degredation caused by G.722.2 vocoder distortion for fullband and subband IP-APA.

subband NLMS and IP-APA adaptive filters with a vocoder-distorted echo path, using the speech signal from figure 7.2 as input. Without vocoder distortion, the ERLE gap between the subband and fullband structures is quite variable. Frequently the ERLE performance is similar for the fullband and subband cases, but the subband structures generally perform better in regions with significant high frequency content, and especially when there is a shift from most of the energy being concentrated at low frequencies to most of the energy being concentrated at high frequencies, such as the region from $t = 1.0$ to $t = 1.5$ seconds. This is likely a reflection of the faster high frequency convergence that subband structures exhibit for speech inputs. When the vocoder is introduced to the transmit path, the differences in ERLE performance between the subband and fullband structures are significantly reduced. While the subband structures still perform somewhat better when there is high frequency content, their ERLE performance is degraded more than that of the fullband. For example at $t = 1.0$ second subband NLMS has an ERLE advantage of more than 11 dB over fullband NLMS when there is no vocoder, but the advantage drops to less than 4 dB when the vocoder is introduced; a similar effect can be observed for the IP-APA case. This is likely a reflection of the fact that the distortion introduced by the G.722.2 vocoder is greater at higher frequencies, so the subband structures are less able to exploit their faster convergence in the high frequency region to achieve an overall higher ERLE.

To compare the performance of different adaptive filtering algorithms in the non-linear channel, a total of eight trials were run using sentences from four male and four female speakers in the TIMIT database. The ERLE values from the eight trials were combined to form figures 7.6 and 7.7, which show the ERLE histograms of the algorithms with and without G.722.2 vocoder distortion in the echo path. When no vocoder is present, the APA adaptive filters outperform the NLMS implementations

in the fullband and subband configurations, which is to be expected for speech input. In keeping with the results of [6], APA also performs better than NLMS in the presence of vocoder distortion. The maximum, minimum, and modal average ERLE were all higher for the APA based filters, however APA was more impacted by the presence of the vocoder than NLMS. When the vocoder was included in the echo path, the maximum ERLE was approximately 10 dB lower for fullband and subband APA based algorithms, but only approximately 5 dB lower for NLMS. The modal averages were affected in a similar manner. The presence of vocoder distortion did not have a significant impact on the relative performance of the fullband and subband structures. In all cases the maximum ERLE for a fullband structures was marginally higher than the corresponding maximum subband ERLE for both the undistorted and distorted cases. For NLMS, the fullband and subband modal averages coincided but for APA and IP-APA, the subband versions had higher modal averages. With and without vocoder distortion, the ERLE of the fullband implementations of APA and IP-APA was more variable than that of the subband versions: the maximum ERLE was greater, but the minimum and modal ERLEs were lower.

Some more insight into the performance of the algorithms can be gained by examining the residual echo signals. Figures 7.8 and 7.9 present the power spectral density (PSD) of the echo signal and the error signals, for the undisorted and distorted channels respectively. The PSD plots are of the concatenation of the eight residual error signals, and were produced using MATLAB's `psd` function, which uses Welch's method to estimate the spectrum.

From figure 7.8 it is interesting to note that for the undistorted echo path, while fullband IP-APA has the lowest residual echo power in the 500 – 1000 Hz region, it has the most residual echo power in the high frequency region, above 6500 Hz, the residual echo signal power is even greater than the echo signal power. Overall

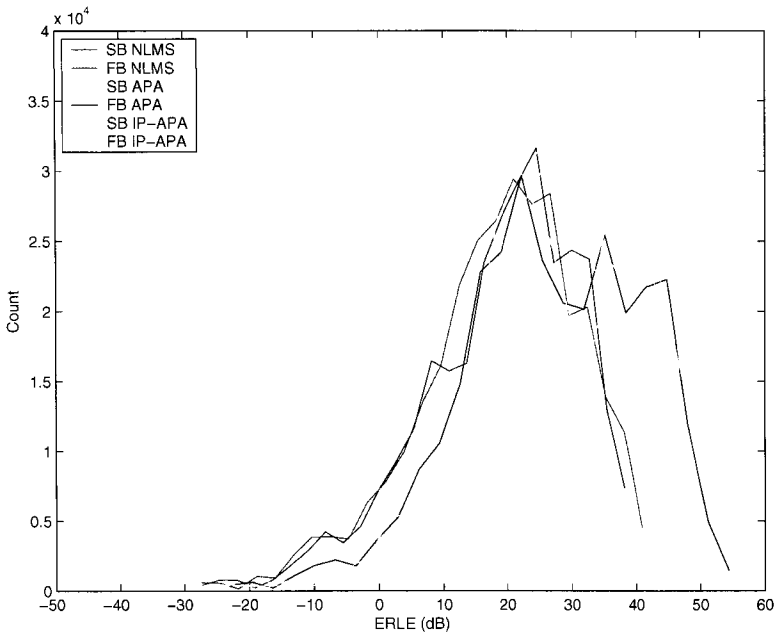


Figure 7.6: ERLE histogram for different adaptive filtering algorithms with no vocoder distortion.

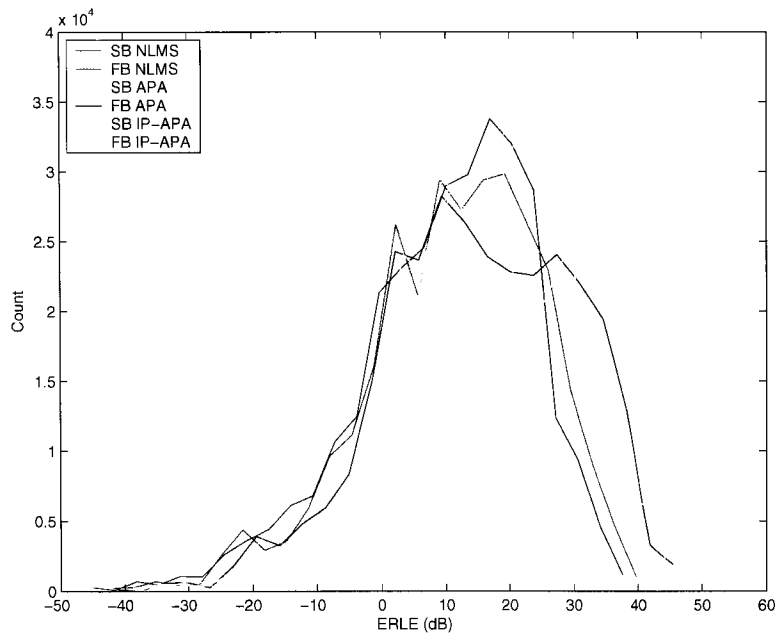


Figure 7.7: ERLE histogram for different adaptive filtering algorithms with G.722.2 vocoder distortion in the echo path.

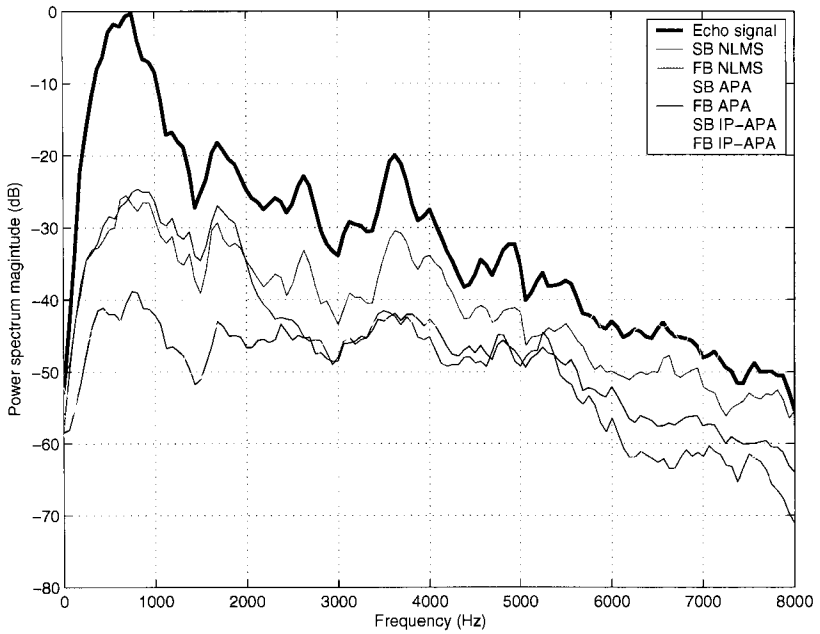


Figure 7.8: PSD of echo and residual echo signals with no vocoder distortion.

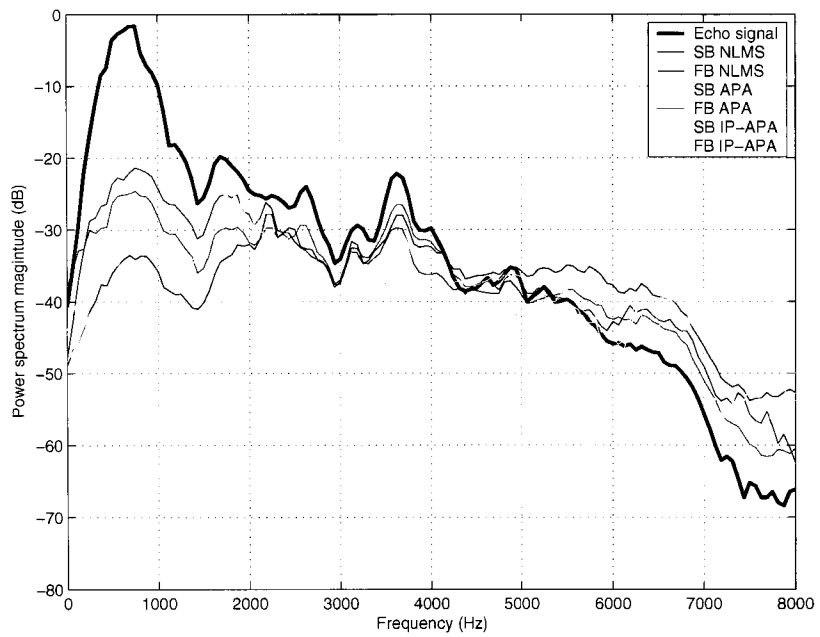


Figure 7.9: PSD of echo and residual echo signals with G.722.2 vocoder distortion in the echo path.

the residual echo signal of IP-APA is spectrally flatter than the other algorithms. This is an indication that IP-APA is achieving a lower overall MSE, and higher maximum ERLE, by sacrificing echo cancellation in the low energy high frequency region for better cancellation in high energy low frequency region. The fullband APA residual error signal is also spectrally flatter than that of NLMS, but not as flat as IP-APA. Unlike the APA based algorithms, NLMS does multiply the signal vector by the inverse correlation matrix to produce a whiter gain vector, so the residual error spectrum follows the error signal spectrum more closely. Similarly, in the subband case, any error whitening can only exist within a given subband, as each band adapts independently, so the level of cancellation is more uniform across the spectrum. It is also notable that the difference in residual echo power between the fullband and subband structures is not consistent across all frequencies. Below approximately 2 kHz, the fullband versions of all three algorithms have lower residual echo power than their subband counterparts, and above 2 kHz the residual power is lower in the subband case. These effects persisted when vocoder distortion was introduced to the echo path, as can be seen in figure 7.9. The gap between fullband and subband is however smaller in the nonlinear echo path, especially for the higher frequencies. The difference between algorithms is also smaller in the distorted echo path, although the fullband IP-APA still has the flattest residual echo power spectrum.

Comparing figures 7.8 and 7.9 it can be seen that, as expected the impact of the vocoder distortion on the cancellation performance is not uniform across all frequencies. In the region around 500 - 1000 Hz the residul echo power is 20 – 35 dB lower than the echo power, depending on the algorithm, compared to the undistorted case where the residual echo power is 25 – 45 dB below the echo power. In contrast, above 1 kHz the residual echo power in the distorted channel approaches the echo power, and above 5 – 6 kHz the residual echo power is higher than the echo signal power

for all of the implementations. Also note that the orignal echo signal power in the high frequency region for the distorted path is significantly reduced compared to the undistorted case, as a result of the lowpass filtering and high band signal generation process of the G.722.2 vocoder.

7.3 Summary

This chapter investigated the effects of vocoder distortion on acoustic echo cancellation. Vocoders in the echo path impose non-linear, time-varying distortion on the echo signal, which can degrade the performance of a linear echo canceller. For the G.722.2 vocoder, it was found that speech with low frequency harmonic content can be well modelled so there is little distortion, and the echo cancellation is not significantly affected. Even when the vocoder is operating at the lowest rate, the ERLE for this type of signal is still acceptable. In contrast, broadband speech segments containing high frequency energy are not well modelled by the vocoder, and echo cancellation performance is significantly degraded. Since subband structures generally outperform their fullband counterparts when there is high frequency energy, and vocoders distort high frequencies more than low, the gap between fullband and subband structures is smaller with vocoder distortion. For a given algorithm, the overall ERLE performance offered by the fullband and subband structures in the distorted echo path was comparable. The faster tracking of the subband algorithms seen in chapter 5 did not, therefore, carry over to improve the echo cancellation in the presence of vocoder distortion. In general however, the faster tracking algorithms performed better, which is in keeping with the results for the narrowband vocoder case presented in [6]. The fast tracking IP-APA performed the best, although the gap between the algorithms narrowed when the vocoder was present in the echo path.

Chapter 8

Complexity Reduction and Stabilisation of IP-APA

In the previous chapters the IP-APA was shown to offer fast initial convergence and superior tracking ability, providing the best echo cancellation performance for both changing and vocoder distorted echo environments. The main shortcoming of the algorithm is the high complexity associated with it.

In this chapter, two enhancements to the IP-APA algorithm are proposed. First an online regularisation is developed based on existing regularisation schemes for APA. Next, a new, fast (reduced computational complexity) version of IP-APA is developed, based on the Gauss-Siedel Fast Affine Projection Algorithm (GS-FAP). The online regularisation from the first part stabilises the required matrix inversion, allowing the computationally efficient Gauss-Seidel algorithm to be employed.

8.1 Regularisation of Improved Proportionate APA

In the description of APA and IP-APA presented in Algorithm 2 in chapter 2, the regularisation matrix $\delta\mathbf{I}$ is a matrix with a small positive constant along the diagonal, and its purpose is to prevent numerical instability due to ill-conditioning in the correlation matrix for coloured inputs. For the original GS-FAP implementation in [26] the regularisation is only applied during initialisation; the matrix inversion is not stabilised, leaving it vulnerable to numerical precision effects. While regularisation is desirable, selecting a fixed regularisation parameter is challenging, especially given the variable nature of speech power. If the regularisation is too small, the matrix inverse is not adequately stabilised, possibly leading to divergence. Oversampled subband adaptive filters are especially problematic in this regard, as the highly coloured subband signals that result from oversampling lead to a poorly conditioned correlation matrix to be inverted. However, if the regularisation is too large adaptation is slowed, leading to poor convergence and tracking.

In [67] an online regularisation is proposed that fulfils two purposes. Unlike a fixed regularisation, an online regularisation can effectively stabilise the matrix inverse for variable power input signals such as speech. The parameter proposed in [67] is based on signal energy, and therefore also serves as a step-size control, slowing adaptation when near-end signal power is high, as in low SNR or doubletalk conditions. A similar online regularisation is desirable for IP-APA, as its fast convergence and tracking capabilities also mean that it can quickly diverge in response to numerical instability or near-end disturbances such as doubletalk.

A regularisation for IP-APA will presently be developed by following the derivation of [67], but including the individual step-size matrix $\mathbf{A}(n)$. In the derivation $\delta(n)$ will

denote the regularisation parameter for standard APA, and $\delta'(n)$ will denote the regularisation for IP-APA.

First recall that $\mathbf{A}(n)$ is a diagonal matrix with diagonal elements a_l, $l \in \{0, \ldots, N-1\}$, computed as:

$$\begin{aligned} a_l(n) &= \frac{\kappa_l(n)}{\|\underline{\kappa}\|_1} \\ &= \frac{1-\alpha}{2N} + (1+\alpha)\frac{|\hat{h}_l(n)|}{2\|\underline{\hat{h}}(n)\|_1 + \epsilon}. \end{aligned} \tag{8.1}$$

It is easy to verify that $\mathbf{A}(n)$ has a trace of unity, and can be approximated by the mean of its diagonal as:

$$\mathbf{A}(n) \approx \frac{1}{N}\mathbf{I} \tag{8.2}$$

Note that when $\alpha = -1$, the diagonal elements in $\mathbf{A}(n)$ are all equal, $a_l = 1/N$, and IP-APA reduces to traditional APA.

The regularisation in [67] is based on minimising the system distance, therefore the cost function used to derive the optimal regularisation parameter is the norm of the system distance given by:

$$\underline{\Delta} = \underline{h}(n) - \underline{\hat{h}}(n)$$

$$\delta_{opt}(n) = \delta(n) \text{ such that } \mathcal{E}\{\|\underline{\Delta}(n)\|^2 \text{ is minimised}$$

If the echo path is not changing, the system distance can be calculated recursively as [67]:

$$\underline{\Delta}(n+1) = \underline{\Delta}(n) - \mathbf{A}(n)\mathbf{X}(n)(\mathbf{X}^H(n)\mathbf{A}(n)\mathbf{X}(n) + \delta'(n)\mathbf{I})^{-1}\underline{e}(n) \tag{8.3}$$

Applying the approximation of (8.2) the whitening matrix inverse becomes:

$$\begin{aligned} [\mathbf{X}^H(n)\mathbf{A}(n)\mathbf{X}(n) + \delta'(n)\mathbf{I}]^{-1} &\approx \left[\frac{1}{N}(\mathbf{X}^H(n)\mathbf{X}(n)) + \delta'(n)\mathbf{I}\right]^{-1} \\ &= N[\mathbf{X}^H(n)\mathbf{X}(n) + N\delta'(n)]^{-1} \end{aligned} \tag{8.4}$$

In [67], the following approximation is made, assuming white input, and $N \gg 1$:

$$[\mathbf{X}^H(n)\mathbf{X}(n) + \delta(n)\mathbf{I}]^{-1} \approx \frac{1}{N\sigma_x^2(n) + \delta(n)}\mathbf{I} \tag{8.5}$$

Applying this to (8.4) with $\delta'(n) = \delta(n)/N$ gives:

$$\begin{aligned}[\mathbf{X}^H(n)\mathbf{A}(n)\mathbf{X}(n) + \delta'(n)\mathbf{I}]^{-1} &\approx N\left[\frac{1}{N\sigma_x^2(n) + N\delta'(n)}\mathbf{I}\right] \\ &= \frac{1}{\sigma_x^2(n) + \delta'(n)}\mathbf{I}\end{aligned} \tag{8.6}$$

Substituting this result into (8.3)

$$\underline{\Delta}(n+1) \approx \underline{\Delta}(n) - \frac{1}{\sigma_x^2(n) + \delta'(n)}\mathbf{A}(n)\mathbf{X}(n)\underline{e}(n)$$

Using the approximation in (8.2)

$$\approx \underline{\Delta}(n) - \frac{1}{N\sigma_x^2(n) + N\delta'(n)}\mathbf{X}(n)\underline{e}(n) \tag{8.7}$$

Setting $\delta'(n) = \delta(n)/N$, the remainder of the derivation follows that of [67]. Thus the online regularisation for IP-APA is $1/N$ times that of APA. In [67] the cost function is differentiated with respect to δ, set to zero to find the minimum cost, and the pseudo-optimal regularisation parameter is shown to be:

$$\delta_{opt}(n) \approx N\sigma_x^2\frac{\sigma_e^2(n) - \sigma_\epsilon^2(n)}{\sigma_\epsilon^2(n)} \tag{8.8}$$

where, σ_e^2 is the power of the error signal, including near-end disturbances, and σ_ϵ^2 is the power of the error component due only to system mismatch, $\epsilon(n) = \mathbf{X}^H(n)\underline{\Delta}(n)$. Since computing $\epsilon(n)$ requires knowledge of $\underline{\Delta}(n)$, which is difficult to estimate in practice, an alternate regularisation is introduced in [68] to stabilise GS-FAP. The computation of $\delta(n)$ is simplified and modified to ensure stability even when there is

little near-end signal power (ie., during singletalk). In [68] the GS-FAP correlation matrix is replaced with a stabilised version given by:

$$\tilde{\mathbf{R}}_{xx}(n) = \mathbf{R}_{xx} + \underline{\delta}(n)\mathbf{I} \tag{8.9}$$

Where the elements $\delta^{(i)}(n)\,, i \in \{0, 1, \ldots, P-1\}$ of the regularisation vector $\underline{\delta}(n)$ are computed as:

$$\delta^{(i)}(n) = \overline{\max\{N|y(n)|^2, \delta_R^{(i)}(n)\}} \tag{8.10}$$

where

$$\delta_R^{(i)} = \left(\sum_{j=0, j\neq i}^{P-1} |x^*(n-j)x(n-i)| \right) - |x^*(n-i)|^2 \tag{8.11}$$

The overline in (8.10) indicates averaging by a dual time constant attack-release filter. The first term in (8.10) slows adaptation when the near end signal power is high, and the second term assures that the GS-FAP matrix inversion is stabilised, even when the near-end signal power is low. In [62], the complexity of the regularisation matrix computation is further reduced by using a scalar, rather than vector $\delta(n)$, and computing it in the simplified form given by:

$$\delta(n) = N \max\{(P-1)\overline{|x(n)|^2}, \overline{|y(n)|^2}\} \tag{8.12}$$

where the overline denotes time averaging. The corresponding regularisation for IP-APA is

$$\delta'(n) = \max\{(P-1)\overline{|x(n)|^2}, \overline{|y(n)|^2}\} \tag{8.13}$$

This form is easy to implement and computationally attractive, as $\overline{|x(n)|^2}$ is already available as the first entry in the correlation matrix. The regularisation of (8.12) has been demonstrated to prevent the nonlinear non-Wiener effects observed for subband APA [62], and to suppress divergence during doubletalk [69] without overly sacrificing convergence speed.

8.2 Improved Proportionate Gauss-Seidel Fast Affine Projection (IP-GS-FAP)

In [31], APA is combined with proportionate NLMS to form the proportionate APA (PAPA). A low complexity version of PAPA is presented for the case of $P = 2$, where the complexity is reduced by omitting the step-size matrix from the computation of the whitening matrix that is inverted, and incorporating the approximated error vector update of FAP. The alternative coefficient vector of FAP is not used, as the time-varying step-size matrix invalidates assumptions uses in the alternative coefficient vector derivation. The low complexity version of PAPA directly solves the set of linear equations to compute the 2×2 whitening matrix inverse.

Algorithm 6 introduces IP-GS-FAP, a version of the IP-APA that uses the fast error vector computation of FAP and the Gauss-Seidel matrix inversion of GS-FAP to produce a reduced complexity version of IP-APA that works for all values of the projection order P. The correlation matrix estimate is obtained by using the correlation vector estimate $\underline{r}_{xx}(n)$ as the first column in a Toeplitz matrix, thereby ensuring that the matrix is positive definite and that the Gauss-Seidel algorithm can be applied. The correlation vector is estimated using a window of size N, and the correlation matrix is multiplied by a factor $1/N$ to approximate the scaling effects of the step-size matrix, analogous to the scaling of the regularisation parameter for IP-APA relative to the APA case.

The algorithm incorporates a time-varying regularisation parameter that stabilises the Gauss-Seidel matrix inversion, and slows divergence during periods of high-level near-end disturbance. It may, however, be desirable to use a more aggressive regularisation parameter during the initial convergence period. During initial adaptation, the inverse correlation matrix estimate used as a starting point for the Gauss-Seidel

iteration is very poor and the step-size matrix and tap-weight vectors are all changing rapidly leading to potential numerical instability. More aggressive regularisation could be achieved using a method similar to the one employed in [68], where the dual time constant attack-release filter increases $\delta(n)$ when the instantaneous value is greater than the average, and decreases it when it is lower. This type of averaging allows $\delta(n)$ to be initialised to a large value to stabilise the early adaptation, without overly degrading convergence speed and steady state tracking.

Algorithm 6 (Improved Proportionate Gauss-Seidel Fast Affine Projection Algorithm (IP-GS-FAP)).

Initialisation

$\underline{\hat{h}}(0) = \underline{x}(0) = \underline{0}_N$ $\qquad$ $\mathbf{X}(0) = \mathbf{0}_{N\times P}$

$\underline{E}(0) = \underline{r}_{xx}(0) = \underline{0}_P$ $\qquad$ $\mathbf{R}(0) = \epsilon\mathbf{I}$

$\underline{b} = [1\ \underline{0}_{P-1}^T]^T$ $\qquad$ $\underline{p}(0) = \underline{b}/\epsilon$

Adaptation

FOR $n \geq 0$

$\underline{r}_{xx}(n) = \underline{r}_{xx}(n-1) + x(n)\underline{\xi}(n) - x(n-N)\underline{\xi}(n-N)$

Update $\mathbf{R}(n)$ using $\frac{1}{N}\underline{r}_{xx}(n)$

$\delta(n) = \max\{(P-1)\overline{|x(n)|^2}, \overline{|y(n)|^2}\}$

Solve $(\mathbf{R}(n) + \delta(n)\mathbf{I})\underline{p}(n) = \underline{b}$ for $\underline{p}(n)$ using one GS iteration

$e(n) = y(n) - \underline{\hat{h}}^H(n)\underline{x}(n)$

$\underline{\hat{h}}(n+1) = \underline{\hat{h}}(n) + \mu\mathbf{A}(n)\mathbf{X}(n)\underline{p}(n)e^*(n)$

$\mathbf{A}(n) = \mathrm{diag}\{\mathrm{a_0(n)}, \ldots, \mathrm{a_{N-1}(n)}\}$

where

$$a_l(n) = \frac{1-\alpha}{2N} + (1+\alpha)\frac{|h_l h(n)|}{2\|\underline{\hat{h}}(n)\|_1 + \epsilon}, l \in \{0, 1, \ldots, N-1\}$$

ENDFOR

Notation

$\underline{\xi}(n)$: uppermost P elements of $\underline{x}(n)$

$\overline{\underline{E}}(n)$: uppermost $P-1$ elements of error vector $\underline{E}(n)$.

$\delta\mathbf{I}$: regularisation matrix.

ϵ: small positive constant to prevent numerical instability.

8.3 Simulation Results

Subband and fullband implementations of the new IP-GS-FAP algorithm with projection order $P = 3$ were compared against standard GS-FAP of order $P = 3$ and NLMS to demonstrate the algorithm's speed of convergence for speech inputs, robustness to doubletalk, and tracking ability.

Figure 8.1 shows the speech signals used for the speech convergence simulations. The echo signal was generated by convolving a speech signal from the TIMIT database with a measured wideband echo path impulse response. Real coloured background noise, recorded in a conference room, was added to the echo signal at a segmental SNR of approximately 30 dB. The power spectrum of the backround noise signal is in figure 6.1. Both the fullband and subband IP-GS-FAP implementations employed the online regularisation of equation (8.13), although the regularisation for the subband implementation was increased by a factor of M/D to better stabilise the inversion of the ill-conditioned correlation matrix of the oversampled subband signals. The regularisation of equation (8.12) was used to stabilise the GS-FAP algorithm. Figure 8.2 shows the system distance convergence for the speech excitation shown in figure 8.1. According to [62], compared to the ideal fixed regularisation, the regularisation of (8.12) tends to oversuppress the adaptation in favour of greater stability. Despite this fact, the stabilised affine projection algorithms still outperform NLMS. Fullband IP-GS-FAP offered the fastest initial convergence and subband IP-GS-FAP had the deepest convergence over the speech segment.

Figure 8.3 shows the system distance for the case when near end speech was added to the echo signal with a segmental NFR of 3 dB; no doubletalk detector was employed, the only control was provided by the online regularisation. While NLMS would not be used without a doubletalk detector, it is useful to illustrate how

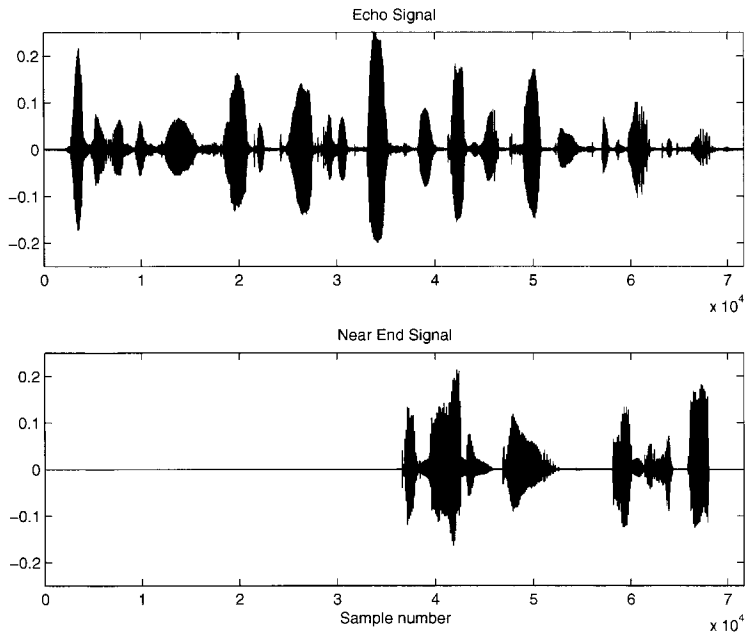


Figure 8.1: Echo and near-end (doubletalk) signals used for simulations.

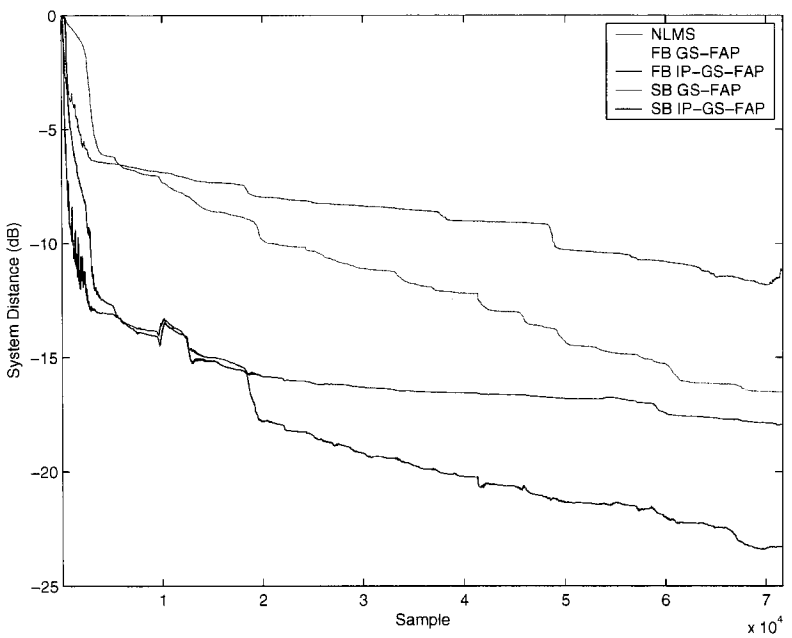


Figure 8.2: System distance simulation results for speech excitation.

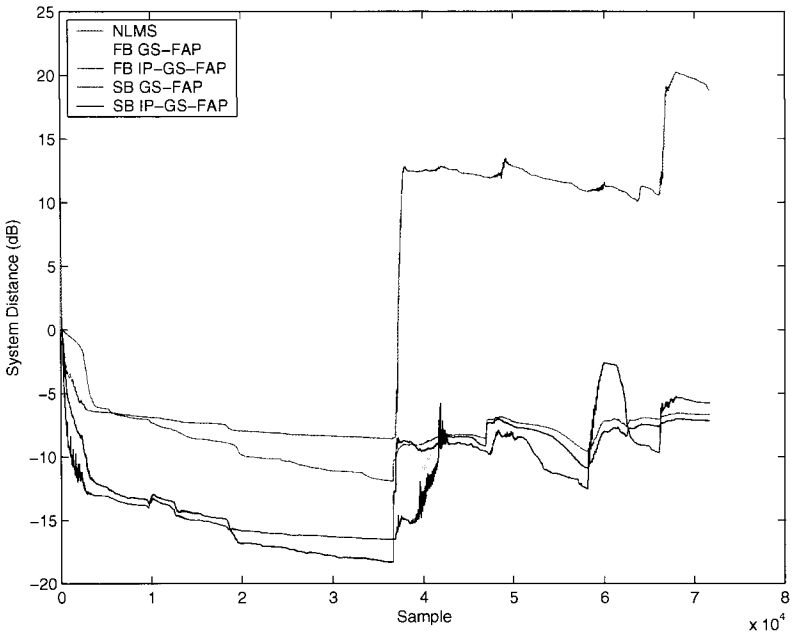


Figure 8.3: System distance simulation results for speech excitation and doubletalk conditions.

quickly it diverges, in the case of a slow reacting doubletalk detector, or a doubletalk detector miss. After fewer than 1000 samples (62 ms) of doubletalk NLMS had diverged from -10 dB system distance to 10 dB. While still affected, the stabilised IP-GS-FAP and GS-FAP algorithms are significantly more robust to the doubletalk than the unstabilised NLMS. During the doubletalk period the system distance of the subband versions remained below -5 dB. In order to make IP-GS-FAP and GS-FAP even more robust to doubletalk, the regularisation parameter can be modified as in [69] where the GS-FAP regularisation is chosen as $\delta(n) = \overline{|x(n)|^2}$ if $\overline{|x(n)|^2} > \gamma\overline{|y(n)|^2}$ and $\delta(n) = 20N\overline{|y(n)|^2}$ otherwise, where γ is a selectable parameter. This calculation of $\delta(n)$ ensures a very conservative adaptation when the near-end signal is high.

Figure 8.4 shows the ERLE performance in a changing echo environment, averaged over 25 trials, and table 8.1 presents the average and standard deviation of the minimum ERLE. The simulation conditions were the same as those in chapter 5:

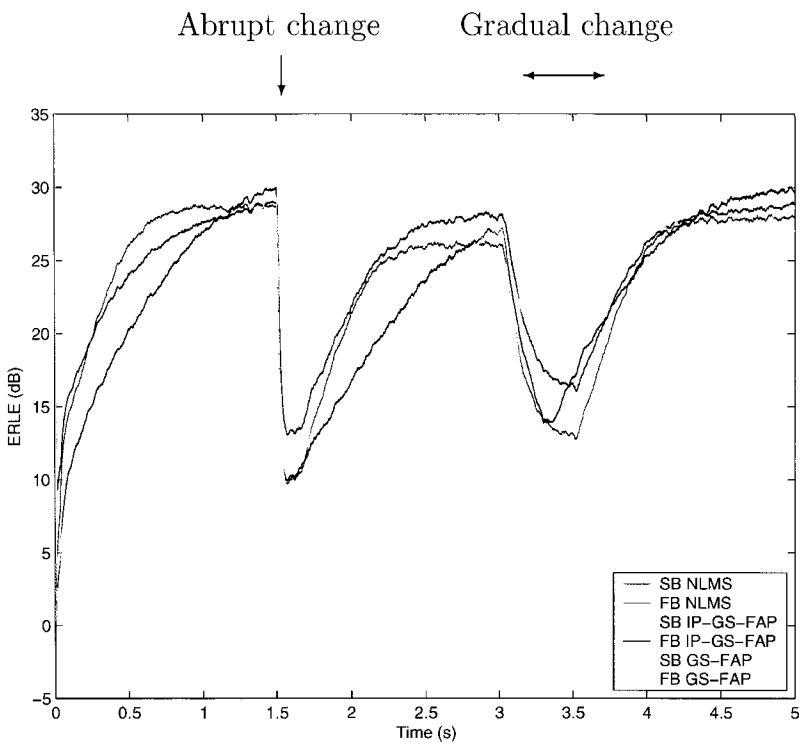


Figure 8.4: ERLE simulation results for white noise excitation and changing echo path conditions, online regularisation.

white noise was used as excitation, there was a sudden echo path change at $t = 1.5$ seconds and a gradual path change from $t = 3$ to $t = 3.5$ seconds. Subband and fullband versions of the online regularised IP-GS-FAP and GS-FAP were compared and fullband NLMS was included to serve as a reference.

When the online regularisation was used, fullband and subband IP-GS-FAP had a higher minimum ERLE than fullband NLMS for the abrupt echo path change, however fullband NLMS had faster initial convergence and reconvergence after the echo path change. Subband NLMS had the highest minimum ERLE during the abrupt change and also converged and reconverged faster than IP-GS-FAP. This is not unexpected, as the online regularisation is intended for variable power excitation, such as speech, and tends to over-suppress adaptation. Despite this fact, fullband and subband IP-GS-FAP still offered the best tracking performance during the gradual

	Minimum ERLE (dB)	
Algorithm	**Abrupt Change**	**Gradual Change**
FB NLMS	8.89 ± 0.24	11.97 ± 0.17
SB NLMS	12.20 ± 0.18	15.31 ± 0.13
FB IP-GS-FAP	9.13 ± 0.18	13.31 ± 0.19
SB IP-GS-FAP	9.84 ± 0.16	17.69 ± 0.10
FB GS-FAP	8.35 ± 0.16	8.33 ± 0.15
SB GS-FAP	8.33 ± 0.14	8.46 ± 0.16

Table 8.1: Average minimum ERLE over 25 trials, online regularisation used for GS-FAP and IP-GS-FAP.

echo path change: the minimum ERLE of fullband IP-GS-FAP was around 1.5 dB higher than fullband NLMS and the minimum ERLE for subband IP-GS-FAP was over 2 dB greater than NLMS. While IP-GS-FAP still offered reasonable performance, the convergence of GS-FAP, especially in the subband, was significantly slowed by the regularisation. The minimum ERLE for subband and fullband GS-FAP during the abrupt and gradual changes was comparable, and was the lowest of the algorithms tested. The minimum ERLE during the gradual change was over 3 dB lower than fullband NLMS.

Figure 8.5 and table 8.2 present the ERLE results for the same algorithms using a fixed regularisation of $\delta = \sigma_x^2/N$ for fullband IP-GS-FAP and $\delta = \sigma_x^2$ for fullband GS-FAP. The subband implementations required a higher value of δ_m to stabilise the matrix inverse: $\delta_m = (M/D)\sigma_{x,m}^2$ was used for subband IP-GS-FAP, and $\delta = 300(M/D)\sigma_{x,m}^2$ was used for subband GS-FAP, where σ_x^2 is the variance taken over the entire signal, and the subscript m denotes the subband number. These values of δ_m were found to provide the fastest stable convergence for white noise inputs.

When the optimum regularisation was used, the results were more comparable to

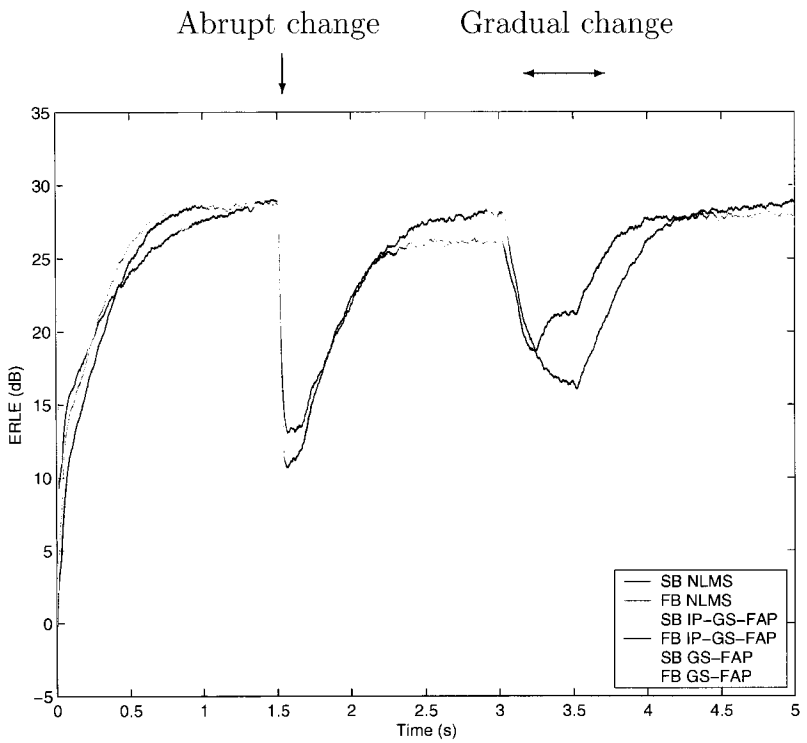


Figure 8.5: ERLE simulation results for white noise excitation and changing echo path conditions, optimal fixed regularisation.

	Minimum ERLE (dB)	
Algorithm	**Abrupt Change**	**Gradual Change**
FB NLMS	8.89 ± 0.24	11.97 ± 0.17
SB NLMS	12.20 ± 0.18	15.31 ± 0.13
FB IP-GS-FAP	9.96 ± 0.22	17.86 ± 0.19
SB IP-GS-FAP	10.69 ± 0.17	20.25 ± 0.15
FB GS-FAP	8.89 ± 0.24	11.97 ± 0.17
SB GS-FAP	8.69 ± 0.26	11.73 ± 0.19

Table 8.2: Average minimum ERLE over 25 trials, "ideal" regularisation used for GS-FAP and IP-GS-FAP.

those presented in chapter 5; the fullband IP-GS-FAP minimum ERLE averages were within one standard deviation of the full IP-APA results. Also, as with the previous results, subband IP-GS-FAP offered the highest minimum ERLE during abrupt and gradual change, and best tracking for the gradually changing path. Unlike the results in chapter 5 subband IP-GS-FAP did not have the constant 5 dB advantage over the other algorithms and structures. This is likely a result of the regularisation, which is needed to stabilise the Gauss-Seidel matrix inversion, suppressing the non-Wiener behaviour; the on-line regularisation in [62] was developed to suppress non-Wiener effects in subband APA. With the optimal regularisation, the performance of fullband GS-FAP is virtually indistinguishable from fullband NLMS, which is to be expected for a white input. On the other hand, subband GS-FAP does not perform as well as subband NLMS. GS-FAP is the only algorithm for which the subband structure does not offer significantly better tracking performance. A possible explanation is that the higher regularisation required to keep subband GS-FAP stable (higher than subband IP-GS-FAP) slows the rate of adaptation, degrading the tracking ability.

8.4 Summary and Discussion

In this chapter stabilisation and complexity reduction enhancements for IP-APA were proposed. The complexity of IP-APA was reduced by applying the fast error vector update and low-complexity matrix inversion techniques used in GS-FAP. An online regularisation algorithm designed for APA was modified to work with IP-APA. The regularisation suppresses adaptive filter divergence when the near-end signal is strong, and stabilises the correlation matrix inversion. Matrix inverse stabilisation is critical for oversampled subband structures where the highly coloured subband signals result in a poorly conditioned correlation matrix. Without regularisation subband

implementations of GS-FAP and IP-GS-FAP quickly diverge. In simulation, when compared with regularised GS-FAP, the regularised version of IP-GS-FAP was shown to offer better tracking in changing echo environments and faster convergence for speech excitation. Despite the more aggressive regularisation needed to stabilise the subband version of IP-GS-FAP, it still offered the best tracking performance of the algorithms tested. The regularised GS-FAP and IP-GS-FAP (fullband and subband) were also shown to be more robust to near-end speech disturbances than NLMS.

Employing the alternative coefficient vector of FAP could further reduce the complexity of IP-GS-FAP by simplifying the coefficient update step so that it requires N multiplies rather than PN. However, as stated in [31], the product of the step-size matrix and the excitation vector is not time-invariant because the step-size matrix varies from one iteration to the next. This conflicts with the derivation of the alternative coefficient vector in [24]. Limited simulations were carried out to determine the impact of using the alternative coefficient vector in IP-GS-FAP. With speech input, including doubletalk scenarios and measured speech data recorded in a changing echo environment, the use of the alternative coefficient vector in fullband regularised IP-GS-FAP did not appear to affect stability or convergence.

Chapter 9

Conclusions and Future Work

This thesis investigated the problem of acoustic echo cancellation in a wideband VoIP environment. Wideband acoustic echo cancellation differs from the narrowband case because the input speech signals are highly coloured, the echo path magnitude response has a greater spectral tilt, and the adaptive filters required to model a wideband echo path require more taps. Subband adaptive filtering has previously been proposed as a means to offset the high complexity and slow convergence experienced by fullband filters in the wideband environment. Oversampled subband structures allow the use of short subband analysis and synthesis filters, reducing the added delay. This is especially important in a VoIP network, where the signal path delay is already high. This thesis examined oversampled subband adaptive filtering and compared different fullband and subband adaptive filtering algorithms to determine which were best suited to the wideband VoIP environment.

It had been previously demonstrated [63] that when adaptive filter input signals are highly coloured, the adaptive algorithm can exploit that colouration to achieve lower mean square errors than the linear time invariant MSE optimal Wiener filter. The subband signals of oversampled subband adaptive filters are inherently coloured due to oversampling. In this thesis it was shown that, because of this subband

signal colouration, for sufficiently high step-sizes and especially with fast tracking algorithms, subband adaptive filters can be made to produce sub-Wiener output error levels for a broader range of inputs than fullband filters. This phenomenon was observed in simulation with synthetic data, and also with recorded experimental data.

Acoustic echo paths can change rapidly and frequently in response to movement in the near end room. Tracking ability is therefore an important quality in an acoustic echo canceller. The effect of echo path variations are frequency dependent: the high frequency region is more affected by echo path changes, making fast tracking especially important in the wideband case. Simulation results demonstrated that subband adaptive filters reconverge faster after an abrupt echo path change and track changing echo paths better than fullband filters. This was attributed to the observation that echo path changes do not affect all frequencies equally, so the subband filters did not need to reconverge in all bands. Each subband filter also has fewer coefficients to adapt than the fullband filter which may have contributed to the better tracking. The fast tracking of proportionate step-size algorithms was confirmed for both fullband and subband implementations. By proportioning more adaptation energy to the taps with larger weights, the proportionate algorithms were able to respond more quickly to the changing echo path than the fixed step-size algorithms. Of the algorithms and structures tested, subband IP-APA offered fastest reconvergence and the best tracking ability in simulated changing echo conditions. This was verified with experimental data in a real changing acoustic echo environment.

Near-end speech or high level background noise can result in divergence of the adaptive filter tap weight vector. To prevent this, echo cancellers typically employ a doubletalk detector to halt adaptation when the near end signal is high. It was found that when the near-end signal is narrowband background noise or narrowband speech, a subband system with an individual doubletalk detector for each band only

halts the adaptation in the bands where the near-end signal energy is high, while a fullband doubletalk detector halts adaptation for the whole filter. As a result the subband system is able to achieve deeper convergence over the same time frame.

When a centralised echo canceller is employed in a VoIP system, speech vocoders in the echo path impose significant time-varying non-linear distortion on the echo signal. When there is no vocoder distortion subband echo cancellers are generally better than fullband echo cancellers at removing high frequency echo content. However, since vocoder distortion is greatest in the high frequency region and for speech frames containing significant high frequency energy, echo cancellation in the high frequency region is degraded the most, and the performance gain exhibited by the subband structures is removed. For a given algorithm, the performance of the subband and fullband structures in the distorted echo path is very comparable. There is, however, variation between algorithms, and it was found that, fast tracking algorithms performed better than slow tracking. This is in agreement with previous work in the narrowband case. The IP-APA was found to offer the best echo cancellation performance in the distorted wideband echo path.

The IP-APA possesses desirable tracking ability, and superior performance in the vocoder distorted path, however its high complexity reduces its practicality for use with the long adaptive filters required for wideband acoustic echo cancellation. To alleviate this concern a modified version of the IP-APA, called IP-GS-FAP, was proposed that uses techniques from the GS-FAP algorithm to reduce the computational burden. An online regularisation technique was modified to work with the new algorithm. The regularisation acts as a time varying step-size to control adaptation and prevent divergence when far-end signal is low and near-end signal is high. The regularisation also acts to stabilise the Gauss-Seidel matrix inversion even when the matrix to be inverted is ill-conditioned, as in the case of an oversampled subband

correlation matrix. Simulations demonstrated that the stabilised IP-GS-FAP offers good tracking ability and fast convergence for speech excitation, and is more robust to near-end disturbances than NLMS.

One area where the work from this thesis could be expanded, is to consider psychoacoustic aspects of human hearing. As demonstrated in [70], for wideband echoes with long delays, as in a wideband VoIP system, steady state ERLE and MSE are not accurate indicators of whether or not an echo is audible. Time and frequency masking and hearing threshold effects have a greater impact on the perceived echo level in the wideband case. This is consistent with the findings in [66] and [37], where it is argued that higher frequency echoes are not as perceptually relevant as those in the low frequency range. Furthermore echo path magnitude responses and wideband speech signals are not homogeneous across the wideband frequency spectrum, both exhibit considerable spectral tilt. High frequencies are absorbed more quickly by the enclosure [4], and speech contains less energy in the high frequency. All of these factors point to possible complexity and perceptual performance gains to be obtained by processing subbands differently: using different algorithms, different echo cancellation techniques (eg., non-linear adaptive filtering), or allocating taps unequally between subbands as in [66]. The idea of treating disjoint frequency regions differently is used in [35], where a wideband stereophonic subband acoustic echo canceller is studied, and it is noted that simpler filter adaptation schemes can be used in upper subbands without a reduction in echo cancellation performance. Another possible extension of this work is the investigation of structures and algorithms for wideband stereophonic or multi-channel echo cancellation. As with wideband telephony, stereophonic telephony improves the subjective quality of conversations, and makes communication more transparent, however good echo cancellation is required to preserve the perceptual benefits.

List of References

[1] A. Gilloire, "Experiments with sub-band acoustic echo cancellers for teleconferencing," *International Conference on Acoustics, Speech, and Signal Processing. Proceedings.*, pp. 2141 – 4, 1987.

[2] P. P. Vaidyanathan, *Multirate Systems and Filter Banks.* Prentice Hall, 1993.

[3] M. M. Sondhi and D. A. Berkley, "Silencing Echoes on the Telephone Network," *Proc. IEEE*, vol. 68, no. 8, pp. 948 – 963, 1980.

[4] C. Breining, P. Dreiscitel, E. Hansler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, "Acoustic echo control. An application of very-high-order adaptive filters," *IEEE Signal Processing Mag.*, vol. 16, pp. 42 – 69, July 1999.

[5] ITU-T, "Talker echo and its control," Recommendation G.131, International Telecommunication Union, Nov. 2003.

[6] Y. Huang, "Effects of Vocoder Distortion and Packet Loss on Network Echo Cancellation," Master's thesis, Carleton University, 2000.

[7] M. Sondhi, "An adaptive echo canceller," *Bell Syst. Tech. Jour.*, vol. 46, pp. 497–511, 1967.

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

[9] J. James, B. Chen, and L. Garrison, "Implementing VoIP: A voice transmission performance progress report," *IEEE Commun. Mag.*, vol. 42, no. 7, pp. 36 – 41, 2004.

[10] T. Yensen, M. Parperis, I. Lambadaris, and R. Goubran, "Determining acoustic round trip delay for VoIP conferences," in *Proc. IEEE Second Workshop Multimedia Signal Processing '98*, pp. 161 – 166, Dec. 1998.

[11] 3GPP, "Adaptive Multi-Rate Wideband (AMR-WB) speech codec; General description," Technical Specification TS 26.171 V6.0.0 (2004-12), 3GPP, Dec. 2004.

[12] ITU-T, "Wideband coding of speech at around 16 kbit/s using Adaptive Multi-rate Wideband (AMR-WB)," Recommendation G.722.2, International Telecommunication Union, Jan. 2002.

[13] J. Markel and A. Gray, *Linear Prediction of Speech.* Berlin and New York: Springer, 1976.

[14] B. Bessette, R. Salami, R. Lefebvre, M. J. ek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. J. vinen, "The Adaptive Multirate Wideband Speech Codec (AMR-WB)," *IEEE Trans. Speech Audio Processing*, vol. 10, pp. 620–6, Nov. 2002.

[15] N. S. Jayant, J. D. Johnston, and Y. Shoham, "Coding of wideband speech," *Speech Commun.*, vol. 11, no. 2-3, pp. 127–138, 1992.

[16] M. M. Sondhi and W. Kellermann, "Adaptive echo cancellation for speech signals," in *Advances in speech signal processing* (S. Furui and M. M. Sondhi, eds.), pp. 327 – 356, New York: Marcel Dekker, 1992.

[17] S. Haykin, *Adaptive Filter Theory.* Prentice Hall, 4 ed., 2002.

[18] S. Weiss, L. Lampe, and R. Stewart, "Efficient implementations of complex and real valued filter banks for comparative subband processing with an application to adaptive filtering," *Proceedings of First International Symposium on Communication Systems and Digital Signal Processing*, vol. 1, pp. 32 – 5, 1998.

[19] B. Widrow and M. Hoff, "Adaptive switching circuits," *Wescon/89. Conference Record*, pp. 709 – 17, 1989.

[20] D. T. Slock, "On the convergence behavior of the LMS and the normalized LMS algorithms," *IEEE Trans. Signal Processing*, vol. 41, no. 9, pp. 2811 – 25, 1993.

[21] D. Morgan, "Slow asymptotic convergence of LMS acoustic echo cancelers," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 2, pp. 126 – 36, 1995.

[22] K. Ozeki and T. Umeda, "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties," *Electronics and Communications in Japan (English translation of Denshi Tsushin Gakkai Zasshi)*, vol. 67, no. 5, pp. 19 – 27, 1984.

[23] M. Rupp, "Family of adaptive filter algorithms with decorrelating properties," *IEEE Trans. Signal Processing*, vol. 46, no. 3, pp. 771 – 5, 1998.

[24] S. Gay and S. Tavathia, "The fast affine projection algorithm," *International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, vol. 5, pp. 3023 – 6, 1995.

[25] H. Ding, "A stable fast affine projection adaptation algorithm suitable for low-cost processors," *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, vol. 1, pp. 360 – 3, 2000.

[26] F. Albu, J. Kadlec, N. Coleman, and A. Fagan, "The Gauss-Seidel fast affine projection algorithm," *IEEE Workshop on Signal Processing Systems*, pp. 109 – 14, 2002.

[27] R. Barrett and M. B. et. al, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. Society for Industrial and Applied Mathematics, 1994.

[28] S. Makino, Y. Kaneda, and N. Koizumi, "Exponentially weighted stepsize NLMS adaptive filter based on the statistics of a room impulse response," *IEEE Trans. Speech Audio Processing*, vol. 1, no. 1, pp. 101 – 8, 1993.

[29] D. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancelers," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 5, pp. 508 – 18, Sept. 2000.

[30] J. Benesty and S. Gay, "An improved PNLMS algorithm," *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, vol. 2, pp. 1881 – 4, 2002.

[31] T. Gänsler, S. Gay, M. Shondhi, and J. Benesty, "Double-talk robust fast converging algorithms for network echo cancellation," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 6, pp. 656 – 63, 2000.

[32] W. Kellermann, "Kompensation akustischer echos in frequenzteil-bändern," *Aachener Kolloquium 1984*, pp. 322–325, 1984. (in German).

[33] I. Furukawa, "Design of canceller for broad band acoustic echo.," *Proc. Int. Teleconference Symposium*, pp. 232 – 239, 1984.

[34] W. Kellermann, "Analysis and design of multirate systems for cancellation of acoustical echoes," *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings.*, vol. 5, pp. 2570 – 2573, Apr. 1988.

[35] P. Eneroth, T. Gänsler, S. Gay, and J. Benesty, "Studies of a wideband stereophonic acoustic echo canceler," in *Proc. IEEE WASPAA '99*, pp. 207 – 210, Oct. 1999.

[36] D. R. Morgan and J. C. Thi, "A Delayless Subband Adaptive Filter Architecture," *IEEE Trans. Signal Processing*, vol. 43, pp. 1819–1830, Aug. 1995.

[37] S. Sakauchi, Y. Haneda, S. Makino, M. Tanaka, and Y. Kaneda, "Subjective assessment of the desired echo return loss for subband acoustic echo cancellers," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E83-A, no. 12, pp. 2633 – 2639, 2000.

[38] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: analysis, experiments, and application to acoustic echo cancellation," *IEEE Trans. Signal Processing*, vol. 40, pp. 1862 – 1875, Aug. 1992.

[39] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing.* Prentice Hall, 1983.

[40] M. Harteneck, S. Weiss, and R. W. Stewart, "Design of near perfect reconstruction oversampled filter banks for subband adaptive filters," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 46, no. 8, pp. 1081 – 1085, 1999.

[41] S. Weiss, A. Stenger, R. Stewart, and R. Rabenstein, "Steady-state performance limitations of subband adaptive filters," *IEEE Trans. Signal Processing*, vol. 49, pp. 1982 – 91, September 2001.

[42] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing Principles, Algorithms, and Applications.* Prentice Hall, third ed., 1996.

[43] W. Chin and B. Farhang-Boroujeny, "Subband adaptive filtering with real-valued subband signals for acoustic echo cancellation," *IEE Proc. Vision, Image and Signal Processing*, vol. 148, no. 4, pp. 283 – 8, 2001.

[44] M. Knappe and R. Goubran, "Steady-state performance limitations of full-band acoustic echo cancellers," *IEEE International Conference on Acoustics, Speech and Signal Processing. Proceedings.*, vol. 2, pp. 73 – 6, 1994.

[45] A. Birkett and R. Goubran, "Limitations of handsfree acoustic echo cancellers due to nonlinear loudspeaker distortion and enclosure vibration effects," *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 1995.

[46] P. Eneroth and T. Gänsler, "Analysis of subband impulse responses in subband echo cancelers," tech. rep., Department of Applied Electronics, Signal Processing Group Lund University, 1999.

[47] D. Duttweiler, "A twelve-channel digital echo canceler," *IEEE Trans. Communications*, vol. CM-26, no. 5, pp. 647 – 53, 1978.

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

[49] A. Sugiyama, J. Berclaz, and M. Sato, "Noise-robust double-talk detection based on normalized cross correlation and a noise offset," *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings.*, vol. 3, pp. 153 – 6, 2005.

[50] J. Cho, D. Morgan, and J. Benesty, "An objective technique for evaluating doubletalk detectors in acoustic echo cancelers," *IEEE Trans. Signal Processing*, vol. 7, no. 6, pp. 718 – 24, 1999.

[51] T. Jia, Y. Jia, J. Li, and Y. Hu, "Subband doubletalk detector for acoustic echo cancellation systems," *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings.*, vol. 5, pp. 604 – 7, 2003.

[52] J. Gordy and R. Goubran, "A subband doubletalk detector for echo cancellation in hands-free environments," *IEEE Vehicular Technology Conference. Proceedings*, vol. 3, pp. 1397– 1401, Sept. 2005.

[53] A. S. Spanias, "Speech Coding: A Tutorial Review," *Proc. IEEE*, vol. 82, no. 10, pp. 1541 – 1582, 1994.

[54] C. Laflamme, J.-P. Adoul, H. Y. Su, and S. Morissette, "On reducing computational complexity of codebook search in CELP coder through the use of algebraic codes," *IEEE International Conference on Acoustics, Speech and Signal Processing. Proceedings.*, vol. 1, pp. 177 – 180, Apr. 1990.

[55] X. Lu and B. Champagne, "A centralized acoustic echo canceller exploiting masking properties of the human ear," *IEEE International Conference on Acoustics, Speech and Signal Processing. Proceedings.*, vol. 5, pp. 377 – 380, 2003.

[56] X. Lu and B. Champagne, "Pitch analysis-based acoustic echo cancellation over a non-linear channel," in *Proc. 11th European Signal Processing Conf. (EUSIPCO)*, vol. 1, pp. 159–162, Sep 2002.

[57] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus." Linguistic Data Consortium, Philadelphia `http://www.ldc.upenn.edu/Catalog/LDC93S1.html`, 1993.

[58] Third Generation Partnership Project (3GPP), "TS 26.204 speech codec speech processing functions; adaptive multi-rate - wideband (AMR-WB) speech codec; ANSI-C code," tech. rep., 3GPP, 2004. Available online at: `http://www.3gpp.org/ftp/specs/html-info/26204.htm`.

[59] M. Reuter and J. Zeidler, "Nonlinear effects in LMS adaptive equalizers," *IEEE Trans. Signal Processing*, vol. 47, no. 6, pp. 1570 – 9, 1999.

[60] A. Beex and J. Zeidler, "Non-linear effects in adaptive linear prediction," *Proceedings of the Fourth IASTED International Conference Signal and Image Processing*, pp. 21 – 6, 2002.

[61] K. Quirk, L. Milstein, and J. Zeidler, "A performance bound for the LMS estimator," *IEEE Trans. Information Theory*, vol. 46, no. 3, pp. 1150 – 8, 2000.

[62] H. Sheikhzadeh, R. L. Brennan, and K. R. L. Whyte, "Near-end distortion in over-sampled subband adaptive implementation of affine projection algorithm," in *Proc. EUSIPCO 2004*, pp. 413 – 416, Sep. 2004.

[63] A. A. Beex and J. R. Zeidler, "Steady-state dynamic weight behavior in (N)LMS adaptive filters," in *Least-Mean-Square Adaptive Filters* (S. Haykin and B. Widrow, eds.), John Wiley & Sons Inc., 2003.

[64] P. L. De Leon and D. M. Etter, "Experimental results with increased bandwidth analysis filters in oversampled, subband acoustic echo cancelers," *IEEE Signal Processing Letters*, vol. 2, pp. 1 – 3, Jan. 1995.

[65] O. Hoshuyama, R. A. Goubran, and A. Sugiyama, "A generalized proportionate variable step-size algorithm for fast changing acoustic environments," *IEEE International Conference on Acoustics, Speech and Signal Processing. Proceedings.*, vol. 4, 2004.

[66] E. J. Diethorn, "Perceptually optimum adaptive filter tap profiles for subband acoustic echo cancellers," *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 1995.

[67] V. Myllyla and G. Schmidt, "Pseudo-optimal regularization for affine projection algorithms," *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings.*, vol. 2, pp. 1917 – 20, 2002.

[68] E. Chau, H. Sheikhzadeh, and R. Brennan, "Complexity reduction and regularization of a fast affine projection algorithm for oversampled subband adaptive filters," *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings.*, vol. 5, pp. 109 – 12, 2004.

[69] F. Albu and C. Kotropoulos, "Modified gauss-seidel affine projection algorithm for acoustic echo cancellation," *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings.*, vol. 3, pp. 121 – 4, 2005.

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