THE UNIVERSITY
of LIVERPOOL

Echo Control Techniques in Public Switched Telephone Networks

Thesis submitted in accordance with the requirements of
the University of Liverpool for the degree of
Doctor in Philosophy

by
David J. Jones

Department of Electrical Engineering and Electronics
The University of Liverpool.

October 1998.

Acknowledgements

I would like to thank my supervisor, Dr. Ken Evans, for the time he has dedicated to helping me with this project. Throughout my time at Liverpool University, our conversations have been a great source of inspiration and relief from the pressures of a PhD.

I would also like to thank my industrial supervisor, Bob Reeves, for giving me the opportunity to work in a real industrial environment. Throughout this project, our discussions have served to enlighten me on the nature of real-world echo cancellation.

I would like to thank BT for supplying the facilities and financial support that has made this work possible. Thanks also to the University of Liverpool, Department of Electrical Engineering and Electronics for providing facilities during my studies and to EPSRC for financial assistance.

Thanks also go to Derek Molyneux, Scott Watson, Mat Pollard, Fabrice Plante, Jason Devaney, and other members of the other members of the DSP research group at the University of Liverpool, past and present.

Finally, I would like to thank Michelle for her support and for encouraging me to keep going during the last few months. I couldn't have finished without you.

ABSTRACT

This thesis describes some of the problems associated with the control of speech echoes in a telephone network. Although echo is present in all telephone calls made using a PSTN, it only becomes noticeable when the round-trip delay between the near and far-ends of the system exceeds 35-50ms. For round-trip delays below this limit, the presence of echo is likely to make the connection sound 'hollow' or reverberant. Normally in the UK, this limit is only exceeded on international calls or calls made using digital mobile telephones, and hence echo control is only applied to such connections.

Although various approaches may be used to remove unwanted echo, this thesis is exclusively concerned with the echo cancellation technique. This uses adaptive filtering to synthesise a replica echo, which is then subtracted from the actual echo. In practice however, it is unlikely that complete cancellation of the echo will occur, even if the optimum filter is used. One reason for this is that a linear echo canceller is unable to remove the distortion added to the echo by the companding process at the near-end of the network. Although the quantisation noise has different spectral properties from the original echo, its power varies in proportion to that of the echo and thus, it is likely to be perceived as echo if transmitted to the far-end talker.

In a network echo canceller, the uncancelled echo caused by quantisation noise is often suppressed using a centre-clipper. An alternative to centre-clipping is to use an adaptive filter that is capable of modelling the distortion. In principle, this might be achieved by companding the output of the adaptive filter but this is likely to lead to unwanted spikes in the canceller residual due to a quantisation level mismatch between the predicted and actual echo. For this technique to be successful, the adaptation process must be improved. In order to achieve this, an algorithm has been designed to estimate the echo before it was corrupted by the companding process, and this estimate is then used to form the error term in the adaptation process. Unfortunately, the effectiveness of this technique deteriorates as the near-end background noise level increases.

Several other techniques for suppressing the residual echo have been examined. These can all be effective at reducing the residual echo level but an unwanted side effect called 'noise modulation' is produced when near-end background sounds are present. Although less disturbing than uncancelled echo, noise modulation is likely to be audible by the far-end talker. A comfort noise system has been developed that enables any noise modulation to be almost completely masked. This system uses a GSM Voice Activity Detector to indicate times when background sounds are present in isolation and to obtain parameters that represent their spectral characteristics. The resulting comfort noise is not only spectrally shaped to approximate that of the background sounds but also has a similar time variation to the original sounds. It has been found that certain type of background sound may result in unnatural sounding comfort noise. A technique to detect such sounds has been proposed that uses parameters obtained from the VAD.

Contents

1. Introduction

1.1 Introduction

Speech is the most natural form of communication between humans and it is therefore not surprising that speech telephony is the most popular form of telecommunication. Indeed, with over 26 million telephone subscribers in the UK, speech telephony is BT's core business, accounting for over 90% of revenue [WEST96]. Today, global networks have evolved to a level where it is possible to communicate with half a billion telephone subscribers world-wide, from virtually any place at any time, simply by dialling a few numbers. Although other forms of telecommunication, for example electronic mail, have increased in popularity during recent years, they are in general more difficult to use, less natural, and do not permit the same degree of expression as speech communication. Thus, speech is likely to remain the most important form of telecommunication.

In an ideal communication system, the 'connection' between the two subscribers should be completely transparent. The transmitted speech should be free from impairments or defects, and in addition, the connection should be full-duplex because a real conversation is an inherently two-way process. Clearly, the transmission characteristics of the network should reflect this ideal. In practice however, speech transmitted over a telephone network will not sound identical to the original, and the full-duplex nature of the connection may cause a number of problems. Varying

degrees of degradation will always be present in a connection, whose characteristics depend upon the transmission method and any processing that may be employed within the network. In general, the subscribers' quality expectations determine which types of impairment or degradations are acceptable. For example, some impairments will be deemed acceptable either because they do not severely affect the usability of the connection or because subscribers have been conditioned to them. Conversely, the presence of other types of impairment may result in a connection being virtually unusable.

The impairments present in a telephone connection generally fall into two categories: those that are caused by one-way transmission properties and those that are due to the full-duplex nature of the connection. The one-way transmission characteristics mainly influence the perceived quality of the earpiece/loudspeaker output. For example, the remote subscriber's speech may suffer unwanted amplitude and/or phase distortion that varies with time and frequency. Also, in the absence of far-end speech, the perceived quality of a connection may be affected by thermal noise, unwanted noise from other sources or cross-talk from other channels. The two-way communication process may be disrupted by several types of impairment, although subscribers may not necessarily perceive them as introducing a reduction in the full-duplex capabilities of a connection.

An important one-way transmission impairment that may occur in both analogue and digital transmission systems is propagation delay. In a half-duplex communication link, the propagation delay may not pose a significant problem. However, in a full-duplex system, the subscribers may find that a connection with excessive transmission delay is unusable because it upsets the dynamics of the conversation. If both parties understand the propagation delay, then an agreed protocol could be adopted to make the connection useable, albeit in a less natural way. In practice, connections that introduce a significant amount of delay are likely to suffer from an additional type of degradation, namely echo. It has been found that when echo is audible, subscribers have great difficulty in using a connection even if they understand the echo generation mechanism.

All modern telecommunications networks are constructed using digital transmission systems that are based on the use of pulse code modulation (PCM) and time division multiplexing. This enables speech (and data) to be transmitted without many of the one-way degradations associated with long distance analogue transmission. As networks have been modernised to use such systems the quality of the connections has increased, and the expectations of subscribers have increased correspondingly. However, the problems of two-way degradations such as transmission delay and echo remain, and arguably, are now even more of a problem. The use of digital systems has increased the delay in telecommunications networks and thus the presence of echoes is now likely to be more frequent and more problematic. Moreover, the introduction of echo control devices may also add to the two-way imperfections present in a connection. For example, echo suppressors work by inserting a large loss into the transmission path when echo is present, and this limits the full-duplex capability of the connection.

1.2 Research Aims

The presence of echoes in long distance or international telephone calls is not a new phenomenon. Consequently, the application of 'echo control' to improve the quality of a connection is not a new subject area. Indeed, there is a vast array of published literature describing various methods of reducing the effects of echo. Today, the most important form of echo control employs the echo cancellation technique that uses adaptive filtering to generate an echo replica which is used to effect the cancellation.

This thesis presents a complete review of the echo generation mechanism within a Public Switched Telephone Network (PSTN) and of the most important echo control methods. Echo cancellers must operate in an environment where the echo is subject to background sounds from the near-end of the network and quantisation noise introduced by the codecs in the echo path. It is known that linear echo cancellers are unable to remove this quantisation noise, which under certain circumstances may be interpreted as echo by the far-end subscriber. Although some network echo cancellers are known to use 'residual echo control' there appears to be very little published material in this area. The aim of the research has therefore been to investigate various

ways in which the significance of residual echo can be diminished. The techniques that have been identified have several associated problems. In particular, the unwanted phenomenon of 'noise modulation', encountered with all of the techniques described in this thesis, is addressed.

1.3 Thesis Structure

Chapter 2 introduces a simplified structure for a PSTN and the differences between the local access network and the national/international network are discussed. The echo generation mechanism is introduced within the context of this simplified PSTN, and the consequences of both propagation delay and echo, when simultaneously present are examined. Finally, several echo control techniques including echo suppression and echo cancellation are introduced and their relative merits compared.

In chapter 3, the basic principles of adaptive filtering are discussed in relation to the echo cancellation problem. Although this thesis is not specifically concerned with adaptive filtering, any discussion of echo cancellation would be incomplete without consideration of the adaptive filtering process. Minimisation of the mean squared error leads to a set of optimum coefficients that form the optimum Wiener Filter. The well-known LMS algorithm, which seeks to emulate the performance of a Wiener filter, is described and its performance is examined in the context of a network echo canceller. It is demonstrated that an echo canceller based solely on linear filtering is unable to remove the quantisation noise that is generated at the near-end of the network. It cannot therefore be, by itself, a complete solution to the echo control problem.

In principle, one way of removing the quantisation noise is to introduce the same distortion into the predicted echo waveform as digitisation introduces to the true echo by using knowledge of the companding process. Unfortunately, the residual echo is now likely to contain unwanted spikes that occur whenever there is a mismatch between the quantisation levels of the predicted and actual echo waveforms. Chapter 4 describes a modified NLMS algorithm that may be used to reduce the

misadjustment of the filter in the presence of quantisation distortion, without requiring a reduction in step size. Its is shown that when using this algorithm, companding of the filter output may be used more successfully to reduce the power of the residual echo. However, it is shown that this technique is only likely to be successful when there are no background sounds from the near-end of the network. Unfortunately, the presence of such sounds reduces the performance of the enhanced algorithm to that of the standard LMS algorithm.

Chapter 5 discusses several residual echo control techniques and examines the distortion of the near-end talker and any background sounds that may be present. It is found that although these techniques can be effective at reducing the level of the uncancelled echo, they all produce an unwanted side-effect known as noise-modulation.

Chapter 6 describes a comfort noise system that can avoid the noise modulation introduced by the techniques described in the previous chapter. The comfort noise system uses a GSM voice activity detector (VAD) to indicate times when only background sounds are present. During these times, the spectral characteristics of the background sounds are calculated and stored for use in comfort noise generation. The resulting comfort noise is not only spectrally shaped to match that of the background but also has a temporal variation that is similar to the actual background sounds. The comfort noise system has been tested using several different types of background sounds that are believed to be representative of those encountered in a real network and the results of these tests are also presented in chapter 6.

During the testing of the comfort noise system, it was found that the presence of certain types of background sound could result in unrealistic sounding comfort noise. These sounds, are usually present for short periods of time and have strong periodic components, one example would be the sound of a ringing telephone. It is very important that these sounds are detected and excluded from the comfort noise system. A technique is proposed in chapter 7 that uses the partial correlation

coefficients, that are calculated as a 'by-product' of linear prediction used in the VAD, to detect the presence of such sounds.

Finally, the conclusions and recommendations for future work are presented in chapter 8.

2. Network Echo Cancellation

2.1 Introduction

The generation of echoes in a full-duplex voice communications system such as the Public Switched Telephone Network (PSTN) has the potential to severely degrade the quality of a connection. Echo is present on all calls across a PSTN, but it only degrades the perceived quality when certain factors become significant. In these circumstances, some form of echo control must be applied, ideally, in such a way that subscribers are unaware that it is taking place.

This chapter introduces the basic principles of network echo generation and control, and begins by describing the main features of a PSTN that are relevant, including its structure and the speech transmission process. The echo generation mechanism will then be outlined and it will be seen that echo can arise both within the network itself and at any customer equipment connected to it. The degree to which the echo is disturbing depends upon the echo delay and the echo attenuation, and for this reason, the effects of these factors will be discussed.

Where echoes are perceptible, the application of some form of echo control is necessary to prevent the quality of the connection from being unsatisfactory. In the past, various techniques such as permanent loss insertion and echo suppression

(switched loss insertion) have been used for echo control, and some of the methods that have been used are described here. Finally, the basic principles of echo cancellation will be introduced. It will be shown that echo cancellation has significant advantages over echo suppression methods, particularly in relation to maintaining the full-duplex nature of a conversation whilst simultaneously attenuating any echo that may be present.

2.2 The structure of a PSTN

In a PSTN, each subscriber is connected to a local exchange. When a call is set up between two subscribers attached to the same exchange, i.e. a local call, the connection can be made internally within the exchange. However, when a call is made between two subscribers attached to different exchanges, i.e. a long distance call, the connection is more complicated because of the need to establish a link between the local subscribers' exchanges.

2.2.1 Switching in a PSTN

One way to construct a national or long-distance network would be to connect each exchange to all other exchanges, i.e. use a non-hierarchical network. However, this is not a practical solution for a network with a large number of exchanges. For example, in the UK there are approximately 6300 exchanges, which if connected using a non-hierarchical topology would require almost 20 million separate links [BATE91].

An alternative solution is to use a hierarchical network. This may be achieved by fully interconnecting a small number of the exchanges or switches, say 100 [BATE91], with each of the others being connected to just one of these hundred. Thus, subscribers may be connected over long distances by switching the call through a series of exchanges in a hierarchical network. Note that a real network is likely to be more complicated than suggested by the previous statements because there are more than two levels of hierarchy. The routing of international calls follows

a similar pattern – the call is directed to one of several different International Switching Centres (ISC's) which then switch the link over a submarine cable or satellite to a foreign country.

The switching principle for both long-distance and local calls is illustrated in Figure 2.1, which shows several connections in a hypothetical network. A long distance national call from Ipswich to Edinburgh might be switched through exchanges in Manchester, Birmingham and London before arriving at Ipswich. A local call between two subscribers X and Y connected to an exchange in Cheshire, say, is switched internally within that exchange.

Call routing through a hierarchy of exchanges

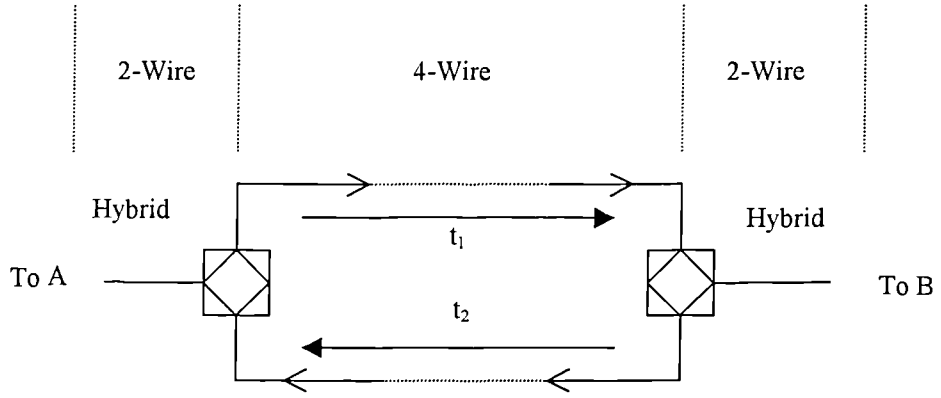
Figure 2.1

2.2.2 Two and Four-Wire Transmission

Now that the method of connecting the subscribers has been introduced, the type of links that are used to connect them can be described.

Each subscriber is connected to a local exchange via a 'two-wire' connection, which consists of a pair of copper (or aluminium) conductors that are twisted together. This is a bi-directional connection that carries both the local talker's speech from the microphone and the remote talker's speech to the earpiece. For local or short-distance calls a bi-directional transmission path is satisfactory as the speech is not noticeably attenuated or distorted. Over long distances, the attenuation will be much larger and hence in a digital network the signals will need to be regenerated. This is best achieved by using a separate transmission path for each direction and hence the use of a so-called 'four-wire' transmission system.

Thus, a four-wire transmission system is used to transmit the speech signals through the network that connects the local exchanges, as shown in Figure 2.2. Although the local two-wire circuit physically consists of metallic conductors, the national (and international) four-wire circuits no longer use pairs of wires for each path. Fibre-optic cables, co-axial cables or microwave radio links, which have many channels multiplexed onto them are more commonly used. However, the term 'four-wire' is still in common use and should be taken to mean the use of a separate transmission path for each direction.



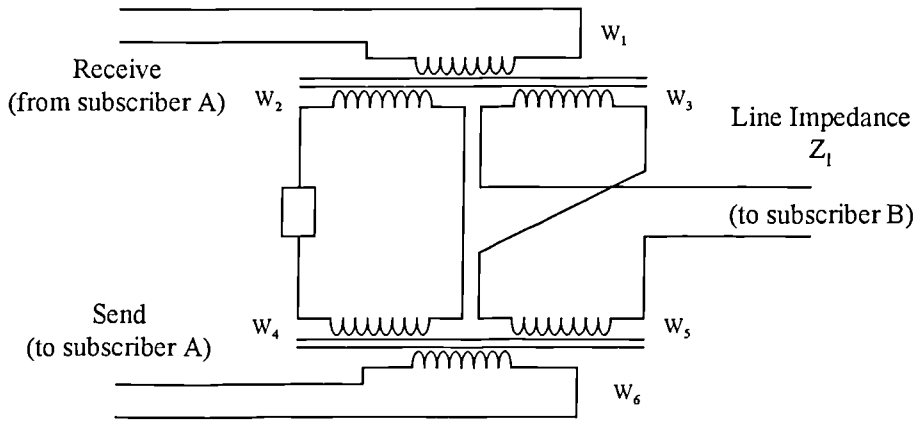
The two-wire/four-wire transmission system

Figure 2.2

For long-distance and international calls, the switching is performed in the four-wire circuit. The switching of local calls is also performed using a four-wire configuration in modern digital local exchanges, but in older analogue systems, the switching was performed in the two-wire section of the system.

2.2.3 Two to Four-Wire Conversion

The local two-wire connection is converted to a four-wire system using a device known as a 'hybrid', as shown in Figure 2.2. A hybrid is a four port device, in which two of the ports are attached to the four-wire circuit, one of the ports to the two-wire local circuit and the final port is connected to a balance network. There are many different possible hybrid circuits and one such circuit constructed using transformers [HILL73] is shown in Figure 2.3.



A typical hybrid transformer circuit

Figure 2.3

The intended function of this circuit is to route the Receive signal to the subscriber but not to the Send port since this would generate an echo. At the same time, the signal from the subscriber is to be routed to the Send port. However in reality, even in the ideal case only 50% of the received power is transferred to the subscriber and 50% is absorbed in the balance impedance. Similarly, only 50% of the power from the subscriber is transferred to the Send port and 50% is dissipated in the output impedance of the Receive port.

In a practical, i.e. non-ideal telephone network, the impedance of the two-wire circuit varies for different customers attached to the same local exchange. For example, one subscriber may be located further away from the exchange containing the hybrid than a different subscriber, and thus the impedance of the two-wire circuit is likely to be different. In the UK there are approximately 25 million customers using BT's network alone [BATE91], and this makes 'tuning' the balance circuit for every

customer impractical. Thus, the selection of the balance network impedance is a compromise for all the different two wire connections. An impedance mismatch is therefore highly probable and leakage across the hybrid will occur and this will result in the generation of an echo.

Note that because the impedance of the two-wire circuit cannot be represented as a single resistor in series with a single capacitor, which is the typical form of the balance impedance, the impedance mismatch is frequency dependent. The signal appearing at the Send port is therefore a filtered version of the signal present at the Receive port. The magnitude of the power attenuation is known as the Trans Hybrid Loss (THL) or more commonly the Echo Return Loss (ERL). The exact value depends upon the network and the turns ratio of the transformers. However, it is usually assumed to have a minimum value of 6dB [HILL73]. In modern networks, the average ERL is found to be approximately 10dB [ITUT97].

2.3 Digital Transmission of Speech

Whilst the two-wire connections to the local exchanges are analogue, all modern national and international networks employ digital transmission and this provides high quality long distance transmission between local exchanges. The digital information is transmitted in PCM format in the national and international networks. Note that some connections also use schemes such as DCME (Digital Circuit Multiplication Equipment) [ITUT91] and ADPCM (Adaptive Differential PCM) [ITUT90] for the international 'jump', because it allows a higher traffic capacity over the same link.

The analogue-to-digital and digital-to-analogue conversion processes take place at the local exchange using a logarithmic quantisation scheme, which is used to achieve an approximately constant signal to quantisation noise ratio over a wide range of input signal powers. The quantisation process is of particular interest here because any echo control system must operate using echo that has been corrupted by quantisation noise. As will be seen, the presence of quantisation noise is more

significant for connections that employ echo cancellers than those that use other kinds of echo control.

Before discussing the characteristics of logarithmic quantisation, uniform quantisation will be examined briefly.

2.3.1 Uniform Quantisation

In uniform quantisation, a constant spacing, Δ, separates the quantisation levels that are placed between the minimum and maximum allowed amplitudes. It is usually reasonable to assume that the quantisation error is equally likely to take any value between $\pm\Delta/2$, even if the amplitude distribution of the input waveform is not uniform. This assumption can be justified by considering how the amplitude distribution of the quantisation error changes with level spacing Δ.

If the distribution of the input waveform is non-uniform and the level spacing is large, the distribution of the quantisation error across each quantisation level is also non-uniform. However, if the level spacing is very small the distribution across each quantisation level is approximately uniform. The smallest number of levels used in practice is 256 and this gives a level spacing sufficiently small for the approximation to be valid. In addition to being valid in practice, this assumption simplifies the analysis of uniform quantisation.

The use of any quantisation process will result in a distorted version of the original signal appearing at the quantiser output. For a uniform quantiser with a sufficient number of levels, the distortion component of the output waveform will have a white spectrum and, provided that the input signal is not too small, will have constant power for all input signals. For this reason, the distortion is also known as quantisation noise. It may be shown that the signal to quantisation noise ratio (SQNR) for a linear quantiser is given by:

$$\mathrm{SQNR}_{\mathrm{linear}} = 10\log\left(\frac{\sigma_x}{x_m}\right)^2 + 4.77 + 6n \ \ (\mathrm{dB}) \qquad (2.1)$$

where:

σ_x^2 = input power (input amplitude must not exceed $\pm x_m$)

x_m = maximum amplitude allowed by quantiser

n = number of bits in the quantiser

This equation states that the SQNR is proportional to the input power, and improves by 6dB for every extra bit that is used. The maximum SQNR for a linear quantiser with 256 levels, that is achieved when the input signal is uniformly distributed with a power of ${x_m}^2/3$, is approximately 50dB. However, measurements have shown that speech has a non-uniform amplitude distribution, and that its power level varies over a wide range [RABI78].

It has been found that the amplitude distribution of speech may be approximated by a Laplace distribution, or more accurately by a Gamma distribution [RABI78]. For this discussion, only the Laplace distribution will be considered because it is simpler to manipulate and in any case, calculations based on the two distributions yield almost identical results [RABI78]. The Laplace distribution is given by:

$$p_x(x) = \frac{1}{\sigma_x\sqrt{2}} e^{-\frac{\sqrt{2}|x|}{\sigma_x}} \qquad (2.2)$$

where σ_x is the standard deviation of the distribution.

Now for the same 256 level quantiser, the maximum SQNR will be approximately 40dB (assuming $\pm 4\sigma_x = x_m$) if the input is distributed according to equation (2.2). Although the speech may occupy the full range of the quantiser with very infrequent

clipping, it will have a smaller power than a uniformly distributed input with the same maximum range. Thus, because the quantisation noise power is constant the maximum SQNR will be smaller than previously. In addition, the SQNR will be low for significant periods of time because small input amplitudes are more likely than large amplitudes.

The large unwanted variation of SQNR with input power that is experienced with uniform quantisation may be avoided by using a logarithmic quantisation scheme. In this scheme, an approximately constant SQNR is achieved by increasing the number of levels available for small amplitudes at the expense of a reduction in the number of levels available for large amplitudes. Thus, as the power of the input signal becomes smaller the power of the quantisation noise also decreases. The most commonly used logarithmic quantisation schemes are μ-Law which is used in North America and Japan, and A-Law which is used by the rest of the world. On connections between countries that use different schemes A-Law is used, and the country using μ-Law carries out the conversion between the two. If both countries use the same method then this method is used to encode the speech on international connections.

2.3.2 Logarithmic Quantisation

This section discusses the properties of logarithmic companding and in particular those of the A-Law, as this is used more frequently on international telephone connections. However, the properties of μ-Law are similar.

As stated previously, the purpose of using logarithmic quantisation is to improve the SQNR for speech signals and this is achieved by 'compressing' the signal. The compression process effectively amplifies the input waveform using a variable gain, which is inversely related to the input amplitude. The resulting compressed signal is then quantised by an 8-bit (in the case of μ/A-Law) linear quantiser and transmitted over the network. At the remote end of the system, the samples must be uncompressed using an expander whose transfer function is the inverse of the

compressor transfer function. This process of compressing and expanding the waveform is known as 'companding'.

For A-Law companding, the compressor characteristic is formally given by the following equation

$$y(t)=\begin{cases}\dfrac{A}{1+\ln A}\left(\dfrac{x(t)}{x_m}\right) & \left|\dfrac{x(t)}{x_m}\right|<\dfrac{1}{A} \\ \dfrac{\mathrm{sgn}[x(t)]}{1+\ln A}\left(1+\ln A\left|\dfrac{x(t)}{x_m}\right|\right) & \dfrac{1}{A}\le\left|\dfrac{x(t)}{x_m}\right|\le 1\end{cases} \qquad (2.3)$$

where

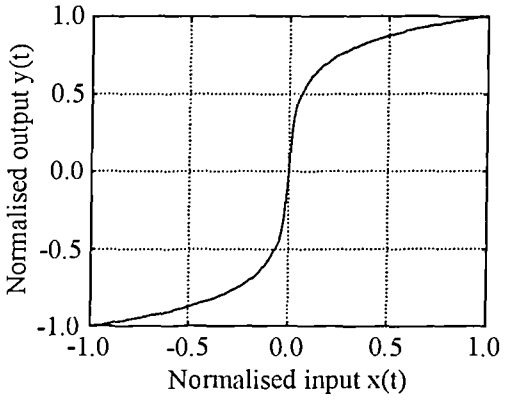
$x(t)$ = the instantaneous input amplitude

$y(t)$ = the instantaneous output amplitude

x_m = the maximum allowed input amplitude

A = compression factor = 87.6 for standardised A-Law

The transfer function given by (2.3) is plotted in Figure 2.4 for inputs normalised to ±1.0.



True A-Law compression characteristic

Figure 2.4

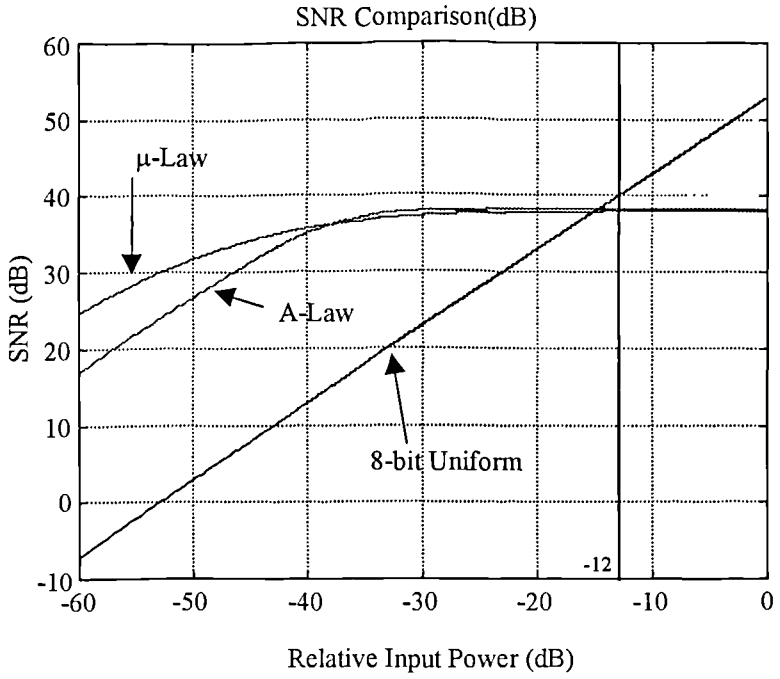
It can be seen that when the compressed output is uniformly quantised there are more levels associated with small input amplitudes than with large amplitudes. The

quantised version of the compressor output may be considered to be equivalent to the unquantised version plus quantisation noise, whose power is independent of the input power. At the remote end of the network the quantised signal is expanded and now, the quantisation noise power is no longer constant. It may be shown that [JONE94] for a normalised input whose amplitude is distributed according to equation (2.2), the SQNR for the A-Law is approximately given by:

$$SQNR_{a-law} = \frac{3L^2\sigma_x{}^2}{(1+\ln A)^2\left(\frac{1}{A^2} - \frac{e^{-\frac{\sqrt{2}}{\sigma_x A}}}{A^2} + \sigma_x^2\right)} \qquad (2.4)$$

where L is the number of quantisation levels ($2^8 = 256$) and all other symbols have the same meaning as before. It should be noted that this equation assumes clipping does not occur because it does not predict the further distortion that would be added to the quantisation noise.

Figure 2.5 shows how the SQNR varies with input power. Also shown for comparison, are the SQNR for μ-Law and the SQNR for an 8-bit uniform quantiser.



SQNR comparison for A-Law, μ-Law and uniform quantisation

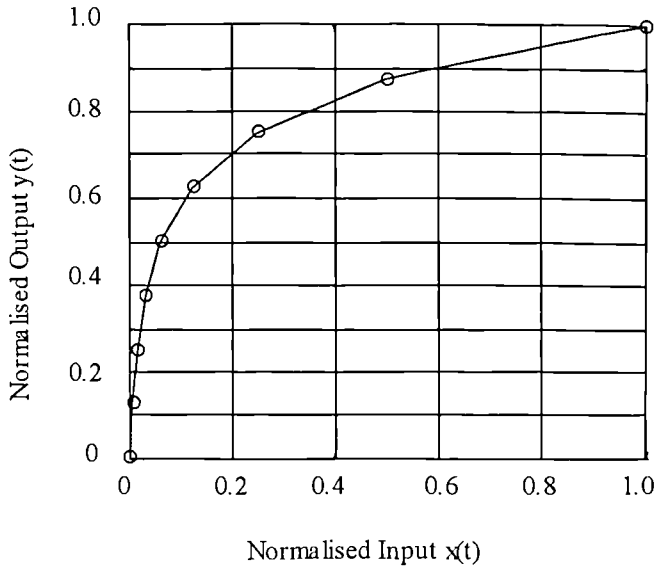
Figure 2.5

Figure 2.5 shows that for both μ and A-Laws, the change in SQNR is much smaller than the corresponding change in input power. For input powers in the range –40dB to –12dB the SQNR changes by less than 5dB, whereas the SQNR for the linear quantiser changes by 28dB. Note that the curves are not valid above –12dB because at this point the SQNR will begin to fall due to overload.

It can be seen that the SQNR performance using the μ and A-Law companding is broadly similar, but with two notable exceptions. Firstly, the SQNR for μ-Law is approximately 8dB larger than A-Law for small input powers where μ-Law has a smaller level spacing than A-Law. Secondly, for large input powers the SQNR for A-Law is slightly higher than for μ-Law. Compared to uniform quantisation, both μ and A-laws have a superior SNR performance until the input power rises above –15dB. Although Figure 2.5 suggests that μ-Law has a larger SQNR for small input powers, the way in which the networks have been designed result in a SQNR that is inferior to A-Law. PCM networks in most countries of the world have been designed to operate using a 32-channel multiplex (at the lowest level) of A-Law encoded

samples, where two of the channels are reserved for signalling. However, in North America and Japan, PCM networks were designed to operate using a 24-channel multiplex of μ-Law encoded samples with the LSB of each companded word being 'stolen' for signalling. When the LSB is being used for signalling, the SQNR is reduced so that it is smaller than that obtained using A-Law.

In practice, the compression functions of both μ and A-Law are piecewise linear approximations to the continuous functions, as shown in Figure 2.6 [BELL82], and this permits simple implementation by using look-up tables or on-the-fly calculation.



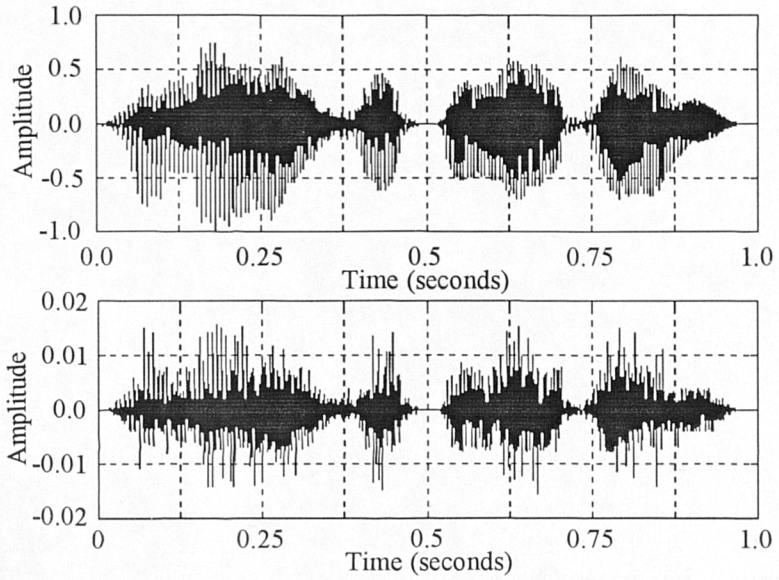
Piecewise-linear approximation to the true A-Law characteristic

Figure 2.6

The companding curve is split into 14 linear segments (16 for μ-Law) in such a way that it approximates the continuous function. The use of this approximated function slightly alters the SNR curves - they become 'scalloped' because the quantisation noise changes abruptly at the segment boundaries [BELL82].

Although there is normally little correlation between the quantisation noise and the original speech waveform, when listened to the quantisation noise has a 'speech-like'

quality. Figure 2.7 shows a speech waveform and the quantisation noise that is generated after companding using A-law companding.



a) Original speech, b) A-law quantiation noise

Figure 2.7

It can be seen that the envelope of the quantisation noise is similar to that of the original speech, although at a much reduced level. In the same way, the average power of the quantisation noise calculated over a short period (e.g. 20ms) varies in a similar way to that of the average speech power. If the quantisation noise is listened-to in isolation, it may even be possible to understand the original speech despite it having an entirely different spectral content. Thus, if an echo canceller removes the part of the echo that is linearly related to the far-end signal but permits the transmission of the associated quantisation noise then this may be heard as a distorted echo.

As will be discussed in chapter three, the quantisation error present in the echo affects the performance of any echo canceller, and the maximum echo attenuation obtainable by linear methods is ultimately limited by the quantisation process.

2.4 Echo Generation

Now that transmission of speech in the network has been discussed, the echo generation mechanism can be described. Suppose that the local subscriber A is talking to subscriber B at a remote location as in Figure 2.2, and that the impedance of the balance network of the remote hybrid is mismatched. As subscriber A talks, the speech signal propagates through the network to the remote hybrid, where some of the signal is leaked across the hybrid and returned to the talker. The leakage across the hybrid from the Receive port to the Send port, in Figure 2.3, causes the talker to hear an echo. The time delay between the talker speaking and hearing the echo is determined by the Round Trip Delay = $t_1 + t_2$, where t_1 and t_2 are the outward and inward, one-way propagation times.

One of the sources of delay in a telecommunications system arises when signals must propagate over long distances. The following table shows typical one-way transmission times for different media [ITUT93b].

Transmission Medium	One-way Transmission Time
Terrestrial coaxial cable (FDM and Digital transmission)	4μs/km
Optical Fibre (Digital transmission)	5μs/km
Submarine coaxial cable	6μs/km
Satellite (14000km orbit)	110ms
Geostationary Satellite (36000km orbit)	260ms

Note that the one-way transmission times given for the satellites are the propagation times between two earth stations, i.e. the sum of the 'up' and 'down' transmission times. Using these figures it can be seen that a call from the UK to the USA routed through a trans-Atlantic submarine cable will incur a minimum one-way delay of 18ms (3000km). For calls routed over a satellite, its altitude determines the

transmission time, which in general will always be larger than for an equivalent landline. In practice, the one-way transmission time over a satellite link is likely to by larger than 300ms [ITU93b]. This is because the 260ms time given here does not include any processing delay that may occur before the signal can be sent over the link.

There are other sources of delay in a communications system other than propagation delays, and these are usually attributed to processing of the signal. For example, delay is introduced in a local exchange when a PCM coder digitises the signal (0.3ms), or when it passes through a PCM/ADPCM/PCM transcoder (0.5ms). Such delays are usually very small in comparison to the propagation time over long landlines or satellite and so are usually ignored.

The transmission delays in each direction, t_1 and t_2, will only be equal if the call is routed over the same physical transmission medium for both directions. If for example, a call is routed via a satellite in one direction and over a landline in the opposite direction then t_1 and t_2 will not be equal. However, subscribers would only be aware of the round-trip delay and would not know that this inequality exists.

2.4.1 The Effects of Delay in the Absence of Echo

It is important to realise that when the round trip delay is large the presence of echo is not the only factor that affects the quality of the connection. Even in echo free conditions, the delay has a pronounced effect on the dynamics of a conversation although users may not attribute the resulting difficulties to the delay [RIES63].

For example consider what happens if user B attempts to interrupt user A during a conversation. Due to the transmission delay, A will keep talking for a period of time that is related to the one-way propagation delay between B and A. When user B's interruption reaches A, user A may stop talking to listen to B. However by this time B has also stopped talking because A has apparently ignored the interruption. As the round-trip delay becomes larger, it has been found that the probability of

interruptions becomes higher and that the conversation becomes more 'broken-up'. The problems caused by the introduction of delay are often misinterpreted as a problem with the other subscriber, for example, slow responses, excessive interruptions or complete failures to respond, and therefore the connection is not rejected as being unsatisfactory. Only when the round-trip delay becomes very large, for example in the region of 2400ms, do the users attribute the difficulties to delay [REIS63]. However, once users have recognised that the difficulties are caused by delay, it has been found that connections that were previously deemed satisfactory are now rejected.

In order to minimise the difficulties caused by delay, the ITU-T specifies that for general network purposes the one-way transmission delay should never exceed 400ms [ITUT93b]. However, this limit may be exceeded under some circumstances. For example, calls which are routed over satellites in both directions and international calls between digital cellular systems that are connected by long terrestrial facilities. Another extreme example, is in the Skyphone system [BOYD88], [LEWI94], where calls from aircraft to earth-station are routed over an Inmarsat satellite and may, additionally, traverse an Intelsat communications satellite.

2.4.2 The Effects of Echo and Delay

There are several types of echo that may occur in a PSTN. Talker echo is so called because the talker hears an echo of his or her own voice. Another type of echo, called listener echo, arises when the talker echo is reflected from the talker hybrid to the listener's end of the network. The listener then hears reverberation of the talker's speech. In general, control of talker echo also removes listener echo and hence listener echo is usually ignored.

If the round trip delay is small (<30ms), the talker will not be aware of a distinct echo, but its presence may bring a reverberant quality to the connection. Perhaps surprisingly, it has been found that some degree of echo or reverberation is preferable to the total absence of echo. Indeed to prevent a connection from sounding 'dead', a

form of echo known as sidetone is provided inside the customer's telephone equipment. An attenuated version of the talker's voice is allowed to feed back to the earpiece on the handset, and is automatically used by the speaker to adjust the loudness of his/her speech.

As the round-trip delay becomes larger, a distinct echo will become audible for a given ERL. An ITU-T workgroup has concluded that echo, although noticeable, presents no difficulty to users when the one-way delay is less than 25ms [G.131 Control of Talker Echo]. Generally, this means that network echo control is only applied to international calls, where large transmission delays are expected. There are of course exceptions. In large countries like the USA for example, echo control is necessary on long distance national calls where the round trip delay exceeds the limits as defined in the standards [ITUT93b].

Another type of echo that may occur is called acoustic echo. Rather than being generated at a hybrid, acoustic echo is caused by coupling between the loudspeaker and microphone of a communications device. Significantly, acoustic echo may arise even if there is no hybrid present, for example in a hands-free digital mobile telephone that is connected directly to the four-wire circuit. In other words, instead of being generated in the network the echo originates from the user equipment. In this situation, it is likely that multiple echoes or reverberations will occur because the far-end talker signal will be reflected by different surfaces in the environment.

2.5 Echo Control Techniques

When the round-trip delay is larger than approximately 30-50ms, some form of echo control becomes essential in order to prevent the echoes from interfering with the communication process. Currently, four approaches may be used in order to provide control of the hybrid echo. These are:

i) the removal of the hybrids by converting all two-wire circuits to four-wire,

ii) adding a net loss to the transmission path,

iii) echo suppression, and

iv) echo cancellation.

The following sections describe these four different methods, and discuss their relative merits.

2.5.1 Removal of the Hybrids

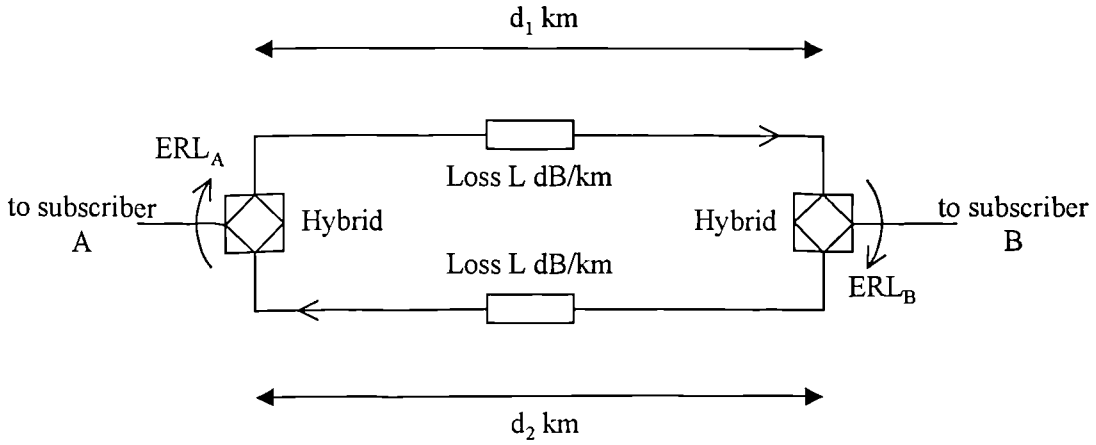
One solution to the echo problem is to remove the hybrids and convert all the two-wire circuits to four-wire connections. Although this would eliminate echo generated at the hybrids is would not necessarily guarantee echo free conditions. Acoustic echo could still occur as this is generated at the customer equipment rather than in the network.

Additionally the conversion of all two-wire connections to four-wire for the sole purpose of echo elimination is not economically feasible. This may be illustrated by considering that in the UK alone there are approximately 36 million metallic conductor pairs connecting subscribers to local exchanges [BATE91]. A considerable investment would be required to install another 36 million conductor pairs and upgrade local exchanges to cope with a four-wire local circuit. Further costs would also be incurred in maintenance of these extra connections. Moreover, there may not be enough physical duct space to cope with the extra wires, and this would again increase the cost of upgrading. In addition to upgrading the network, all the customer equipment would have to be replaced at considerable cost and inconvenience.

However at some time in the future hybrid echo may no longer be the main source of echo in a PSTN, as services may be introduced that require a direct four-wire connection. This is already happening in mobile telephony, where the handsets are interfaced to the operators' network using a four-wire connection [MOFF87].

2.5.2 Adding a Net Loss

One of the first approaches to echo control that was used in analogue networks was to insert a loss into the transmission path in each direction, as shown in Figure 2.8.



Echo control by loss insertion

Figure 2.8

As the length of the four-wire circuit increases, the propagation delay increases, and it has been found that more attenuation is then required to reduce the perceived loudness of the echo [ITUT96a]. Thus, the magnitude of the loss must vary in relation to the length of the transmission path. If L is the loss per kilometre required to give satisfactory echo performance, and d is the length of the path between the hybrids, the total loss in each direction is given by L'=L×d.

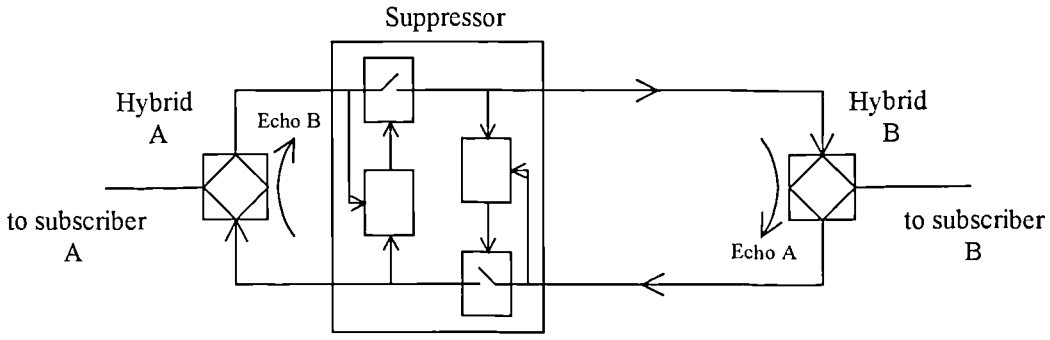
In Figure 2.8, the echo of subscriber A's speech is attenuated by $ERL_B + L(d_1 + d_2)$ dB, whereas the speech transmitted from B to A is only attenuated by Ld_2 dB. The obvious disadvantage of this technique is that as the round-trip delay increases, the need for attenuation becomes greater, in other words, L must be increased. It has been found that for circuits longer than approximately 1500km [MMSO80], the speech from B is excessively attenuated when the value of L is sufficiently large to give adequate attenuation of the echo.

2.5.3 Echo Suppression

Echo suppression was, for many years, the main technique for controlling echoes in circuits longer than approximately 1500km. Echo suppressors attempt to artificially increase the ERL by inserting a large loss into the echo return path whenever echo is present. The extra attenuation given to the echo signal is known as the Echo Return Loss Enhancement (ERLE). There are two types of echo suppressor configuration that have been in common use – the full configuration and the split configuration [MMSO80], and these are described in the following sections.

2.5.3.1 Full and Split Echo Suppressors

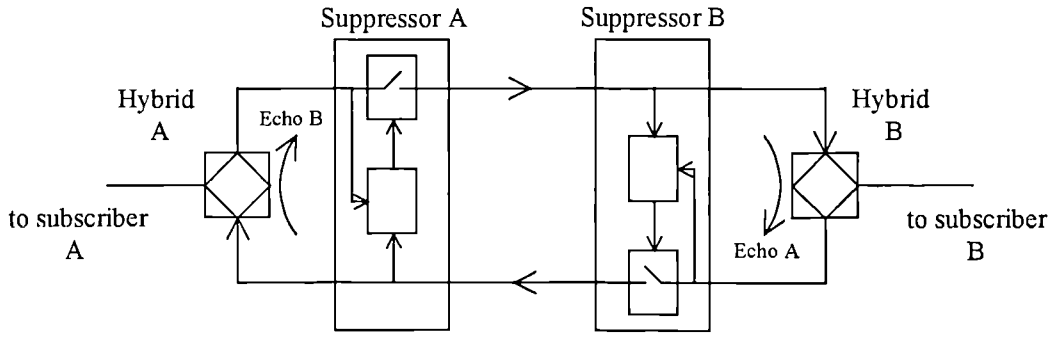
In the 'full' echo suppressor configuration, shown in Figure 2.9, the suppressor is located at one end of the network and performs echo control for customers at both ends.



Full Echo Suppressor Configuration

Figure 2.9

For circuits with a round-trip delay of less than 70ms, satisfactory performance can be obtained using this configuration, but for longer connections, better call quality is obtained using the split configuration [MMSO80]. In the split configuration, shown in Figure 2.10, two echo suppressors are used which are located at opposite ends of the four-wire network.

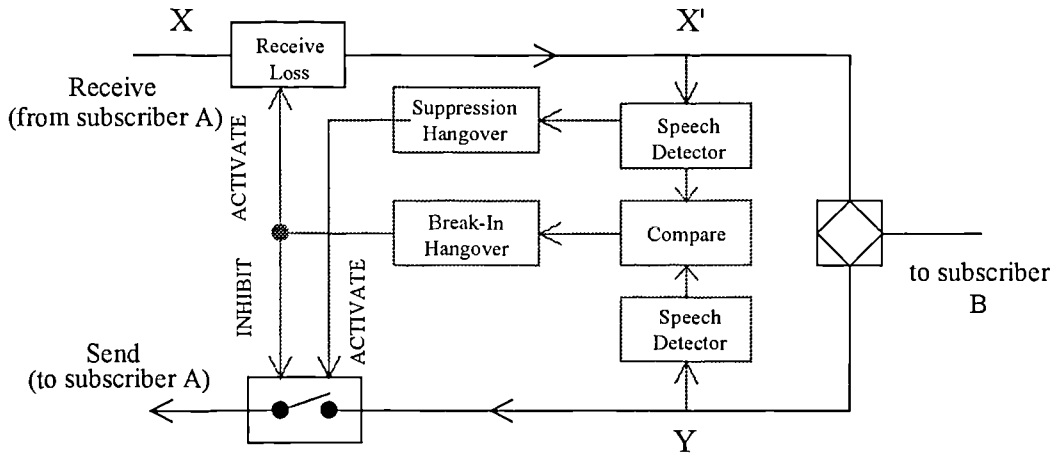


Split Echo Suppressor Configuration

Figure 2.10

The suppressor located near hybrid B attempts to remove the echo of talker A's voice that is generated at hybrid B. Similarly, the suppressor located near hybrid A attempts to control the echo of B's voice generated at hybrid A. Thus each echo suppressor provides echo control for the user at the other end of the system. In relation to a particular suppressor, the near-end of the system is defined as the end that is closest in terms of time delay. Similarly, the far-end is the end that is furthest away from the suppressor. Thus, a split echo suppressor located at the near-end of the network attempts to control the echo of the far-end talkers voice, which is generated at the near-end of the network.

Figure 2.11 shows a schematic diagram of a simple split echo suppressor [MMSON80].



A Simple Split Echo Suppressor

Figure 2.11

When speech is detected from the far-end, a large loss (typically 50dB) is inserted into the send path to attenuate any resulting echo. The suppressor may be located at a considerable distance from the hybrid, and hence the presence of speech at X' will not immediately produce a response at Y. If the loss was immediately switched in, there is a danger that low-level near-end speech could be attenuated before the echo is generated. As this is undesirable, an activation hangover (not shown) is usually employed to stop the loss being instantly switched in. Similarly, if the loss was immediately switched out when the far-end speech at X stops echo could be returned to the far-end. A suppression hangover is also added so that the loss is not removed for a period of time after the offset of far-end speech.

To allow interruptions from the near-end to pass unattenuated, the inserted loss must not be present when the signal from the near-end is mostly due to near-end speaker, rather than echo. When the suppressor is operating, with the loss inserted in the send path, it is important that it is removed as quickly as possible when the near-end talker starts. The condition of operation when both far-end and near-end speech are present is known as 'double-talk'. In general, double-talk is detected when the signal level at Y exceeds the signal level at X', and the suppressor enters the so-called 'break-in' state. Upon the detection of 'double-talk' two things happen:

i) the loss is removed from the send path,

ii) a loss, typically 6 to 12dB, is inserted into the receive path to help suppress echoes during break-in.

The suppressor must remain in the break-in state as long as the near-end talker is active. A hangover of typically 200ms is used to ensure that fluctuations in signal level at X and Y do not cause premature release from break-in. It may be observed that the receive path loss that is inserted during break-in will also help to maintain the break-in state.

Unsurprisingly, it is reported to be difficult to design an echo suppressor that functions correctly under all possible circumstances. These difficulties lead to incorrect suppressor operation with the following undesirable effects:

- Echo transmission due to slow detection of far-end speech, i.e. the send-path loss is not inserted quickly enough.
- Echo transmission due to false break-in. This is when the send-path loss is switched out because the near-end speech has been incorrectly detected.
- Clipping at the start of near-end speech caused by slow entry into the break-in state.
- Clipping at the end of near-end speech segments caused by premature release from the break-in state.

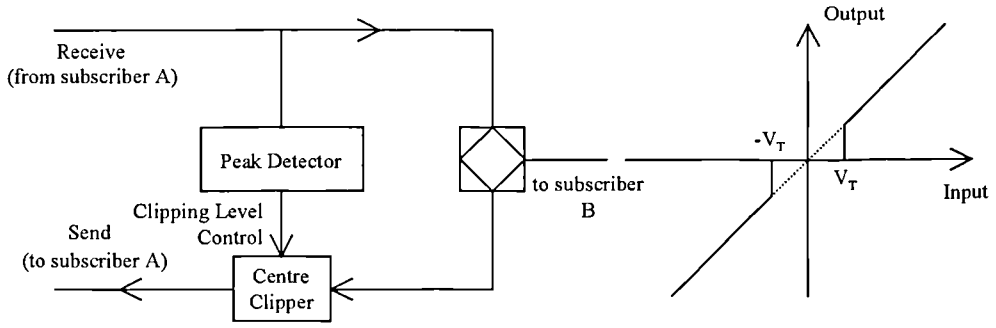
In order to minimise the above mutilation effects, the ITU recommends that echo control devices should be inserted at the end of the four-wire circuit, i.e. close to the hybrid. Such a location would lead to smaller activation and hangover times, and thus should result in less mutilation of the near-end speech. However, as this placement is not usually possible, suppressors are generally located in the international switching centres, which may be at a considerable distance from the hybrid.

On connections with echo suppressors, it has been found that customers are not always aware of the echo as it is adequately suppressed. However, the clipping and mutilation of the near-end speech caused by incorrect switching of the suppressor has been found to be more obvious, particularly on connections where the round-trip delay is greater than 100ms [MMSON80].

2.5.3.2 Centre-Clipping Echo Suppressor

For round-trip delays less than 100ms it has been found that the operation of a split echo suppressor is almost undetectable, but for longer delays the effects of echo and suppressor action as described above become unacceptable. A different kind of echo suppressor called a centre-clipping echo suppressor was proposed [MITC71] to overcome these difficulties.

Whereas echo suppressors of the type previously described insert a large loss in the send-path when echo is present, a centre-clipping echo suppressor only affects portions of the signal, which instantaneously fall below a threshold. In order to remove any echo the threshold should be set to the peak echo amplitude. The transfer function of a typical centre clipper is shown in Figure 2.12, where V_T is the clipping threshold.

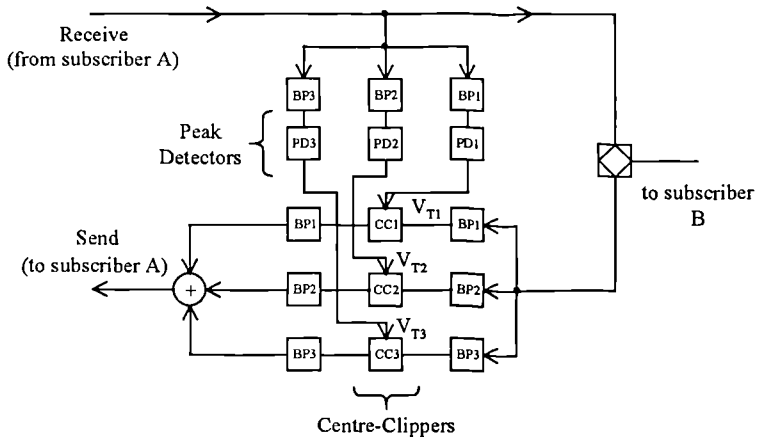


a) Centre-clipping echo suppressor, b) Centre-clipper transfer function

Figure 2.12

When only far-end speech is present, the echo is completely removed if V_T is set correctly. If the near-end speech is larger than V_T then it will be heard at the far-end, although it will be distorted because portions of the signal will lie below the threshold. The advantage of using this method is that double-talk detection of the type needed in a 'standard' echo suppressor is not required.

A further enhancement of the centre-clipper, is the multi-band centre-clipper [MITC71]. The block diagram for a three-band centre-clipper is shown in Figure 2.13.



A multi-band centre-clipper

Figure 2.13

The echo is passed through several different band-pass filters in parallel and each resulting time domain signal is processed using a centre-clipper. After processing, the signals are filtered using identical band-pass filters, to remove some of the distortion products, before being recombined. The thresholds V_{T1}, V_{T2} and V_{T3} are adjusted so that the echo in each band is removed by the clipping process. As with the full-band clipper, no explicit double-talk detection is required. The advantage of this design is that, compared to the single band centre-clipper, the near-end speech is not distorted to the same extent.

2.5.3.3 Background Noise

The full, split and centre-clipping types of echo suppressor have another disadvantage other than that echo may be uncontrolled or the near-end speech mutilated due to inaccurate timing or clipping. When the suppressor operates, all signals including any near-end background sounds or noise are attenuated. Assuming the switching decisions are made perfectly, the far-end subscriber will hear the suppressor action as a variation of the power of the background. This effect is known as 'noise modulation' or 'noise pumping'.

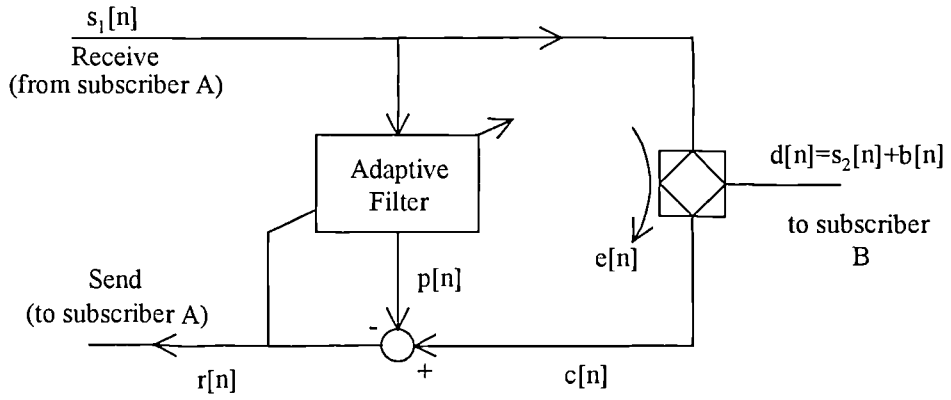
The echo suppression techniques described previously were proposed when analogue networks were widespread. In an analogue PSTN, the speech signals are degraded by noise that is inherent in analogue systems and this helps to mask the imperfections of

an echo suppressor. Today, low noise digital networks are the norm. This means that, in addition to the usual suppressor action, the presence of background noise from the near-end can also influence the perceived quality of a connection. The noise modulation phenomenon is, therefore, now more troublesome.

Although centre-clippers have never been fully developed for echo suppression, they are used in single-band form to control the residual echo that can occur in echo cancellers.

2.5.4 Echo Cancellation

Echo cancellers operate using a different principle from echo suppressors, which enables the mutilation of speech and near-end background sounds to be avoided. Split echo cancellers are located at the same positions in the network as split echo suppressors and, as before, control the echo generated at the near-end, for the benefit of the customer at the far-end of the system. Instead of switching the send path on and off by inserting a large attenuation, echo cancellers attempt to generate a replica of the echo which is subtracted from the near-end signal. If the replica could be made identical to the actual echo then complete cancellation would result. At the same time near-end speech and background sounds would pass undistorted. A block diagram of a simplified cancellation system is shown in Figure 2.14:



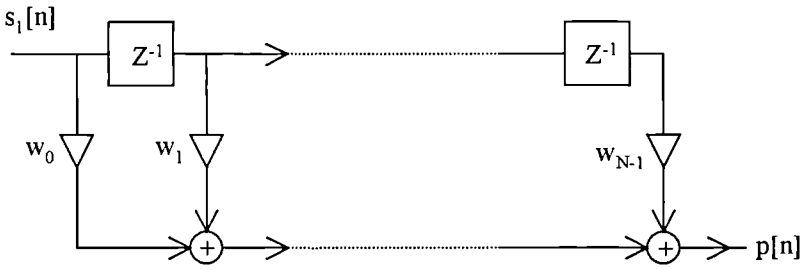
A simplified echo canceller

Figure 2.14

where:

$s_1[n]$	Far-end speech (and background noise)
$s_2[n]$	Near-end speech
b[n]	Near-end background noise
e[n]	Echo
c[n]	Echo + Near-end speech + Near-end Background noise
	$= e[n] + s_2[n] + b[n]$
p[n]	Predicted Echo
r[n]	$= c[n] - p[n]$

Synthesis of the echo replica is often performed using a FIR filter, although other structures may be used. The block diagram of a suitable FIR filter is shown below.



An N^{th} Order Finite Impulse Response (FIR) Filter

Figure 2.15

The output of this filter, which is the predicted echo, is given by:

$$p[n] = \sum_{i=0}^{N-1} w[i] s_1[n-i] \qquad (2.5)$$

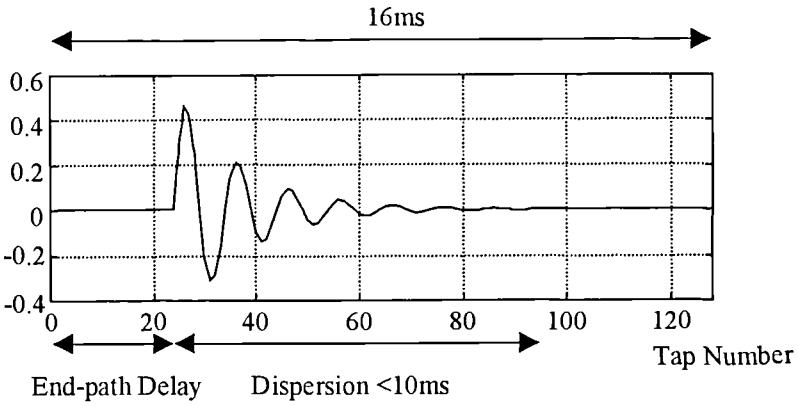
where

$w[i]$	are coefficients that represent the impulse response of the filter,
$s_1[n-i]$	the delayed filter input samples,

$p[n]$ the filter output, and

N the filter order.

For successful echo cancellation, the coefficients must be set so that they represent the impulse response of the near-end echo path. Typically, the impulse response consists of an end-path delay that corresponds to the propagation delay around the near-end path, and a dispersed 'ringing' response that is caused by amplitude and group delay distortion. The form of the impulse response is shown in Figure 2.16.



Generalised Impulse Response of the Echo Path

Figure 2.16

It is important that the end-path delay (which is sometimes called the flat or bulk delay) should not be confused with the round-trip delay. While the round-trip delay may be 100ms or more, as described in section 2.4, the end-path delay will be much shorter. In most European countries total duration of the response to be modelled is less than 32ms, whilst the dispersion is typically 5ms [LEWI92]. Thus, given that the sample rate is 8000 samples/s, an FIR filter with 256 coefficients is needed in order to achieve satisfactory cancellation. In this thesis, the length of the adaptive filter is often referred to as the 'span' of the adaptive filter, i.e. for successful cancellation the impulse response must fit completely within the span of the filter.

One method of obtaining the echo path impulse response would be to measure it directly, i.e. by applying an impulse to the near-end circuit and measuring the

returned waveform. In practice however, the impulse response will not be constant - it varies slowly with time due to the effects of temperature, for example. The impulse response may also change suddenly and substantially, for example when the echo path changes if a call is switched to a different extension. Thus, repeated measurements would need to be taken if the ERLE is to be maintained at the same level. However, repeated measurements using impulses is not acceptable, as they would be heard as 'clicks' or 'pops' at the near-end. In addition, the measurements would be corrupted by the near-end talker and background sounds.

A more elegant solution is to use the speech or other signals that are present on the line instead of impulses to set the coefficients in an adaptive filter. In an adaptive filter, the coefficients are adjusted recursively by a correction term that is proportional to the error between the filter output $p[n]$ and the near-end signal $c[n]$. However, the use of speech signals to 'adapt' the filter coefficients is not without drawbacks. One of the main problems is caused by speech from the near-end corrupting the error measurement, and this can lead to a drastic reduction in echo attenuation if appropriate measures are not taken. Another problem is that an adaptive filter of the type described here is unable to completely cancel the echo due to the presence of quantisation noise.

2.6 Conclusions

Echo is a phenomenon that occurs in all telephone systems but only becomes a problem when the round-trip delay exceeds 35 to 50 milliseconds. In the UK, this limit is only approached on international calls, where echo control techniques such as cancellation or suppression are needed to maintain the quality of the connection. Both echo suppression and echo cancellation have been used for many year to provide echo control on such connections. Due to the superiority of the performance of echo cancellers, present practice is only to introduce cancellers into new systems, and to replace existing suppressors with cancellers.

The following chapters discuss adaptive filtering techniques that may be used in echo cancellation, including filter structures and adaption algorithms. It will be seen that the use of filtering to generate an echo replica and the use of adaption algorithms places limitations on echo canceller performance and introduces other problems. The remainder of this thesis then discusses these problems and suggests ways in which they may be overcome.

3.Adaptive Filters

3.1 Introduction

In the previous chapter, the echo cancellation technique was introduced and it was seen that adaptive filtering was used to generate a replica of the echo. In principle, complete cancellation of the echo could be achieved by subtracting the replica echo from the actual echo. This chapter begins by introducing the basic principles of adaptive filtering that are relevant in the echo canceller application. There are three main configurations of adaptive filter, which may be used in a variety of different applications [HAYK96], [HUGH92], [LEWI92], [MCCO76]. The adaptive filters used in echo cancellation belong to the 'system identification' class, and hence all discussion in this chapter centres on this type. The other two configurations are 'inverse system modelling' and 'linear prediction'. Inverse system modelling is often used to reduce the effects of intersymbol interference in digital receivers, whilst linear prediction, enables the characteristics of a signal to be modelled so that its future values may be predicted from past values.

Although both FIR and IIR filter structures may be used in adaptive filtering, the FIR structure is used in most applications, including echo cancellation. This chapter will therefore concentrate on FIR adaptive filters, with the analysis being based upon that presented in [HAYK96], [HUGH92], [MESS84] and [WIDR76]. Optimum Wiener filters are described and this leads directly to the gradient search algorithm and the

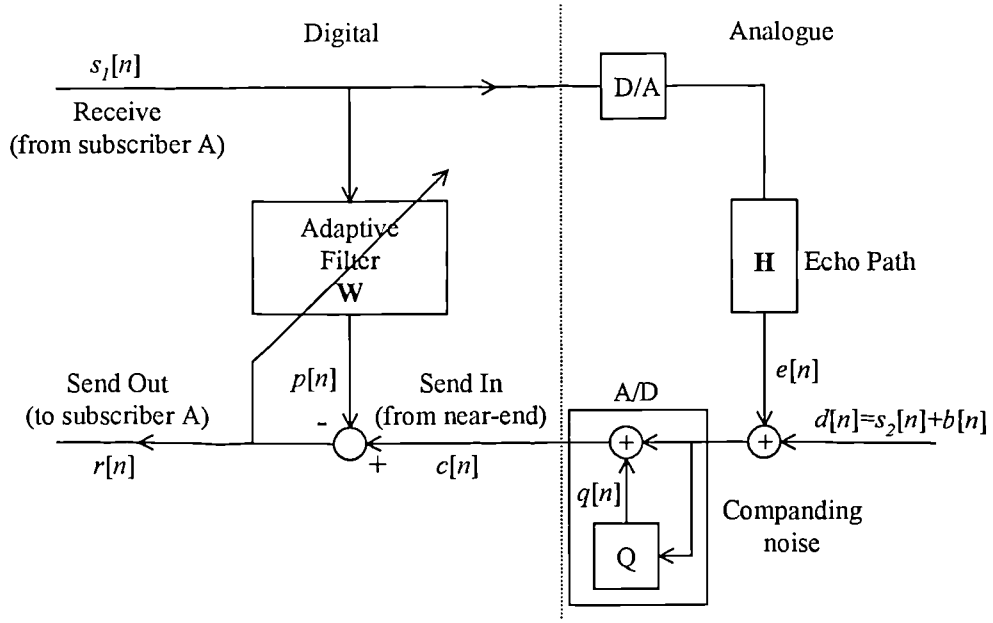
well known Least Mean Squares (LMS) algorithm. The performance of the LMS algorithm is discussed in some detail since this algorithm is commonly used in network echo cancellers, and because it is the method of adaptation used for the echo canceller described within this thesis. Other algorithms that may be used for adjusting the coefficients of a finite impulse response adaptive filter, such as the RLS algorithm are also briefly described and compared with the LMS algorithm.

Echo cancellation using adaptive filters has its limitations when used in a PSTN. For example, a linear adaptive filter cannot exactly cancel the echo, even if the filter is perfectly adapted, because of non-linearities in the echo path and because of the echo dependent quantisation noise that is present. Several techniques for non-linear adaptive filtering have been proposed, and these are briefly considered in section 3.7.6.

Another aspect of echo canceller design, often over-looked, is that the adaptive filter must be controlled in some way. For example, when double-talk occurs the filter coefficients are very likely to diverge and hence increase the echo returned to the far-end talker. Various strategies have been proposed with the aim of reducing the coefficient divergence during double-talk [OCHI77], [DUTT78], [CURT81], [HAYA83], [FURU85], [MINA85], [YE91], [FUJI96]. However, some of these schemes are not suitable for use in a real echo canceller because they are unable to differentiate double-talk from a change of echo path, which would occur when a new hybrid is added in a conference call. This is because double-talk and changes in the echo path both result in an increase in the level of uncancelled echo.

3.2 The System Identification Configuration

Echo cancellation uses an adaptive filter in the system identification configuration, as shown in Figure 3.1.



The system identification configuration for echo cancellation

Figure 3.1

where

$s_1[n]$ = the far-end talker signal (the reference),

$e[n]$ = the echo,

$c[n]$ = the near-end signal comprising the echo e[n], the near-end talker $s_2[n]$ and the near-end background sounds b[n],

$p[n]$ = the predicted echo,

$r[n]$ = the canceller output,

$q[n]$ = the quantisation noise caused by companding process,

$s_2[n]$ = the near-end talker,

$b[n]$ = the near-end background sounds.

This configuration is called system identification because the adaptive filter seeks to identify the transfer function of an unknown system, which in this case is the impulse response of the near-end of the network. If the adaptive model is identical to the impulse response of the echo path, the predicted echo $p[n]$ is equal to the actual echo $e[n]$, and hence the echo is completely cancelled. Note that, in general, even for ideal

adaptation the error will not equal zero because of the presence of the near-end signal $d[n]$ and the quantisation noise $q[n]$.

The system must be adaptive because, when a call is initially set up, the response of the echo path is unknown. The response of the filter must therefore be adapted from its initial arbitrary state to one in which it is a good approximation to the impulse response of the echo path. Thereafter, the adaptive coefficients should be subject to continuous adjustment so that any gradual drift of the echo path is tracked. More substantial re-training will be needed if there is a substantial change of the end-path, as might occur when a call is redirected or a conference connection is established.

The adaptation process attempts to minimise some cost function that is related to the error $r[n]$. As the adaptation proceeds the error decreases and this drives the model to become more accurate, as measured by the cost function, until a steady state is achieved. The method by which the filter is adjusted depends on the type of filter and the cost function. If the adaptive filter of Figure 3.1 is of the FIR type then the predicted echo waveform, $p[n]$ may be written

$$p[n] = \sum_{i=0}^{N-1} w[i] s_1[n-i] \tag{3.1}$$

where N is the number of coefficients in the filter and $w[i]$ are the filter tap weights. Ideally, the difference between the actual echo, $e[n]$, and the predicted echo $p[n]$ would be used to adjust the coefficients of the adaptive filter. However, $e[n]$ is unavailable and therefore $c[n]$ is used instead. In the echo cancellation system $c[n]$ is given by:

$$c[n] = \sum_{i=0}^{\infty} h[i] s_1[n-i] + q[n] + s_2[n] + b[n] \tag{3.2}$$

$$= \sum_{i\ 0}^{N-1} h[i] s_1[n-i] + \sum_{i=N}^{\infty} h[i] s_1[n-i] + q[n] + s_2[n] + b[n]$$

$$= c_1[n] + c_2[n] + q[n] + d[n] \tag{3.3}$$

where

$h[i]=$ The impulse response of the echo path at delay i,

$c_1[n]=$ The component of the echo that is generated by the first N coefficients of the echo path impulse response,

$c_2[n]=$ The component of the echo that is generated by the part of the echo path outside the span of the adaptive filter.

Thus, the error term used in the adjustment may be written:

$$r[n]=c[n]-p[n] \tag{3.4}$$

$$=(c_1[n]-p[n])+c_2[n]+q[n]+d[n] \tag{3.5}$$

$$=(\mathbf{H}-\mathbf{W})^T\mathbf{S}_n+c_2[n]+q[n]+d[n] \tag{3.6}$$

where

$$\mathbf{S}_n=[s_1[n]\ \ s_1[n-1]\ \ s_1[n-2]\ \ .\ .\ .\ \ s_1[n-N+1]]^T \tag{3.7}$$

$$\mathbf{H}=[h[0]\ \ h[1]\ \ h[2]\ \ .\ .\ .\ \ h[N-1]]^T \tag{3.8}$$

$$\mathbf{W}=[w[0]\ \ w[1]\ \ w[2]\ \ .\ .\ .\ \ w[N-1]]^T \tag{3.9}$$

Two drawbacks are immediately evident from the error measurement in equation (3.6). Firstly, the error is corrupted by the near-end talker and background sounds, when present, and by the quantisation noise. It will be seen later that this corruption is likely to influence the performance of the adaptation process. Secondly, an FIR filter of finite length is attempting to approximate the infinite impulse response of the echo path. The implications of this will be considered in section 3.3.

Now that the error has been defined, the question of what form the cost function should take arises. A popular solution is to minimise the ensemble mean square error, usually because it yields tractable mathematics and because the solution has a single minimum point. This is the cost function that will be used here. However, one could

also attempt to minimise the mean absolute value of the error [HAYK96], or the mean of higher powers of the error. Another solution is to use a cost function defined in terms of a time average, rather than an ensemble average and this leads to the development of the RLS algorithm [HAYK96], [HUGH92].

Two equivalent approaches may be taken to obtain the set of coefficients, $\mathbf{W}_{opt}$, which give the minimum mean square error. These are to:

i) Use the principle of orthogonality, or

ii) Use an approach that highlights the quadratic error performance surface.

Although the use of orthogonality enables a solution to be derived quickly and easily, the second method is more useful for our purposes, because it leads directly to the development of gradient descent algorithms and it enables the operation of this type of algorithm to be more readily visualised.

3.3 Optimum Linear Filtering

The set of N coefficients that minimises the mean square error is known as the Wiener Filter, after Norbert Wiener who pioneered work in continuous-time optimum filtering [GARD94], [WIEN49]. In discrete-time Wiener filtering it is assumed that the unknown system that is to be modelled is linear, and that the reference and desired processes, $s_1[n]$ and $c[n]$ respectively, are wide-sense stationary and Gaussian. Now, from equation (3.4) the squared error is given by

$$r[n]^2 = c[n]^2 - 2c[n]\mathbf{S}_n^T\mathbf{W} + \mathbf{W}^T\mathbf{S}_n\mathbf{S}_n^T\mathbf{W} \qquad (3.10)$$

and taking the expectation of the square of the error gives the mean square error. Assuming that the tap input vector, $\mathbf{S}_n$, and $c[n]$ are independent of the filter coefficients $\mathbf{W}$, the mean square error $J[n]$ is given by:

$$J[n] = E\{\varepsilon[n]^2\} = E\{c[n]^2\} - 2\mathbf{B}^T\mathbf{W} + \mathbf{W}^T\mathbf{A}\mathbf{W} \tag{3.11}$$

where **A** and **B** are the auto-correlation matrix and cross-correlation vector, defined by:

$$\boldsymbol{B} = E\{c[n]\mathbf{S}_n\} = E\begin{Bmatrix} c[n]s[n] \\ c[n]s[n-1] \\ . \\ . \\ c[n]s[n-N+1] \end{Bmatrix} \tag{3.12}$$

$$\begin{aligned} \mathbf{A} &= E\{\mathbf{S}_n\mathbf{S}_n^T\} \\ &= E\begin{Bmatrix} s[n]s[n] & s[n-1]s[n] & . & . & s[n-N+1]s[n] \\ s[n]s[n-1] & s[n-1]s[n-1] & . & . & s[n-N+1]s[n-1] \\ . & . & . & & \\ . & . & . & & \\ s[n]s[n-N+1] & s[n-1]s[n-N+1] & . & . & s[n-N+1]s[n-N+1] \end{Bmatrix} \end{aligned} \tag{3.13}$$

The auto-correlation matrix is both Toeplitz and symmetric, and is therefore positive definite, i.e. $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ where **x** is an arbitrary column vector of length N [HAYK96]. This has the consequence that **A** is non-singular, has real positive eigenvalues and may be written in the form:

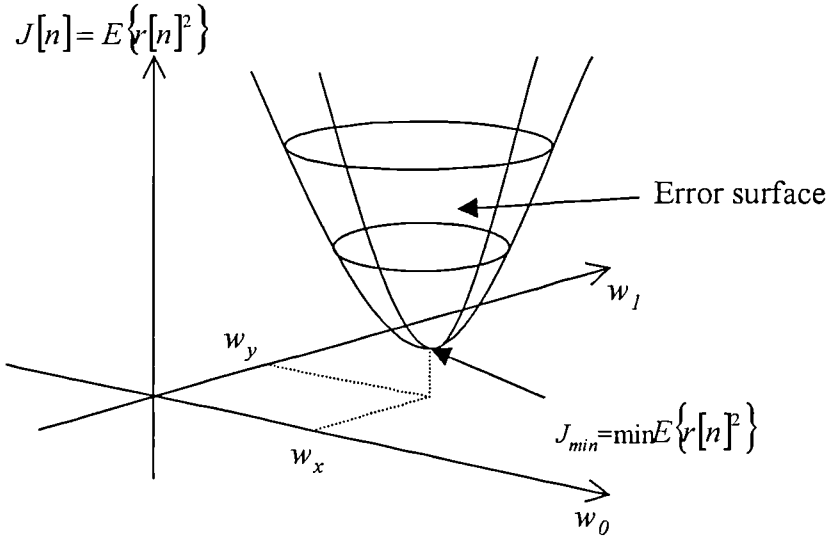
$$\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^T \tag{3.14}$$

where Λ is a diagonal matrix containing the (N-1) eigenvalues of **A**, and **Q** is a matrix of the corresponding eigenvectors.

$$\Lambda = diag[\lambda_1 \quad \lambda_2 \quad . \quad . \quad \lambda_{N-1}] \tag{3.15}$$

It will be shown later that (3.14) allows the convergence properties of the gradient descent, and LMS algorithms, to be written in terms of the eigenvalues of the autocorrelation matrix.

The mean square error described by (3.11) is a quadratic function and takes the form of a multi-dimensional paraboloid. This 'error surface' is very difficult to visualise in more than two dimensions. Figure 3.2 is a plot of the error surface when a two tap filter is used.



The 'error surface' for a two tap filter

Figure 3.2

3.3.1 The Optimum Coefficients

The optimum filter coefficients $\mathbf{W}_{opt}$ are those which minimise the mean square error $J[n]$, and in this case, the minimum MSE, J_{min}, lies at the 'bottom' of the paraboloid, as shown in Figure 3.2. The optimum coefficients may be found by first differentiating the mean square error as defined in (3.11), which gives the gradient vector, ∇, at any point on the error surface.

$$\nabla = -2\mathbf{B} + 2\mathbf{A}\mathbf{W} \qquad (3.16)$$

When the gradient is set equal to zero in (3.16), **W** will be the optimum set of coefficients $\mathbf{W}_{opt}$, and so

$$\mathbf{A}\mathbf{W}_{opt} = \mathbf{B} \tag{3.17}$$

Thus, pre-multiplying (3.17) by $\mathbf{A}^{-1}$ gives an equation for the optimum filter coefficients.

$$\mathbf{W}_{opt} = \mathbf{A}^{-1}\mathbf{B} \tag{3.18}$$

This is the matrix equivalent of the Wiener-Hopf equations [HUGH92] for a discrete time FIR system.

It was shown in equation (3.3) that the echo could be factorised into several different components. Using this equation, the cross-correlation vector may now be written as:

$$\begin{aligned}\mathbf{B} &= E\{c_1[n]\mathbf{S}_n\} + E\{c_2[n]\mathbf{S}_n\} + E\{q[n]\mathbf{S}_n\} + E\{s_2[n]\mathbf{S}_n\} + E\{b[n]\mathbf{S}_n\} \\ &= \mathbf{B}_1 + \mathbf{B}_2 + \mathbf{B}_3 + \mathbf{B}_4 + \mathbf{B}_5\end{aligned} \tag{3.19}$$

If the quantisation noise, near-end speech and background sounds have zero mean and are uncorrelated with the far-end talker, which is true in practice, then $\mathbf{B}_3$, $\mathbf{B}_4$ and $\mathbf{B}_5$ are zero:

$$\mathbf{B} = \mathbf{B}_1 + \mathbf{B}_2 \tag{3.20}$$

The optimum coefficients are now given by:

$$\begin{aligned}\mathbf{W}_{opt} &= \mathbf{A}^{-1}\mathbf{B}_1 + \mathbf{A}^{-1}\mathbf{B}_2 \\ &= \mathbf{H} + \mathbf{A}^{-1}\mathbf{B}_2\end{aligned} \tag{3.21}$$

Equation (3.21) shows that for a filter with N coefficients, the optimum coefficients are, in general, not equal to the first N samples of the echo path impulse response.

This implies that when the samples of the reference process are correlated, it is possible for $c_2[n]$ to be partially or completely cancelled even though the adaptive filter is incapable of modelling the impulse response that generated it. However, if the samples of the far-end reference are mutually uncorrelated, or if the filter can exactly model the echo path, then the optimum coefficients are given by:

$$\mathbf{W}_{opt} = \mathbf{H} \tag{3.22}$$

In all further discussion, it will be assumed that the filter length N is chosen such that $c_2[n]$ equals zero, and hence (3.22) is valid.

3.3.2 The Minimum Mean Square Error

The value of the minimum mean square error (MMSE), J_{min}, obtained using the optimum coefficients may be calculated by (3.18) and (3.11). This yields the following equation for J_{min}.

$$J_{\min} = E\{c[n]^2\} - \mathbf{B}^T \mathbf{W}_{opt} \tag{3.23}$$

If it is assumed that the length of the filter is chosen such that $c_2[n]$ can be completely cancelled, J_{min} can be written as:

$$J_{\min} = E\{d[n]^2\} + E\{q[n]^2\} \tag{3.24}$$

In other words, J_{min} is equal to the power of the near-end signal plus the power of the quantisation noise. In a network cancellation environment, an error power of zero can never be obtained when echo is present even if $d[n]=0$. The quantisation noise $q[n]$ is always present during periods of echo and this prevents J_{min} from equalling zero. Thus, the filter is unable to remove the quantisation noise even if the optimum

set of coefficients, $\mathbf{W}_{opt}$, is used. As will be discussed in the next chapter, the far-end talker will perceive the quantisation noise as distorted echo when the ERL is small.

3.4 The Gradient Search Technique

Several problems arise when attempting to calculate the optimum coefficients. Firstly, the auto-correlation matrix, **A**, and cross-correlation vector, **B**, must be known beforehand. In many applications including echo cancellation, these quantities are unknown and are possibly time varying. Secondly, the matrix inversion needed to calculate the ideal Wiener filter is a computationally intensive process requiring expensive hardware for real-time implementation.

The optimum coefficients can be calculated more efficiently by using the gradient search technique [HUGH92], which is sometimes referred to as the method of steepest descent [HAYK96], [WIDR76]. This procedure still requires advance knowledge of the auto-correlation matrix and cross-correlation vector, but leads to the well-known LMS algorithm that requires no prior knowledge of the signal statistics. In the gradient search technique, the coefficients are adjusted iteratively, so that they move towards their optimum settings in small steps. This approach is also known as the method of steepest descent because the tap weights move towards their optimum values in the direction given by the steepest gradient vector. The coefficients will always converge towards their optimum settings because the error surface has only a single global minimum and, therefore, no local minima.

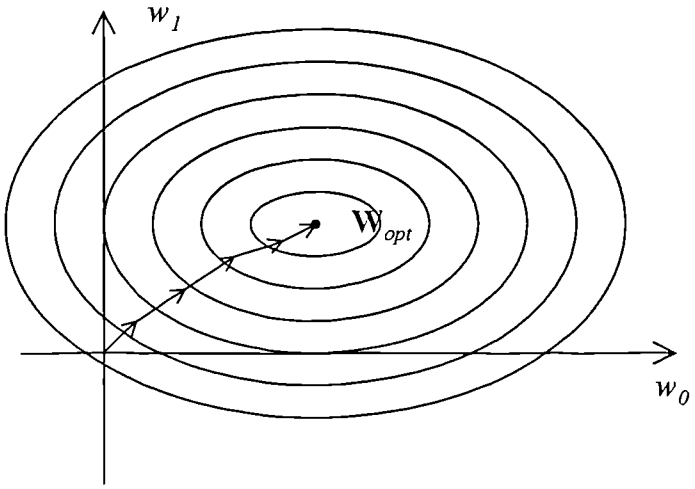
The update equation for finding the optimum coefficients is:

$$\mathbf{W}_{n+1} = \mathbf{W}_n + \mu(-\nabla_n) \tag{3.25}$$

where μ is a parameter called the step size, and

$$\nabla_n = -2\mathbf{B} + 2\mathbf{A}\mathbf{W}_n = 2\mathbf{A}(\mathbf{W}_n - \mathbf{W}_{opt}) \tag{3.26}$$

is the gradient vector computed at each iteration The step-size, μ, controls how quickly the coefficients converge to their optimum values. The convergence process is illustrated in Figure 3.3 for a two tap system. This diagram is identical to Figure 3.2 except viewed along the MSE axis. The ellipses are contour lines that represent points of equal mean square error, $J[n]$.



Convergence of the Gradient Search Algorithm

Figure 3.3

The negative gradient vector at any location on the error surface always points in the direction of steepest descent towards the optimum point, which lies at the bottom of the parabaloid. Hence, after every iteration the coefficients 'descend' the surface towards the optimum point, in steps that are proportional to the step size μ. If the step size is too large the process will not converge and the filter will be unstable.

It may be shown that the gradient search algorithm is guaranteed to converge for all step sizes within the range [HAYK96], [WIDR76]

$$0 < \mu < \frac{1}{\lambda_{\max}} \qquad (3.27)$$

where $\lambda_{\max}$ is the largest eigenvalue of the autocorrelation matrix. It may also be shown that the gradient descent algorithm converges to the optimum solution exactly

[WIDR76]. Once the coefficients equal their optimum values, the gradient becomes zero and so no further changes are made to the coefficient vector.

The gradient search method described here is still of limited practical use in an echo canceller application because knowledge of the auto-correlation matrix **A** and cross-correlation vector **B** is still required. This problem is overcome in the Least Mean Squares (LMS) algorithm.

3.5 The Least Mean Square (LMS) Algorithm

The LMS algorithm is similar to the gradient search method, except that it makes use of an estimated gradient vector when adjusting the coefficients. The estimated gradient is obtained by ignoring the statistical expectation operator in the exact gradient equation. This algorithm is a member of a family known as stochastic algorithms [HAYK96] the word stochastic being derived from a Greek word meaning to aim or guess at.

Using equation (3.26), the exact gradient is given by:

$$\nabla_n = -2E\{c[n]\mathbf{S}_n\} + 2E\{\mathbf{S}_n\mathbf{S}_n^T\}\mathbf{W}_n \qquad (3.28)$$

The estimated gradient, $\hat{\nabla}_n$, is obtained by ignoring the expectation operator in (3.28), and hence becomes:

$$\hat{\nabla}_n = -2c[n]\mathbf{S}_n + 2\mathbf{S}_n\mathbf{S}_n^T\mathbf{W}_n \qquad (3.29)$$

Taking $\mathbf{S}_n$ as a common factor on the RHS of this equation and re-arranging gives

$$\hat{\nabla}_n = -2r[n]\mathbf{S}_n \qquad (3.30)$$

Finally, substituting $\hat{\nabla}_n$ into (3.25) in place of the true gradient, ∇_n, gives the LMS update equation:

$$\mathbf{W}_{n+1} = \mathbf{W}_n + 2\mu r[n]\mathbf{S}_n \qquad (3.31)$$

The update equation now only requires knowledge of the current error $r[n]$, and the vector $\mathbf{S}_n$ containing the current and N-1 previous far-end reference samples. Both of these quantities are readily available.

The following sections examine the convergence and steady state properties of the LMS algorithm. It will be seen that although the LMS algorithm exhibits properties that are similar to the gradient search algorithm, from which it is derived, there are two important differences. The LMS algorithm will only converge to the optimum in the mean sense, i.e. any particular set of coefficients will be non-optimum but their average is the optimum. Once in the 'steady-state', the coefficients will fluctuate about their optimum settings by a factor that is proportional to the step size. Both of these differences arise directly from the use of an estimated gradient.

3.5.1 Convergence of the Mean Coefficient Error Vector

The error between the current set of coefficients and the optimum set is defined to be:

$$\boldsymbol{\varepsilon}_n = \mathbf{W}_n - \mathbf{W}_{opt} \qquad (3.32)$$

An equation for the coefficient error can be found first by using equation (3.6) and (3.31), and then subtracting $\mathbf{W}_{opt}$ from both sides to give:

$$\boldsymbol{\varepsilon}_n = \left(\mathbf{I} - 2\mu\mathbf{S}_n\mathbf{S}_n^T\right)\boldsymbol{\varepsilon}_{n-1} + 2\mu\left(d[n] + q[n]\right)\mathbf{S}_n \qquad (3.33)$$

Taking the expectation of both sides and assuming that the reference samples are independent of the coefficient vector, the average error may be expressed as

$$E\{\boldsymbol{\varepsilon}_n\} = (\mathbf{I} - 2\mu\mathbf{A})^n E\{\boldsymbol{\varepsilon}_0\} \tag{3.34}$$

where $\boldsymbol{\varepsilon}_0$ is the initial error vector. Rewriting $\mathbf{A}$ using equation (3.14) in the above equation gives

$$E\{\boldsymbol{\varepsilon}_n\} = \mathbf{Q}(\mathbf{I} - 2\mu\boldsymbol{\Lambda})^n \mathbf{Q}^T E\{\boldsymbol{\varepsilon}_0\} \tag{3.35}$$

Finally, by pre-multiplying both sides of (3.35) by $\mathbf{Q}^T$, the convergence of the mean coefficient error vector may be written as

$$E\{\boldsymbol{\varepsilon}'_n\} = (\mathbf{I} - 2\mu\boldsymbol{\Lambda})^n E\{\boldsymbol{\varepsilon}'_0\} \tag{3.36}$$

where

$$\boldsymbol{\varepsilon}'_n = \mathbf{Q}^T \boldsymbol{\varepsilon}_n \tag{3.37}$$

An equation similar to (3.36), but without the expectation operator may be obtained for the gradient search algorithm. The convergence of the coefficient vector may be examined by re-writing equation (3.36) to show the i^{th} element of $E\{\boldsymbol{\varepsilon}'_n\}$:

$$E\{\varepsilon'_{i,n}\} = \sum_{k-1}^{N\ 1} (1 - 2\mu\lambda_k)^n E\{\varepsilon'_{i,0}\} = E\{\varepsilon'_{i,0}\} \sum_{k-1}^{N-1} r_k^n \tag{3.38}$$

where r_k^n is given by:

$$r_k^n = (1 - 2\mu\lambda_k)^n \tag{3.39}$$

and λ_k is the k[th] eigenvalue of the auto-correlation matrix. Equation (3.38) shows that N exponentially decaying terms, or modes, of the form given in (3.39), govern the convergence of the mean error vector.

The decay rate for each mode is proportional to its eigenvalue and the step size μ. The largest eigenvalue λ_{max} controls the speed of the fastest converging mode, whilst the smallest eigenvalue λ_{min} sets the speed of the slowest converging mode. Each mode will decay exponentially towards zero only if $|1-2\mu\lambda_k|$ is less than one. Thus in order to guarantee convergence of the mean error vector, ε_n, the step size μ must be bounded by

$$0 < \mu < \frac{1}{\lambda_{max}} \tag{3.40}$$

If μ does not lie within these limits then one or more modes will grow and the adaptive process will become unstable. A time constant for each mode can be defined by recognising that the decay is an exponential process , i.e.

$$r_k = e^{-\frac{1}{\tau_k}} \tag{3.41}$$

Thus expanding (3.41) as a power series and equating the result with (3.39) gives:

$$1-2\mu\lambda_k = 1-\frac{1}{\tau_k}+\frac{1}{2!\tau_k^2}-..... \tag{3.42}$$

where τ_k is the time constant for the k^{th} mode, that describes the rate of convergence. For slow adaptation, the time constant τ_k is large and therefore (3.42) may be approximated as

$$1-2\mu\lambda_k \approx 1-\frac{1}{\tau_k} \tag{3.43}$$

Thus for the mean error vector, ε_n, the time constant of each mode is given approximately by

$$\tau_k = \frac{1}{2\mu\lambda_k} \tag{3.44}$$

It can be seen from the previous equations that a mode will decay faster as its associated eigenvalue becomes larger and as the step size is increased, within the limits of (3.40). One might expect that the overall rate of convergence would be determined by the smallest eigenvalue, however this turns out not to be the case. Using equation (3.44), the slowest converging mode has a time constant of

$$\tau_{(\max)} = \frac{1}{2\mu\lambda_{\min}} \tag{3.45}$$

while for fastest stable convergence, $\mu = 1/\lambda_{\max}$. Hence, the slowest time constant is now is given by:

$$\tau_{(\max)} = \frac{\lambda_{\max}}{2\lambda_{\min}} \tag{3.46}$$

The overall convergence is not dependent upon a single eigenvalue but, rather, equation (3.46) shows that the convergence speed is proportional to the ratio of the maximum to minimum eigenvalues. This ratio is often known as the eigenvalue spread of the auto-correlation matrix.

$$\text{Eigenvalue Spread} = \frac{\lambda_{\max}}{\lambda_{\min}} \tag{3.47}$$

The larger the eigenvalue spread the slower the overall convergence. It may be shown that for a Toeplitz autocorrelation matrix, the maximum and minimum

eigenvalues are bounded by the maximum and minimum values of the power spectral density function [HAYK96]. Thus

$$\min\{S(\omega)\} < \lambda_k < \max\{S(\omega)\} \tag{3.48}$$

where $S(\omega)$ is the power spectral density of the reference signal. As the filter length N tends towards infinity:

$$\lambda_{\min} \rightarrow \min\{S(\omega)\} \tag{3.49}$$

$$\lambda_{\max} \rightarrow \max\{S(\omega)\} \tag{3.50}$$

The overall convergence rate is dependent upon the spectral characteristics of the reference process. For signals such as white noise, which have a flat $S(\omega)$ and therefore an eigenvalue spread of one, all the modes converge at the same speed. Speech signals do not have a flat $S(\omega)$ and hence the eigenvalue spread is much larger than unity. Thus, the convergence will proceed more slowly when speech rather than white noise is used as the reference signal. This is one of the main limitations of the LMS algorithm governing its use in applications such as speech echo cancellation. Note that there are several circumstances under which the LMS algorithm may converge faster when the eigenvalue spread is increased [HAYK96]. The filter may also converge at a different speed depending on the initial coefficient vector value.

The dependence of convergence speed on the eigenvalue spread also has important consequences in the evaluation of echo canceller performance. Before the introduction of ITU-T recommendation G.168 [ITUT97], testing of network echo cancellers was performed with white noise using the procedures described in G.165 [ITUT94]. Although this kind of testing enables convergence properties, steady state behaviour etc, to be characterised, it only gives an approximate picture of how an echo canceller that uses the LMS algorithm might perform with speech. With the introduction of G.168, the white noise has been replaced by a so called 'composite source signal' (CSS), which simulates both the voiced and unvoiced components that are present in real speech. The use of these signals as test waveforms gives a clearer

picture of how the adaptive filter and the associated control logic would perform when real speech signals are used.

3.5.2 Steady State Behaviour of the Coefficient Vector

Equation (3.34) states that if the adaptive process is stable and is convergent, then for a large number of iterations the mean coefficient error vector, $E\{\varepsilon_n\}$, becomes zero. This of course does not mean that the error ε_n actually equals zero, but just that its mean over a sufficiently large number of training sessions tends towards zero. Clearly, the mean could be zero but the coefficients might be fluctuating about their optimum values. The variation of the coefficients may be characterised by calculating the mean Euclidean norm $E|\varepsilon_n|^2$, which is the mean length of the error vector, and is given by:

$$E|\varepsilon_n^2| = E\{\varepsilon_n^T \varepsilon_n\} = \sum_{i=0}^{N-1} E\{\varepsilon_{i,n}^2\} \qquad (3.51)$$

where $\varepsilon_{i,n}$ is the i^{th} error in the coefficient vector at iteration n.

The mean norm therefore represents the total mean squared error of all the filter coefficients. The norm of the error vector given in equation (3.33) is

$$\varepsilon_n^T \varepsilon_n = \varepsilon_{n-1}^T \left(\mathbf{I} - 2\mu \mathbf{S}_n \mathbf{S}_n^T\right)^T \left(\mathbf{I} - 2\mu \mathbf{S}_n \mathbf{S}_n^T\right) \varepsilon_{n-1} + 4\mu^2 \left(d^2[n] + q^2[n]\right) \mathbf{S}_n^T \mathbf{S}_n \qquad (3.52)$$
$$+ \text{other terms that are zero when expectation is taken}$$

Using the relations $\left(\mathbf{S}_n \mathbf{S}_n^T\right)^T = \mathbf{S}_n \mathbf{S}_n^T$, $\left(\mathbf{S}_n \mathbf{S}_n^T\right)^T \mathbf{S}_n \mathbf{S}_n^T = \mathbf{S}_n^T \mathbf{S}_n \mathbf{S}_n \mathbf{S}_n^T$ the above equation becomes

$$\varepsilon_n^T \varepsilon_n = \varepsilon_{n-1}^T \left(\mathbf{I} - 4\mu \mathbf{S}_n \mathbf{S}_n^T + 4\mu^2 \mathbf{S}_n^T \mathbf{S}_n \mathbf{S}_n \mathbf{S}_n^T\right) \varepsilon_{n-1} + 4\mu^2 \left(d^2[n] + q^2[n]\right) \mathbf{S}_n^T \mathbf{S}_n \qquad (3.53)$$

In order to simplify the evaluation of this expression it is assumed that for small step sizes, the following approximation may be used [MESS84], [HAYK96], [KUSH84]:

$$\left(\mathbf{I}-4\mu\mathbf{S}_n\mathbf{S}_n^T+4\mu^2\mathbf{S}_n^T\mathbf{S}_n\mathbf{S}_n\mathbf{S}_n^T\right)\approx\left(\mathbf{I}-4\mu E\left\{\mathbf{S}_n\mathbf{S}_n^T\right\}+4\mu^2E\left\{\mathbf{S}_n^T\mathbf{S}_n\mathbf{S}_n\mathbf{S}_n^T\right\}\right) \qquad (3.54)$$

Now, if we use relation (3.14) and if the reference process is zero mean white Gaussian noise with variance σ_0^2, then (3.53) may be written as

$$\begin{aligned}\boldsymbol{\varepsilon}_n'^T\boldsymbol{\varepsilon}_n' &= \boldsymbol{\varepsilon}_{n-1}'^T\left(\mathbf{I}-4\mu\boldsymbol{\Lambda}+4\mu^2N\sigma_0^2\boldsymbol{\Lambda}\right)\boldsymbol{\varepsilon}_{n-1}'+4\mu^2\left(d^2[n]+q^2[n]\right)N\sigma_0\\ &= \boldsymbol{\varepsilon}_{n-1}^T\boldsymbol{\varepsilon}_{n-1}\left(1-4\mu\sigma_0^2+4\mu^2N\sigma_0^4\right)+4\mu^2d^2[n]N\sigma_0^2\end{aligned} \qquad (3.55)$$

since for a white reference $\boldsymbol{\varepsilon}_n'=\boldsymbol{\varepsilon}_n$ and therefore $\boldsymbol{\varepsilon}_n^T\boldsymbol{\varepsilon}_n=\boldsymbol{\varepsilon}_n'^T\boldsymbol{\varepsilon}_n'$. The average norm may be written as

$$E\left\{\boldsymbol{\varepsilon}_n^T\boldsymbol{\varepsilon}_n\right\}=E\left\{\boldsymbol{\varepsilon}_{n-1}^T\boldsymbol{\varepsilon}_{n-1}\right\}\left(1-4\mu\sigma_0^2+4\mu^2N\sigma_0^4\right)+4\mu^2N\sigma_0^2J_{\min} \qquad (3.56)$$

This equation may now be rewritten in terms of a transient component that decays to zero, and a steady state component. Expressing (3.56) in terms of the norm of the initial error vector, $\boldsymbol{\varepsilon}_0$, gives

$$\begin{aligned}E\left\{\boldsymbol{\varepsilon}_n^T\boldsymbol{\varepsilon}_n^T\right\}&=E\left\{\boldsymbol{\varepsilon}_0^T\boldsymbol{\varepsilon}_0\right\}\left(1-4\mu\sigma_0^2+4\mu^2N\sigma_0^4\right)^n\\ &\quad+4\mu^2N\sigma_0^2J_{\min}\sum_{i=0}^{n-1}\left(1-4\mu\sigma_0^2+4\mu^2N\sigma_0^4\right)^i\end{aligned} \qquad (3.57)$$

If the adaptation process is stable then the first term decays to zero whilst the second term rises to a limiting value from zero. The relative magnitudes and rates of change of the two terms are normally such that little accuracy is lost if the second term is replaced by its limiting value. This gives:

$$E\left\{\boldsymbol{\varepsilon}_n^T\boldsymbol{\varepsilon}_n\right\}=E\left\{\boldsymbol{\varepsilon}_0^T\boldsymbol{\varepsilon}_0\right\}\left(1-4\mu\sigma_0^2+4\mu^2N\sigma_0^4\right)^n+\frac{\mu NJ_{\min}}{1-\mu N\sigma_0^2} \qquad (3.58)$$

The first term on the RHS of (3.58) describes $E\left\{\boldsymbol{\varepsilon}_n^T\boldsymbol{\varepsilon}_n\right\}$ during the initial stages of adaptation, i.e. the transient response. The second term prevents $E\left\{\boldsymbol{\varepsilon}_n^T\boldsymbol{\varepsilon}_n\right\}$ decaying

to zero and hence determines the minimum error that is possible for a given step size, filter length and reference power.

Using (3.58), the adaptation process is stable if the step-size lies within the range

$$0 < \mu < \frac{1}{N\sigma_0^2} \tag{3.59}$$

The step-size that gives the fastest convergence is calculated by differentiating $\left(1 - 4\mu\sigma_0^2 + 4\mu^2 N\sigma_0^4\right)$ with respect to μ and equating the result to zero. This yields:

$$\mu_{opt} = \frac{1}{2N\sigma_0^2} \tag{3.60}$$

which is half the maximum value permitted by (3.59). These equations show that increasing the length of the filter decreases the maximum step size that may be used to obtain stable convergence.

Manipulation of the transient term in (3.58) shows that the convergence of the Euclidean norm has a time constant approximately given by:

$$\tau = \frac{1}{4\mu\sigma_0^2\left(1 - \mu N\sigma_0^2\right)} \tag{3.61}$$

and this results in an initial convergence rate of:

$$10\log\left(\exp\left(-\frac{1}{\tau}\right)\right) \times f_s \text{ dB/s} \tag{3.62}$$

where f_s is the sample rate, which in this case is 8kHz. Notice that the convergence speed for the LMS algorithm is not only dependent upon the eigenvalues of the autocorrelation matrix and step size, but also upon the length of the filter. Equation

(3.61) states that for a fixed step-size, increasing the length of the filter decreases the initial convergence rate.

After the transient component has decayed, the norm of the coefficient error vector approaches a value given by:

$$E\{\boldsymbol{\varepsilon}_{\infty}^{T}\boldsymbol{\varepsilon}_{\infty}\}=\frac{\mu N}{1-\mu N\sigma_0^2}J_{\min} \qquad (3.63)$$

This equation indicates that in order to reduce the steady state error power for a given filter length, $\mu J_{\min} = \mu E\{d[n]^2 + q[n]^2\}$ must be as small as possible. As J_{min} is large during periods of double talk, unwanted fluctuations will be introduced into the coefficient vector and this will cause the level of residual echo to increase. This problem can be alleviated by decreasing μ (or setting μ=0) during periods of double-talk. Other sources of disturbance such as the quantisation noise, $q[n]$, and background sounds, $b[n]$, will also increase $E\{\boldsymbol{\varepsilon}_{\infty}^{T}\boldsymbol{\varepsilon}_{\infty}\}$, and therefore increase the power of the residual echo.

Equation (3.63) also shows that increasing the filter length increases $E\{\boldsymbol{\varepsilon}_{\infty}^{T}\boldsymbol{\varepsilon}_{\infty}\}$. To obtain the same $E\{\boldsymbol{\varepsilon}_{\infty}^{T}\boldsymbol{\varepsilon}_{\infty}\}$ compared with a smaller filter requires the use of a smaller step size, and this in turn may lead to unacceptably slow convergence.

3.5.3 Convergence of the Mean Square Error

Although the coefficients converge towards the optimum in the mean sense, we also need to show that the mean squared error of equation (3.23) converges towards the minimum value that results when the optimum filter is used.

The theoretical MSE, $J[n]$, may be calculated in terms of the coefficient error vector. This is achieved by rearranging the equation for $J[n]$ given in (3.23) to give $E\{e[n]^2\}$, and substituting this (3.11).

$$J[n] = J_{\min} + \varepsilon_n'^T \Lambda \varepsilon_n' \tag{3.64}$$

Equation (3.64) shows that $J[n]$ is governed by the J_{min} and by another term that is known as the excess mean square error. The average excess MSE denoted by $J_{ex}[n]$ is given by

$$J_{ex}[n] = E\{\varepsilon_n'^T \Lambda \varepsilon_n'\} = \sum_{i=1}^{N} \lambda_i E\{\varepsilon_{i,n}'^2\} \tag{3.65}$$

where $\varepsilon_n' = [\varepsilon_{1,n}' \quad \varepsilon_{2,n}' \quad . \quad . \quad \varepsilon_{N,n}']^T$. For a white reference with variance σ_0^2, $J_{ex}[n]$ can be rewritten as

$$\begin{aligned} J_{ex}[n] &= \sigma_0^2 \sum_{i=1}^{N} E\{\varepsilon_{in}'^2\} \\ &= \sigma_0^2 E\{\varepsilon_n^T \varepsilon_n\} \end{aligned} \tag{3.66}$$

The quantity $E\{\varepsilon_n^T \varepsilon_n\}$ was calculated in equation (3.58) and now $J_{ex}[n]$ can be written as

$$J_{ex}[n] = E\{\varepsilon_0^T \varepsilon_0\}\left(1 - 4\mu\sigma_0^2 + 4\mu^2 N\sigma_0^4\right)^n \sigma_0^2 + \frac{\mu N \sigma_0^2 J_{\min}}{1 - \mu N \sigma_0^2} \tag{3.67}$$

$$= J_{tr}[n] + J_{ss}[n] \tag{3.68}$$

As with as $E\{\varepsilon_n^T \varepsilon_n\}$, the average excess MSE consists of two components that describe the initial and final behaviour of the adaptation process. The first term, $J_{tr}[n]$, characterises the transient part of the mean square error that is present during initial convergence. It is clear that the transient component decreases if the conditions of equation (3.59) are satisfied, and has a time constant given by (3.61). The convergence of the MSE exhibits the same behaviour as the mean Euclidean norm of the error vector, i.e. the initial error power decreases at the same speed.

3.5.4 Steady State Behaviour of the Mean Square Error

The second term in equation (3.68) predicts the excess MSE after the transient components have become insignificant. In echo cancellation terms, $J_{ss}[n]$ is the residual echo power that is present after the convergence process has finished. If the echo is completely cancelled $J_{ss}[\infty]=0$ and hence $J[\infty]=J_{min}$. Therefore, the ratio of the excess mean squared error to the minimum mean square error describes the effect of the misadjustment in the coefficient vector during the steady state. The misadjustment M, for a white reference process, is defined to be

$$M = \frac{J_{ss}[\infty]}{J_{\min}} = \frac{\text{Residual Echo Power}}{\text{Near - end Power}} \tag{3.69}$$

$$= \frac{\mu N \sigma_0^2}{1 - \mu N \sigma_0^2} \tag{3.70}$$

For the LMS algorithm the misadjustment is dependent upon step size, filter length and the power of the reference signal. If these factors are constant then the misadjustment remains constant. Thus if near-end talker occurs during adaptation then the residual echo power will increase because of *unwanted coefficient* fluctuation, but the ratio of residual power to near-end power (talker now present) remains constant.

3.6 LMS Performance

A number of tests were performed to investigate the validity of the predictions of the previous sections with regard to the rate of adaptation and the minimum mean squared error of the LMS and gradient search techniques. The characteristics of the LMS and gradient search algorithms have been evaluated to show how the step size and filter length affect the convergence speed and steady state error. For each set of conditions several 'learning curves' were computed, and these show:

i) the minimum mean square error, J_{min}, (blue),

ii) the mean square error for a single run , $J[n]$, (red),

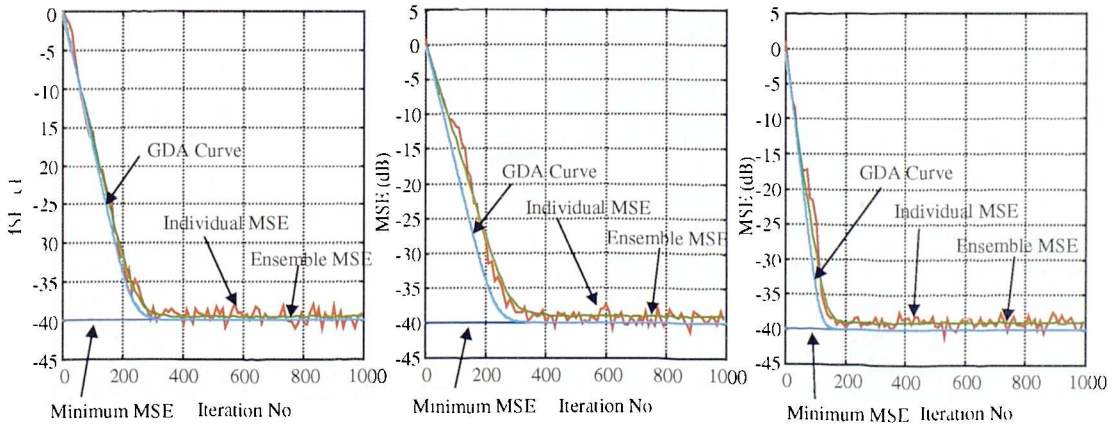
iii) the ensemble mean square error, calculated using 200 separate learning curves (green),

and iv) the theoretical learning curve that results from using the true gradient, i.e. the learning curve of the gradient descent algorithm (cyan).

In all cases the reference process is white noise with unity standard deviation. J_{min} was set using white noise with standard deviation = 0.01. Table 3.1 shows how the initial convergence speed and final misadjustment compare with the theoretical values. The learning curves are shown in Figure 3.4.

Figure	Step Size μ	Filter Length N	Theoretical Convergence Speed (dB/s)	Measured Convergence Speed (dB/s)	Theoretical Misadjustment	Measured Misadjustment
(a)	0.01	10	-1250	-1265	11%	11.4%
(b)	0.01	20	-1111	-1103	25%	25.5%
(c)	0.02	10	-2223	-2237	25%	26.1%

Table 3.1

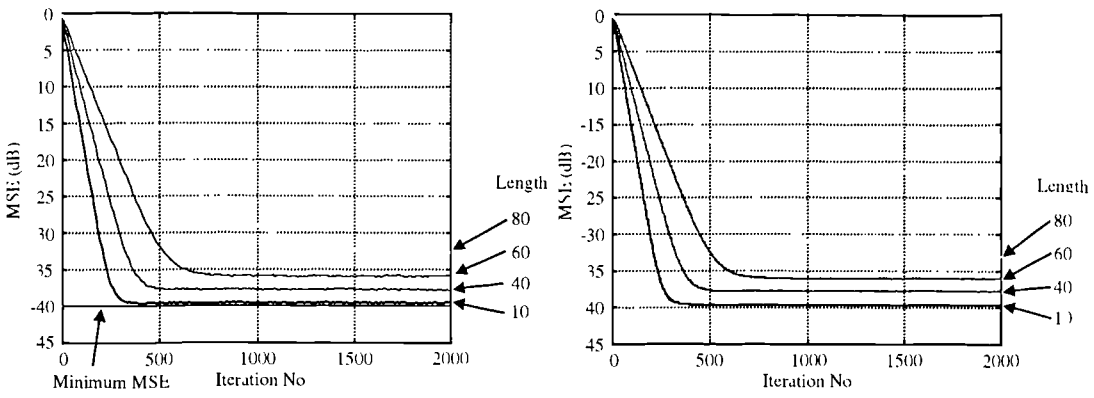


a) μ=0.01, N=10, b) μ=0.01, N=20, c) μ=0.02, N=10

Figure 3.4

The graphs show that the gradient descent algorithm converges to the MMSE, with the convergence speed being independent of the filter length. Initially, the ensemble learning curve for the LMS algorithm converges at a similar speed, but as the MSE decreases it begins to lag behind because of the effects of gradient noise. It may be seen that in the steady state the MSE is always greater than the theoretical minimum, due to the effect of the steady state component in equation (3.68). Table 3.1 shows that for the three different conditions, the measured misadjustment is approximately equal to the theoretical values. As predicted, doubling the step size will approximately double the convergence speed, but will also approximately double the misadjustment. Similarly doubling the filter length approximately doubles the final misadjustment, and decreases the convergence speed.

The effect of filter length on convergence speed can be seen more clearly in Figure 3.5. The measured and theoretical learning curves obtained using a 200 ensemble average were calculated using four different filter lengths and a step size of 0.01.



a) measured characteristic, b) theoretical characteristic

Figure 3.5

As expected, the convergence speed slows with increasing filter lengths. The measured and theoretical characteristics are in good agreement.

3.7 Other Adaptation Algorithms

Previous sections have discussed the operation and performance of the basic LMS algorithm. There are of course other techniques that may be used to adjust the filter coefficients. This section briefly discusses some important variations of the LMS algorithm. Also briefly described is the Recursive Least Squares algorithm, which attempts to minimise a different cost function. The relative merits of these techniques are compared with the baseline LMS algorithm.

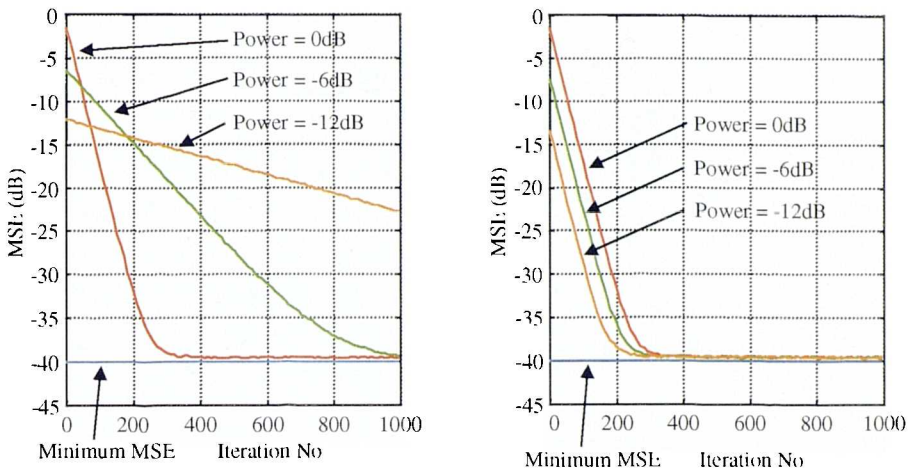
3.7.1 The Normalised, NLMS, Algorithm

A simple modification that can be made to the basic LMS algorithm is to normalise the step size μ in relation to the power of the reference signal. The purpose of normalisation is primarily to make the convergence speed independent of the far-end reference power. Additionally, normalisation of the step size prevents the LMS algorithm from becoming unstable as the reference power increases.

Normalisation is usually accomplished by modifying the step size using:

$$\mu' = \frac{\mu}{\sigma_0^2 + b} \tag{3.71}$$

where b is a constant that is chosen to prevent the normalised step size, μ', from becoming too large, and σ_0^2 is an estimate of reference signal power. Figure 3.6 shows how the learning rate varies for the LMS algorithm as the reference power is changed.



a) without normalisation, b) with normalisation

Figure 3.6

It may be seen that without normalisation the convergence speed slows considerably as reference power decreases. With normalisation the convergence speed is identical for all signal powers (within the limits of numerical accuracy). Obviously if there is no far-end signal there cannot be any convergence and hence the constant b is chosen to prevent μ' from becoming too large.

For the normalised algorithm the total coefficient vector MSE in the steady state is approximately given by (assuming that the reference signal is present and b=0)

$$E\left\{\varepsilon_\infty^T \varepsilon_\infty\right\} = \frac{\mu N}{1-\mu N} \cdot \frac{J_{\min}}{\sigma_0^2} \tag{3.72}$$

and the misadjustment is given by

$$M = \frac{\mu N}{1-\mu N} \tag{3.73}$$

This equation suggests that the misadjustment is constant, dependent only upon the step size and filter length. In other words the residual echo power is always M times the minimum mean square error value. The normalisation also prevents the adaptation process becoming unstable as the reference power increases. Equation (3.59) determines the maximum step size for a given reference power, and shows that

the maximum permitted step size decreases with increasing reference power. Thus, the adaptation process would become unstable if the maximum permitted step size were to become less than the step size being used to adapt the coefficients.

3.7.2 Variable Step Size (VSS) LMS Algorithms

In the environment of a real PSTN the echo path will not be stationary, i.e. it will slowly change as a call progresses. Any echo canceller must therefore be able to adapt to this slow change in the echo path. One method that may be used to analyse the behaviour of the LMS algorithm in such an environment [WIDR76] is to split the mean squared error into two terms, that are due to coefficient vector noise, and coefficient vector lag. Unsurprisingly, the tracking capabilities of the LMS algorithm are dependent upon the step size. The analysis in [WIDR76] shows that there is an optimum step size where the MSE due to gradient noise is equal to the MSE due to lag.

The trade-off that exists between adaptation speed, steady state misadjustment and tracking performance, for a fixed step size can be relaxed by making the step size variable. Ideally, the step size should be large during initial convergence and small enough during 'steady state' to track slow changes of the impulse response. If there is a large change in the echo path, as could occur if a call were to be switched to a different extension, then the power of the echo returned to the far-end would suddenly increase. In these circumstances, the step-size should be large so that the new impulse response can be learned as quickly as possible. Various methods have been proposed for the implementation of variable step size algorithms, some of which are briefly discussed here.

In the VSS algorithm proposed by [HARRI86] each coefficient has its own adaptation constant, whose size is controlled by examining the sign of the corresponding gradient estimate. If the sign of several consecutive gradient estimates is identical, either initial convergence is occurring or the optimum value for the tap has changed. In this case, the step size is increased, unless it has reached its

maximum permitted value. Similarly, if there are several consecutive sign changes, it is likely that the coefficient is fluctuating about the mean value and in this case, the step size of that coefficient is reduced. The choice of decrement is critical for a particular coefficient, if the step size decreases too quickly adaptation will effectively stop and the potentially reduced MSE will not be realised.

In the VSS algorithm proposed by [KWON92] a single adaptation constant is used for all coefficients, whose size is proportional to the square of the error $r[n]$. The adaptation constant is updated using:

$$\mu_{n+1} = \alpha\mu_n + \gamma r^2[n] \tag{3.74}$$

where α is a forgetting factor and γ is a small positive constant that determines the degree of influence that the square error has on the adaptation constant.

In another technique proposed by Benveniste et al. and discussed by [HAYK96], the adaptation constant is adjusted in a manner that is analogous to the adaptation of the filter coefficients. In this technique, the step size is adjusted using the following equation

$$\mu_{n+1} = \mu_n + \alpha\hat{\nabla}_\mu(n) \tag{3.75}$$

where $\hat{\nabla}_\mu(n)$ is an estimated gradient (in terms of the coefficient vector and the reference input vector), that is obtained by differentiating the MSE with respect to the adaptation constant. The estimated gradient describes how the step size should be changed in order to minimise the MSE for the current coefficient vector.

Although these techniques may lead to improved convergence and tracking behaviour, their effectiveness is still limited because, for the LMS algorithm, the adaptation of each coefficient is controlled by a mixture of modes. Better

performance might be obtained by using a separate adaptation constant for each mode.

3.7.3 Block LMS Algorithm

In this technique [HAYK96] the reference and desired signals are sectioned into blocks and are used to adapt the filter coefficients on a block by block basis, rather than sample by sample. Analysis of the algorithm shows that because of the use of data blocks, an averaged gradient whose accuracy improves with increasing block length, is used to adjust the filter coefficients. In spite of this, analysis also shows that the convergence speed and misadjustment of the BLMS algorithm are identical to those of the standard algorithm [HAYK96] because convergence speed is still determined by the eigenvalues of the reference process. In addition, the choice of step size in the BLMS method is more restricted than in the LMS algorithm. This may lead to slower convergence or, if the block size is large, to instability, i.e. the filter is unable to converge because the required criteria cannot be satisfied.

Where the BLMS algorithm is superior to the LMS algorithm is in the way it can be implemented. Specifically, the use of data blocks enables the calculation of $p[n]$ (linear convolution) and the gradient estimate (linear correlation) to be performed in the frequency domain using Fast Fourier Transforms (FFTs), and this results in a reduction in the computational load. It may also be shown that each mode in the frequency domain corresponds directly to a single frequency domain coefficient, and this enables convergence speed to be optimised for each mode.

3.7.4 Self Orthogonalising Adaptive Filters

It was demonstrated in section 3.5.1 that the convergence properties of the LMS algorithm are highly dependent upon the spectral characteristics of the reference waveform. When the reference is white, the eigenvalues of the auto-correlation matrix are identical and the LMS algorithm exhibits the fastest possible convergence obtainable for a given step size. However, when the reference is non-white, the

eigenvalue spread of the auto-correlation matrix is greater than unity and the LMS algorithm takes longer to converge for a given step size.

In a self-orthogonalising adaptive filter, the reference samples are 'transformed' such that they become statistically independent. In other words, the spectrum of the reference process is whitened. For an ideal self-orthogonalising filter, the adaptation process performs as if white noise was used as the reference waveform. In [MULG88], the input vector $\mathbf{S}_n$ is 'transformed' by a matrix **P** so that the new tap input vector is given by:

$$\mathbf{Z}_n = \mathbf{P}\mathbf{S}_n \tag{3.76}$$

where the auto-correlation matrix of $\mathbf{Z}_n$ is given by:

$$E\{\mathbf{Z}_n\mathbf{Z}_n^T\} = \mathbf{I} \tag{3.77}$$

It may be shown [MULG88] that if the auto-correlation matrix is written in terms of its eigenvalues, a suitable transform is:

$$\mathbf{P} = \Lambda^{-\frac{1}{2}}\mathbf{Q}^{-1} \tag{3.78}$$

Where Λ is an N×N matrix whose diagonals are the eigenvalues of **A**, and **Q** is an N×N matrix containing the corresponding eigenvectors.

To illustrate the othogonalising properties, consider the LMS update equation in which the tap input vector $\mathbf{S}_n$ has been 'transformed' using the matrix **P**.

$$\mathbf{W}_{n+1} = \mathbf{W}_n + 2\mu\mathbf{Z}_n r[n] \tag{3.79}$$

where $0<2\mu<1$ and all other symbols have their previously defined meanings. Equation (3.34) now becomes:

$$E\{\varepsilon_{n+1}\} = [1-2\mu]E\{\varepsilon_n\} \tag{3.80}$$

Equation (3.80) shows that such an adaptive system is convergent if $0<2\mu<1$ and that the convergence properties are independent of the reference statistics. This adaptation algorithm is obviously of limited practical use because the auto-correlation matrix must be known beforehand. For a more detailed discussion of practical self-orthogonalising adaptive filters, see [HAYK96] and [MULG88].

3.7.5 The Recursive Least Squares (RLS) Algorithm

The Wiener theory that was used to derive the LMS algorithm is based on the use of ensemble averages, and hence the resulting filter coefficients are optimum only in a statistical sense. The RLS algorithm is based on time averages rather than ensemble averages, which results directly from the use of the following cost function

$$\sum_{i=0}^{n}(c[i]-p[i])^2\alpha^{n-i} \tag{3.81}$$

where $0<\alpha<1$ and is used to determine the influence that old samples have on the summation. The optimum coefficients are described by the normal equations which are the equivalent of the Wiener-Hopf equations (3.17). The use of time averages enables an iterative technique to be used to estimate the inverse auto-correlation matrix, and this leads to the following equations which describe the RLS algorithm

$$\mathbf{g}_n=\sum_{i=0}^{n}\alpha^{n-i}\mathbf{S}_i\mathbf{S}_i^T \tag{3.82}$$

$$\mathbf{g}_n^{-1}=\frac{1}{\alpha}\left[\mathbf{g}_{n-1}^{-1}-\frac{\mathbf{g}_{n-1}^{-1}\mathbf{S}_n\mathbf{S}_n^T\mathbf{g}_{n-1}^{-1}}{\alpha+\mathbf{S}_n^T\mathbf{g}_{n-1}^{-1}\mathbf{S}_n}\right] \tag{3.83}$$

$$r[n]=c[n]-\mathbf{W}_{n-1}^T\mathbf{S}_n \tag{3.84}$$

$$\mathbf{W}_n=\mathbf{W}_{n-1}+\mathbf{g}_n^{-1}\mathbf{S}_n^T r[n] \tag{3.85}$$

A very short convergence time can be achieved using the RLS algorithm, which is in the order of N iterations. In addition, the convergence is independent of the reference process eigenvalues. The enhanced convergence properties are achieved because the adaptation process uses all the available data rather than just the instantaneous data as in the LMS algorithm. One disadvantage of the RLS technique described by equations (3.82) to (3.85) is that the processing requirements are significantly greater than for the LMS algorithm. Whereas the LMS only requires $2N$ multiplies the RLS requires $2N^2$. For example in the implementation of a 256 tap filter, the processing requirement for the RLS algorithm is 128 times greater than needed for the LMS algorithm. However, the largest drawback to the RLS algorithm is not one of computational load, as the use of fast RLS algorithms [HAYK96] can reduce the number of multiplies to the order of 7N. The major disadvantage is that the RLS algorithm and its fast variants are numerically unstable, and this generally rules out their use in a speech echo canceller that may be required to operate robustly over a period of many hours.

3.7.6 Non-Linear Adaptive Filtering

When the methods described in sections 3.3, 3.4 and 3.5 are applied in an echo canceller, the echo path is being modelled as a linear process. In consequence, even if the optimum coefficients determined by these methods are used, components of the echo that result from network non-linearities will not be removed. When the ERL is small, the far-end subscriber will perceive the quantisation noise as distorted echo and this is just as unacceptable as the uncancelled echo of a linear system. Network echo cancellers generally use a centre-clipper, which acts like a low-level echo suppressor, to remove such uncancelled echo.

The companding quantisation process may be regarded as an important source of non-linearity in telephone connections because its behaviour does not obey the superposition principle. Moreover, the non-linear part of companding may be regarded, generally, as irreversible, because it is impossible to know the compander input value given only its output value.

An alternative to the use of a centre-clipper might be to use a non-linear adaptive filter that is able to model, and therefore permit cancellation of, both the linear and non-linear distortions that occur in the echo path. Non-linear adaptive filtering is commonly and successfully used in systems that support full-duplex digital data transmission. In such systems, if the echo of the transmitted signal appears at the receiver input then the waveform transmitted from the far-end will be corrupted. It has been reported that channel attenuations of 40dB are commonly encountered [HUGH92] and therefore the echo power can be significantly larger than the power of the far-end transmissions. Any non-linearity in the echo path may prevent a linear canceller from reducing the echo power sufficiently to enable error free demodulation of the far-end transmissions.

Three types of non-linear adaptive filter are described in [HUGH92], these are: look-up table filters, Volterra-series filters and neural networks. Although these techniques might be used to compensate for non-linearities caused, for example, by the iron cored transformers in a hybrid, or by a mismatch between the companding curves of a pair of codecs, it seems unlikely that they will be generally useful for diminishing the dominant, and irreversible, distortion introduced by quantisation.

3.8 Conclusions

The basic principles of adaptive filtering and their application to echo cancellation have been introduced in this chapter. In an echo canceller, the adaptive filter estimates the echo waveform, which is then subtracted from the received echo waveform. Ideally, the echo is completely cancelled when the estimated echo equals the actual echo. In practice, for a number of reasons, the predicted echo will not be an exact replica of the actual echo so that cancellation will not be complete and an echo residue will result.

Several important points have emerged. The optimum filter coefficients are calculated by minimising the ensemble mean square error between the actual echo and the predicted echo. For an adaptive FIR filter of order N, the optimum

coefficients will, in general, not be equal to the first N samples of the echo path impulse response. This is because when the samples of the reference process are mutually correlated, they are also correlated with the component of the echo generated by the impulse response that lies outside the time span of the adaptive filter. Thus, it may be possible to obtain complete or partial cancellation of the echo generated outside the span of the adaptive filter, depending upon the statistics of the reference waveform.

Additionally, the development of equation (3.21) leads to the conclusion that, provided the filter order is sufficiently large, the optimum coefficients will always be equal to the first N samples of the echo path impulse response regardless of the reference waveform. For this unique solution to exist the determinant of the auto-correlation matrix, **A**, must be non-zero and hence the matrix must be invertible. When the determinant is zero, the matrix is said to be singular and in this case, there will be an infinite number of solutions to the equations described by (3.17). This situation could arise, for example, when the reference signal is a single sinusoid.

However, any practical implementation of the LMS algorithm will be realised in a finite precision environment and thus, the auto-correlation matrix will become effectively singular when the determinant falls below some small positive number. As the determinant of the auto-correlation matrix approaches this numerical limit, the total number of solutions increases, i.e. the optimum solution belongs to a set of optimum solutions. Thus in a finite precision system ill-conditioning of the reference waveform autocorrelation matrix causes the 'target' solution of the LMS algorithm to deviate from the unique optimum obtained when the reference is white. The result of this is that the optimum coefficients calculated during one segment of a speech waveform are not likely to be optimum subsequent segments, and thus the level of echo cancellation is reduced. One way in which this problem can be solved is to add a small amount of noise, known as 'dither', into the reference waveform, or more usually, to introduce leakage into the coefficients [HAYK96].

The gradient search algorithm has been described and was used to derive the LMS algorithm. The main advantages of the LMS algorithm are that by selecting a suitable step size it is always convergent, and that it is simple to implement. Additionally, it imposes a very modest computational load, for a real-time implementation, of only $2N$ multiply-adds where is N the number of filter taps. The analysis shows that, because of use of an estimated gradient vector, the LMS algorithm only converges to the optimum solution in the mean sense. Once in a 'steady-state' the coefficients fluctuate about their optimum values by a factor that is governed both by the adaptation step size and by the filter order. In a similar way, the zero-mean steady-state error fluctuates with a standard deviation determined by the same two factors. In general, the convergence of the LMS algorithm becomes slower as the step size decreases and filter length becomes larger. This makes it unsuitable for use in applications where a particularly long duration impulse response is to be modelled, for example in acoustic echo cancellation. However, the main limitation of the LMS algorithm, in the echo canceller application, is that when speech signals are present, the convergence speed is slow and several seconds are required before the filter approaches the optimum solution. A variety of other algorithms that exhibit a superior convergence with speech signals are available. Two examples are self-orthogonalising filters and the RLS algorithm.

When the LMS algorithm is used in an echo canceller, the presence of near-end speech will cause the filter coefficients to diverge during 'double-talk' and hence increase the level of echo returned to the far-end. All echo canceller designs must include some way of controlling the process, so that double-talk does not cause a rapid increase in the echo power returned to the far-end. In a similar way, the presence of near-end background sounds limits the cancellation that may be achieved when adapting using a single step size. Such algorithms are not considered in this thesis, and should be the subject of further work. However, one such algorithm that employs two filters [OCHI77] has been implemented in a real-time echo canceller [JONE98]. It was found that this technique provides robust protection against the effects of double-talk, at the expense of extra filtering.

The use of the optimum filter is also unable to remove the quantisation distortion generated at the near-end of the network. At the far-end of the network, this quantisation noise will be perceived as distorted echo. Although non-linear filtering techniques exist, it is unclear to what extent they might be able to reduce the echo residue introduced by quantisation distortion. The following chapter addresses the problem of how to remove the quantisation noise without the use of a centre-clipper, and how to improve the LMS adaptation process when the echo is corrupted by companding distortion.

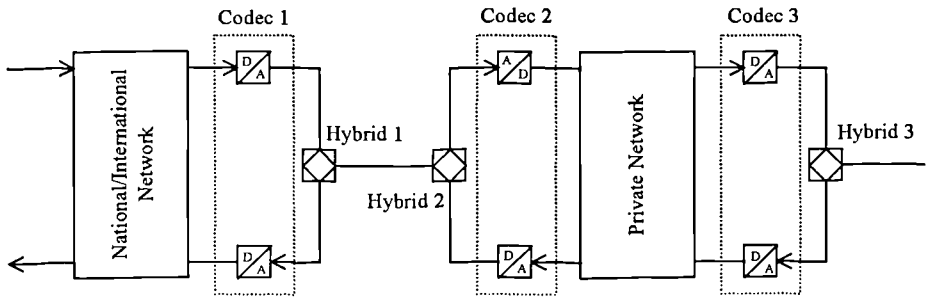
4. An Enhanced LMS Algorithm

4.1 Introduction

It was seen in the previous chapters that an echo canceller operates in an environment where the echo returned from the near-end of the network is subject to both linear and non-linear distortions. In a PSTN, an important source of non-linear distortion arises from the use of the companding that is applied to the transmitted speech signals. Because a linear adaptive filter is only able to model linear distortion, i.e. amplitude and group delay distortion, the presence of the quantisation noise/distortion has two adverse effects.

The first adverse effect is that the maximum Echo Return Loss Enhancement (ERLE) is limited to the instantaneous SQNR of the echo waveform, as described in chapter three. If the adaptive filter coefficients are identical to the impulse response of the linear part of the near-end circuit then the filter will predict the sample values of the unquantised echo exactly. This will leave only the quantisation noise of the echo as the residual. The quantisation noise power increases as the ERL decreases because the amplitude of the echo becomes larger. Hence, the companding noise, which as discussed in chapter 2 sounds speech-like to the far-end talker, will be interpreted as a strong echo. It should be noted that under certain circumstances, there may be more than one pair of PCM codecs in the echo path [ITUT96c] and this will affect the quantisation noise power and its characteristics. For a normal international call,

quantisation noise is introduced by the compander in the local exchange that connects the subscriber to the network. Now consider a call made to a private four-wire network that is interfaced to the national network using a two-wire connection as shown below:



Multiple PCM Codecs in the Echo Path

Figure 4.1

As with the previous normal international call, quantisation noise is again introduced by the codec (Codec1) in the local exchange that connects the private network to the national network. However, due to the presence of Codec 2 and Codec, quantisation noise will also be introduced into both the receive and send paths of the private network. Thus, the power of the echo quantisation noise returned to the far-end talker will be larger than for a normal international call. The work described in this chapter assumes that this situation does not arise.

The second adverse effect arising from the companding noise is that the echo canceller's adaptive filter will be misadjusted from its optimum setting, with the coefficients fluctuating during adaptation. It was shown in chapter three that the magnitude of the fluctuations is proportional to the quantisation noise power and proportional to the step-size of the adaptation algorithm. Therefore, in addition to the unwanted quantisation noise, the residual echo is also likely to contain a component that is a linearly distorted, time varying, version of the echo.

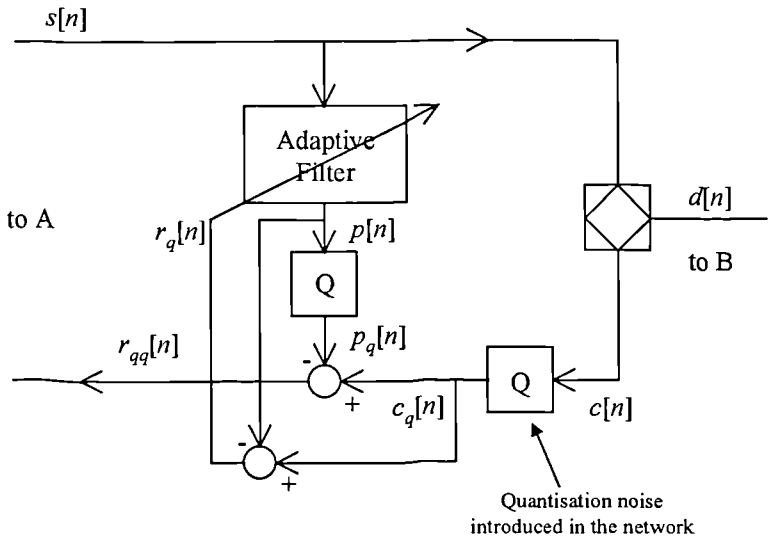
In principle the magnitude of the fluctuation of the coefficients, and hence the misadjustment, can be made arbitrarily small by reducing the adaptation step size. If

the replica echo is identical to the true echo, and an ideal compander is the only source of non-linear distortion, it may be possible to remove the echo completely by subjecting the replica to the same non-linear distortion. In a network echo canceller, the residual echo is processed using a Non-Linear Processor (NLP), usually a centre-clipper, to remove any residual echo. The centre-clipper acts as a low-level echo suppressor and therefore limits the full-duplex nature of the connection when it is operating. If the canceller could be made to remove the entire echo as described above, a non-linear processor would not be needed.

The purpose of this chapter is to introduce a technique that can be used to improve the steady-state performance of the LMS algorithm. The possibility of avoiding the need for a NLP as described above was the main reason for the investigation of this technique The enhancement is achieved by estimating the error before corruption by quantisation noise, and using this estimate in place of the usual error term in the adaptation algorithm. The aim of the algorithm is to obtain a smaller residual echo power than would normally be possible for a given step size without any penalty in convergence speed. It will be demonstrated that the performance of the adaptation algorithm has similar convergence speed to the standard technique when the adaptation constant is large, but that it has a smaller misadjustment.

4.2 Companding of the Adaptive Filter Output

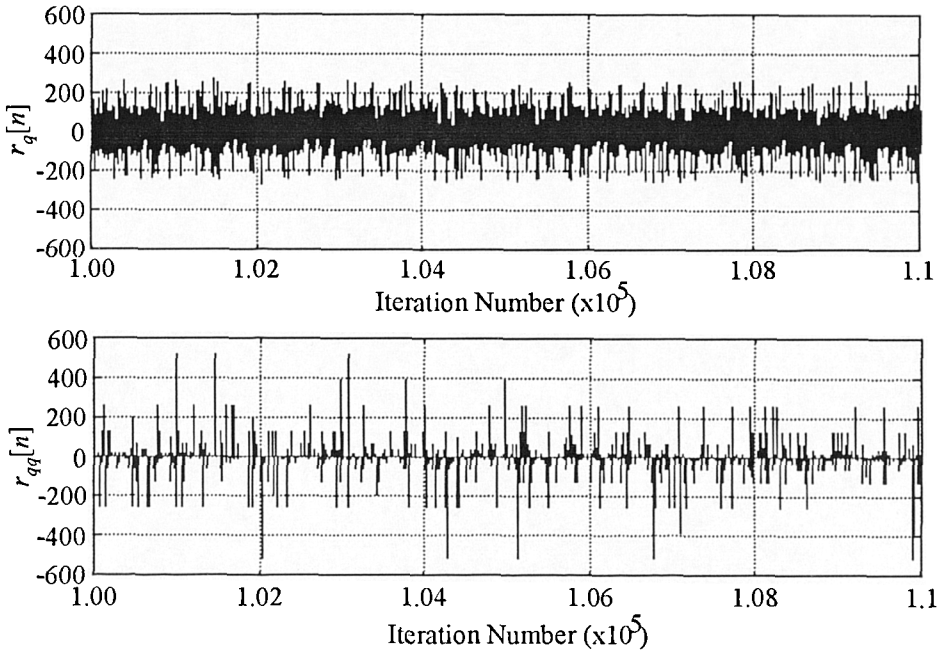
This section examines the characteristics of the echo residual when the predicted echo is subjected to the same non-linearity as found in the network. Figure 4.2 shows the block diagram of such a system, in which the non-linearity is introduced by an ideal compander.



Companding of the Filter Output to Reduce the Residual Echo

Figure 4.2

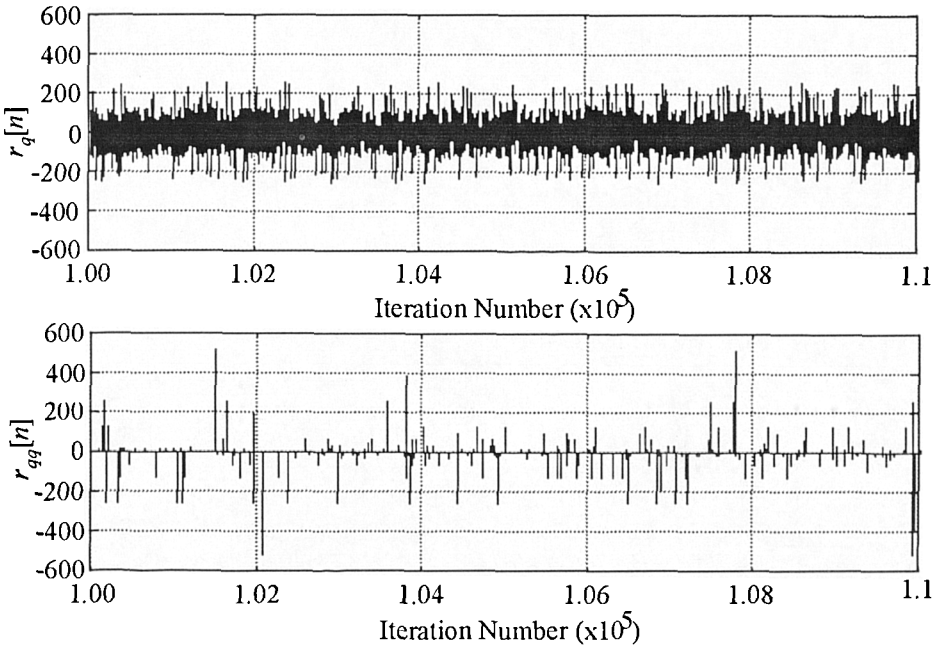
In Figure 4.2 $p_q[n]$ and $c_q[n]$ are the quantised versions of the filter output and echo respectively. There are two error terms calculated by this system: $r_q[n]$ is the error used in adaptation of the coefficients, and $r_{qq}[n]$ is the error that results when the filter output is quantised. Figure 4.3a and Figure 4.3b show the steady state residual echo signals, $r_q[n]$ and $r_{qq}[n]$, that are obtained when adapting a 128 tap FIR filter using the NLMS algorithm with a step size of 0.01. The reference waveform for this test was white Gaussian noise with a standard deviation of 8000.



a) without filter companding, b) with filter companding, μ=0.01

Figure 4.3

Figure 4.4 is similar to Figure 4.3 except that the step size has been reduced to 0.001.



a) without filter companding, b) with filter companding, μ=0.001

Figure 4.4

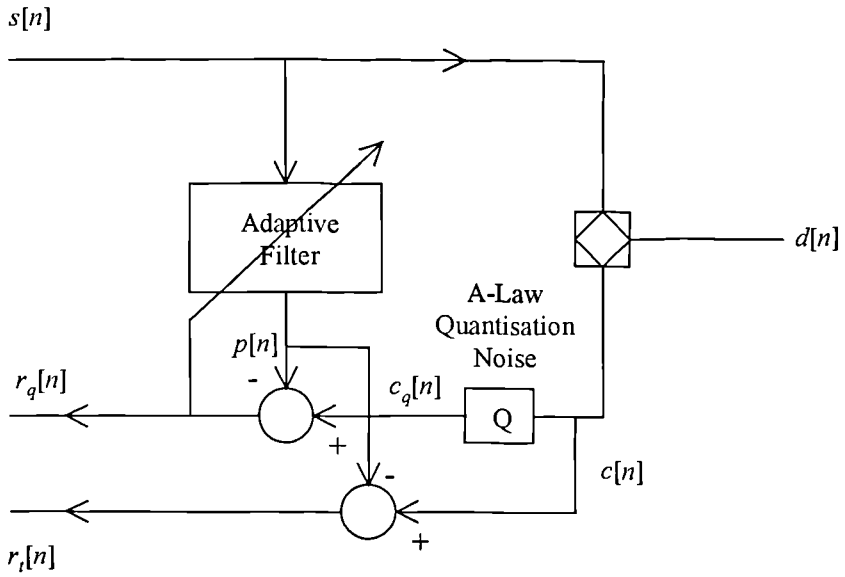
These figures show that companding the filter output has only resulted in partial elimination of the quantisation noise from $r_{qq}[n]$. Although the residual is removed at many instants of time, $r_{qq}[n]$ has become more 'impulsive' than $r_q[n]$, because of the remaining occasional quantisation level mismatches between the filter output $p[n]$ and the echo $c[n]$. In other words, when $p[n]$ lies in the same quantisation level as $c[n]$, the quantisation noise is indeed removed from $r_{qq}[n]$. However, due to fluctuations of the adaptive coefficients, the filter output and actual echo will sometimes fall in different quantisation intervals. In this case, the filter output $p[n]$ and actual echo $c[n]$ will be quantised to different levels and this causes a spike to appear in the error $r_{qq}[n]$. The amplitude of the spikes is limited to those amplitudes permitted by the level spacing of the quantisation process. Thus if a mismatch occurs at large amplitudes the resulting spike will be larger than for a mismatch at small amplitudes. Figure 4.4 shows that when a smaller step size is used the frequency and amplitude of the spikes decreases, because the coefficient fluctuation is smaller and hence erroneous prediction of the correct quantisation level occurs less frequently. If the residual could be reduced to such a train of isolated spikes then they could be easily identified and removed, without affecting near-end speech transmission.

If the filter coefficients were less misadjusted, then subjecting the predicted echo to the same non-linearity as found in the network would be more successful at the removal of the quantisation noise. Decreasing the step size may reduce the misadjustment, but this could result in long convergence times. For most practical step sizes, the misadjustment will be too large to allow the method to work successfully. One reason why the coefficients are misadjusted is that the quantisation noise is corrupting the LMS correction process. This suggests that one way to reduce the misadjustment might be to reduce the effect of the quantisation noise. Any technique that may be used to increase the prediction accuracy should not sacrifice convergence speed, or require a substantial increase of processing power.

4.3 Improving the Accuracy of the Predicted Echo

The technique described in this chapter attempts to improve the accuracy of the predicted echo by adapting the filter using an estimate of the 'true error'. Figure 4.5

shows an echo canceller system that explains the terminology to be used here and, in particular, what is meant by 'true' error.



The 'true' error and 'canceller' error in an idealised echo canceller

Figure 4.5

In Figure 4.5 the 'true error' and 'canceller error' are defined by:

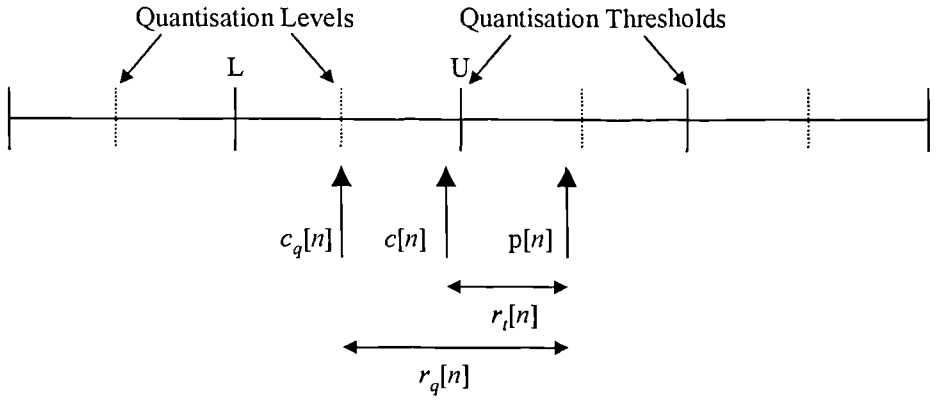
$$\text{True Error:} \qquad r_t[n] = c[n] - p[n] \tag{4.1}$$

$$\text{Canceller Error:} \qquad r_q[n] = c_q[n] - p[n] = r_t[n] + q[n] \tag{4.2}$$

Normally, the canceller error $r_q[n]$, which is a sum of the true error and quantisation noise is used as the error term in the LMS algorithm. However, the presence of quantisation noise $q[n]$ causes random fluctuations in the filter coefficients and therefore limits the accuracy of $p[n]$ for a given step size. If the true error $r_t[n]$ could be used, the steady-state fluctuations would be smaller, and therefore $p[n]$ would be more accurate. Unfortunately, in a real canceller system this quantity is unknown and hence cannot be used. It is therefore proposed here that, rather than using the very noisy error value $r_q[n]$ for tap adaptation, a less noisy estimate of true error should be used. This estimate is chosen such that for all events, with the same measured values of $c_q[n]$ and p[n], half will have a larger true error whilst half will have a smaller true

error. This median estimate is used in preference to an estimate based on the mean or other possible statistics because it is readily calculated and has modest requirements in relation to processor power. To find the median estimate of true error for each sample time it is necessary to know the probability density function (PDF) of the true error. It is to be expected that some improvement of the adaptation process would be obtained using even a poor estimate of the PDF but, of course, better performance would be obtained by using a better approximation.

Consider Figure 4.6, which shows several quantisation levels and thresholds (from segment seven of the A-law for example), and the values of the quantised echo and predicted echo at some instant in time.



The quantisation levels and thresholds

Figure 4.6

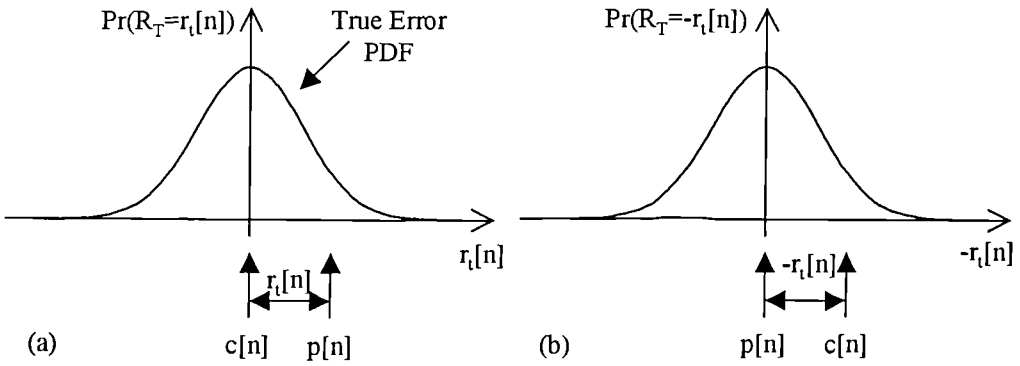
Due to the quantisation process, the unknown true echo $c[n]$ must lie within the same quantisation region as $c_q[n]$. In other words, $c[n]$ must lie between the lower and upper quantisation thresholds, denoted by L and U in the diagram. Normally the value at $c_q[n]$ is used in forming the error term. However, for the situation illustrated here, $c[n]$ is likely to be closer to U than L because $p[n]$ is greater than U. Similarly, $c[n]$ would be more likely to be closer to L than U if $p[n]$ was less than L. This is because p[n] is attempting to predict $c[n]$, which is the echo before corruption with quantisation noise, and $c[n]$ is more likely to be close to $p[n]$ than distant from it. If the filter has been adapted such that the error of $p[n]$ is very small then it would be better to use the distance between $p[n]$ and the nearer threshold, L or U, as the

measure of error in preference to the distance between $p[n]$ and the quantisation level $c_q[n]$. However, if the PDF of the true error is known then a better estimate of true echo value can be found than either the quantisation level or the quantisation threshold. If the estimated echo is denoted by $c_g[n]$ then the estimated true error, $r_g[n]$ is given by:

$$r_g[n] = c_g[n] - p[n] \tag{4.3}$$

A hypothesised form of the PDF of the true error, and the relationship between $c[n]$, $p[n]$ and $r_t[n]$ is shown in Figure 4.7a. The PDF of the true error is assumed to be symmetrical and to have zero mean. It is also assumed that there is highest probability of the true error $r_t[n]$ lying in the value range $v....v+\delta v$ when v is close to zero and that the probability of $r_t[n]$ lying in this range decreases as the magnitude of v becomes larger. It is also assumed that when the standard deviation of the true error is constant the PDF is independent of the values of $p[n]$ and $c[n]$.

Figure 4.7a shows the error distribution for a particular value of $c[n]$ which might be used to find the probability of $r_t[n]$ lying in the range $v....v+\delta v$.



True error distributions for a a) given $c[n]$, b) given $p[n]$

Figure 4.7

Unfortunately since $c[n]$ is unknown in a real network this form of the distribution is of little use. However, centring the error distribution on $p[n]$, as in Figure 4.7b which shows the distribution for a particular value of $p[n]$, is equivalent to centring it on $c[n]$ when calculating the probability that $r_t[n]$ lies within the range $v....v+\delta v$ since it

is assumed that the PDF is symmetric and independent of $c[n]$ and $p[n]$ for constant true error standard deviation. The situation of Figure 4.7b is of interest in the echo canceller since it allows an estimate of the true echo value to be found, which is to be used in place of $c_q[n]$ for adaptation equations. It will be demonstrated that using this estimated value in the formation of the error term can yield significant improvements over the standard algorithm.

However, before calculation of the median estimate is discussed two factors need to be considered. Firstly, what is the distribution of the true error, and secondly, how may its standard deviation be calculated? These issues are addressed in the following sections.

4.4 The Distribution of the True Error

In an echo canceller whose coefficients are adjusted by the LMS algorithm or one of its derivatives, it is likely that the true error will be normally distributed, even if the reference, $s[n]$, and echo, $c[n]$, are not. The true error, $r_t[n]$, is the difference between the unquantised echo $c[n]$ and filter output $p[n]$, as given by equation (4.4).

$$r_t[n] = c[n] - \sum_{i=0}^{N-1} w_n[i]s[n-i] \qquad (4.4)$$

Each adaptive coefficient, $w_n[i]$, may be written as the sum of two components:

$$w_n[i] = a[i] + b_n[i] \qquad (4.5)$$

where $a[i]$ is the optimum coefficient for tap i (i.e. its mean value), and,

$b_n[i]$ is the random misadjustment at tap i (i.e. fluctuations about the mean)

Initially, when the filter is untrained with $w_0[i]=0$ for 0<i<N-1, we have $a[i]=-b_0[i]$ for all i. Now suppose that the adaptive filter has been trained, has reached its 'steady-state' and that there is zero misadjustment. The true error is zero and thus:

$$r_t[n] = c[n] - \sum_{i-0}^{N-1} a[i]s[n-i] = 0 \qquad (4.6)$$

because $b_n[i]$ is zero for all i. In other words

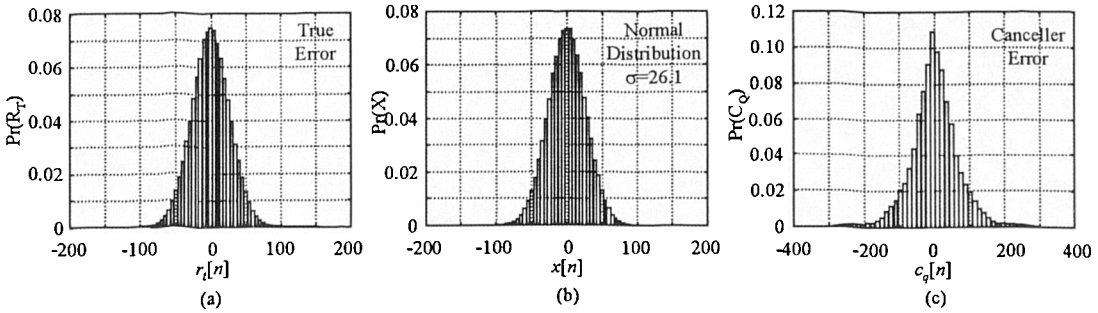
$$\sum_{i-0}^{N-1} a[i]s[n-i] = c[n] \qquad (4.7)$$

However, when the coefficients are being adapted using the LMS algorithm the steady state misadjustment will be non zero, and thus the true error is now given by:

$$r_t[n] = \sum_{i=0}^{N-1} - b_n[i]s[n-i] \qquad (4.8)$$

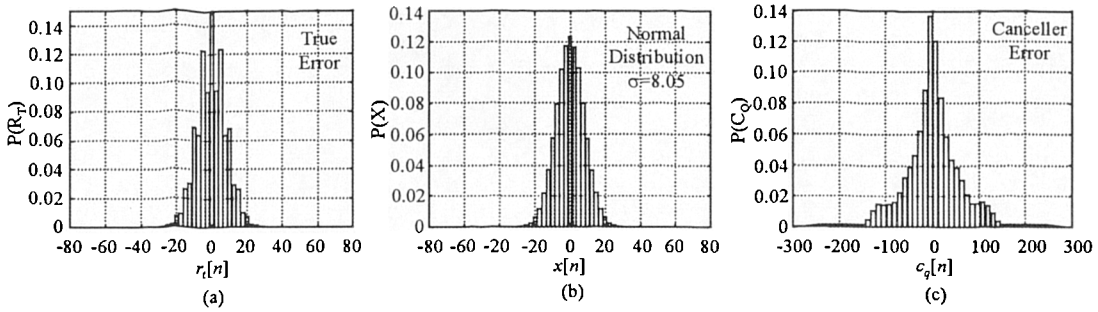
The most common form of the Central Limit Theorem [CHAT83] states that, when samples of size P are taken from a population of independent random variables, the sample means will form a distribution that tends toward Gaussian as P increases, regardless of the distribution of the original population. Equivalently, samples consisting of the sum of the P random variables will tend to Gaussian as P increases. In this application $r_t[n]$ may be viewed as one sample which is a weighted sum of random variables. For these samples, the random variables are the coefficient errors $b_n[i]$ and the weighting factors are the sample values $s[n-i]$. The ensemble distribution of $r_t[n]$ is then expected to tend toward Gaussian when N is large, as here. The ensemble is considered to be formed by using the same, or similar signal sequence $s[n]$ but different training instances having different sets of coefficient errors $b_n[i]$. The expectation that $r_t[n]$ will be Gaussian is enhanced by the observation that at each sample time, n, and each training instance, the $b_n[i]$ are themselves found to have a Gaussian-like distribution.

The distributions of the canceller error $r_q[n]$ and true error $r_t[n]$ when the adaptation process has converged are shown below, and are compared with a normal distribution that has the same standard deviation as the true error. Figure 4.8 shows the distributions for μ=0.15 and Figure 4.9 shows the distributions obtained when μ=0.015. Note that for both these figure, the far-end reference waveform was white, uniformly distributed noise.



a) True Error, b) Gaussian Distribution, c) Canceller error with μ=0.15

Figure 4.8



a) True Error, b) Gaussian Distribution, c) Canceller error with μ=0.015

Figure 4.9

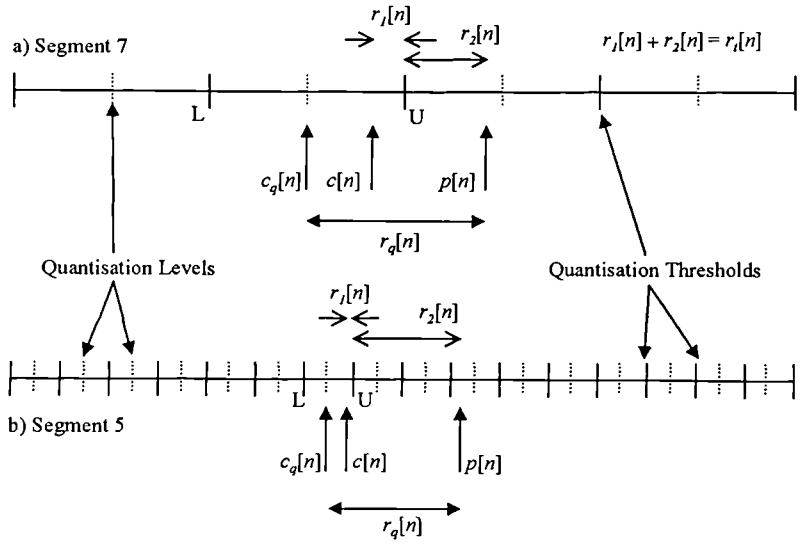
As the step size becomes smaller, the 'width' of both errors decreases. However, the distribution of the true error remains of Gaussian appearance whilst that of the canceller error tends towards the distribution of the echo quantisation noise. This is to be expected, because if the filter coefficients are identical to their optimum values, the canceller error is identical to the echo quantisation noise.

These figures show that true error appears to be normally distributed. This fitting by 'eye' is sufficient for our intended application since it is expected that the algorithm performance will not be sensitive to the type of distribution used. It would be possible to investigate and compare the performance achieved with a variety of probability distributions, for example uniform or triangular etc. However, consideration of the way that $p[n]$ is generated and measurements of $p[n]$ indicate that the distribution of the true error is Gaussian and therefore this is the distribution that is investigated here.

4.5 Estimation of True Error Standard Deviation

The previous section discussed the assumption that the true error would be normally distributed. In order to use the true error distribution to estimate this error at given instants in time, its standard deviation or power must be estimated. The method of estimating the standard deviation described here relies on the observation that the predicted echo $p[n]$ often falls within a quantisation level that is different from $c_q[n]$, which is the corresponding quantised echo sample value. It is assumed that the PDF of the true error, $r_t[n]$, is such that smaller errors are more probable than large errors, and that small values of $c[n]$ are more likely than larger values. Now, when $p[n]$ and $c_q[n]$ both lie within the same quantisation level very little information is available to update an estimate of the standard deviation. However when they lie in different levels, it has been found that the distance between $p[n]$ and the nearest quantisation threshold may be used to estimate the standard deviation of $r_t[n]$.

Figure 4.10a shows a situation that might arise when $c_q[n]$ lies in segment 7 of the companding characteristic, with $p[n]$ and $c_q[n]$ in different quantisation levels so that the condition is satisfied for the power estimate to be updated. Figure 4.10b shows a similar situation in which $c_q[n]$ falls in a lower segment, in this case segment 5. In the following discussion $p[n]$ is always taken to be above U but the same reasoning would apply equally to the case of $p[n]$ being below L.



True error power estimation in quantisation segments 7 and 5

Figure 4.10

After a short period of adaptation, the standard deviation of true error will have fallen to a value that is smaller than the quantisation level spacing ($\Delta_7 = U - L$) for segment seven. This means that when $p[n]$>U, it is much more likely that $c[n]$ will be close to U rather than elsewhere within the quantisation level. The distance between $p[n]$ and U ($r_2[n]$), is approximately equal to the true error and it is therefore better to use $r_2[n]$ to update the standard deviation estimate rather than $r_q[n]$. In effect, when $c[n]$ is close to a threshold and when many similar samples are considered, $r_2[n]$ is drawn from the full distribution of all possible true error values. As $c[n]$ moves away from U towards the quantisation level, $r_2[n]$ will come only from the tail of the error distribution and will therefore occur less frequently. It is clear that as the true error power falls during adaptation, the standard deviation of the $r_q[n]$'s attributed to segment 7 tends towards $\Delta_7/2$, and it would become less and less accurate to use $r_q[n]$ for updating the standard deviation estimate. However, if $r_2[n]$ is used, rather than $r_q[n]$, then although this leads to an underestimation of the true error, the magnitude of the error that is incurred is acceptably small. When the distribution of true error is Gaussian, numerical investigation of the use of $r_2[n]$ for standard deviation calculations indicates that even if all updates arise from segments where the standard deviation is much lower than the quantisation interval the estimated standard deviation is no more than 20% lower than the true value.

Figure 4.10b shows the situation that might arise for segment 5, at the same stage of adaptation as the segment 7 case described above. Here, the quantisation spacing is smaller than for segment 7 and the standard deviation of the true error is likely to be similar to the quantisation interval. Thus, the likelihood of the standard deviation being larger than the quantisation spacing increases as the magnitude of c[n] becomes smaller. Now, $r_2[n]$ will be drawn from the full distribution of errors regardless of whether or not $c[n]$ is close to U, $r_2[n]$ will be approximately equal to the true error $r_t[n]$ and hence the resulting standard deviation estimate will be close to the true value.

Because the PDF of speech has its highest density at low amplitudes, most updates of the true error power estimate will result from $c_q[n]$ values that are in the lower quantisation segments and hence will give good accuracy. Moreover, most updates will occur in the lower segments because, for these segments, a higher proportion of samples will have $p[n]$ and $c_q[n]$ in different quantisation levels which is the condition that must be satisfied if an update is to be possible. It is of course also necessary that there is averaging of the error measurement over a sufficiently large number of samples.

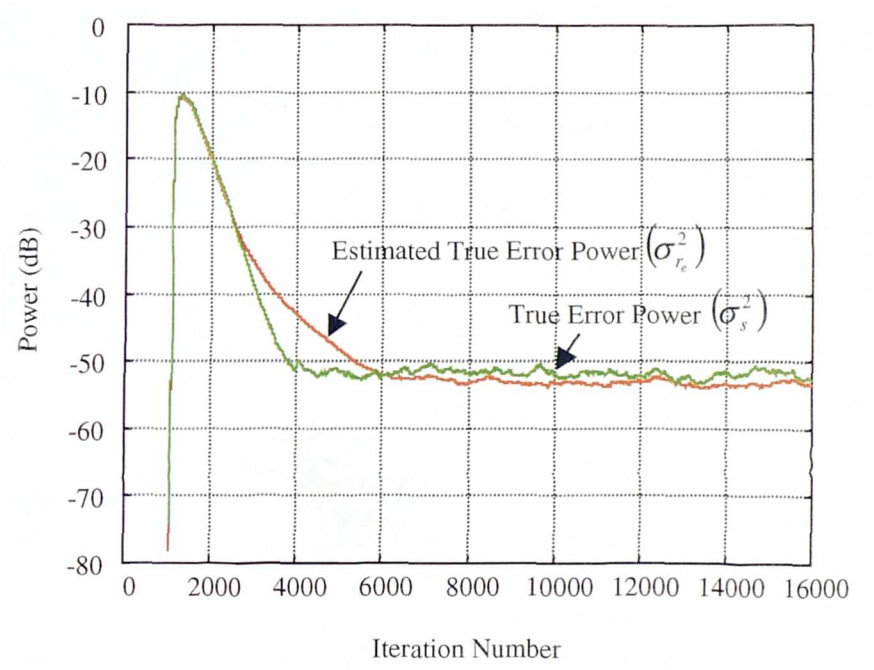
Thus, whenever $p[n]$ lies in a different quantisation level to $c[n]$, the square of $r_2[n]$ may be used as an estimate of the true error, for the purposes of estimating its standard deviation or power. Accordingly, equation (4.9) is used to calculate the true error power estimate $\sigma_{r_e}^2[n]$:

$$\sigma_{r_e}^2[n] = \begin{cases} \alpha\sigma_{r_e}^2[n-1] + (1-\alpha)(p[n]-U)^2 & \text{if } p[n] > U \\ \alpha\sigma_{r_e}^2[n-1] + (1-\alpha)(L-p[n])^2 & \text{if } p[n] < L \end{cases} \qquad (4.9)$$

where (1-α) is the factor which determines how $\sigma_{r_e}^2$ tracks the actual true error power. As α is made larger the power estimate becomes more accurate, but at the expense of tracking speed and vice-versa. (1-α) is usually very small, typically a value of 0.00625 is used in the power calculation. The performance of the estimation process is described in the next section.

4.5.1 Performance of the Estimation Process

Figure 4.11 shows how the estimated power compares with the actual true error power when adapting the filter coefficients using the standard LMS algorithm (note that the 1024 sample delay is due to the delay in the echo path). The test was carried out with the step size μ=0.15, α=0.99375, ERL=6dB and white Gaussian noise as the far-end reference signal. For clarity the powers are shown relative to the power of the far-end signal, i.e. the far-end power is 0dB.



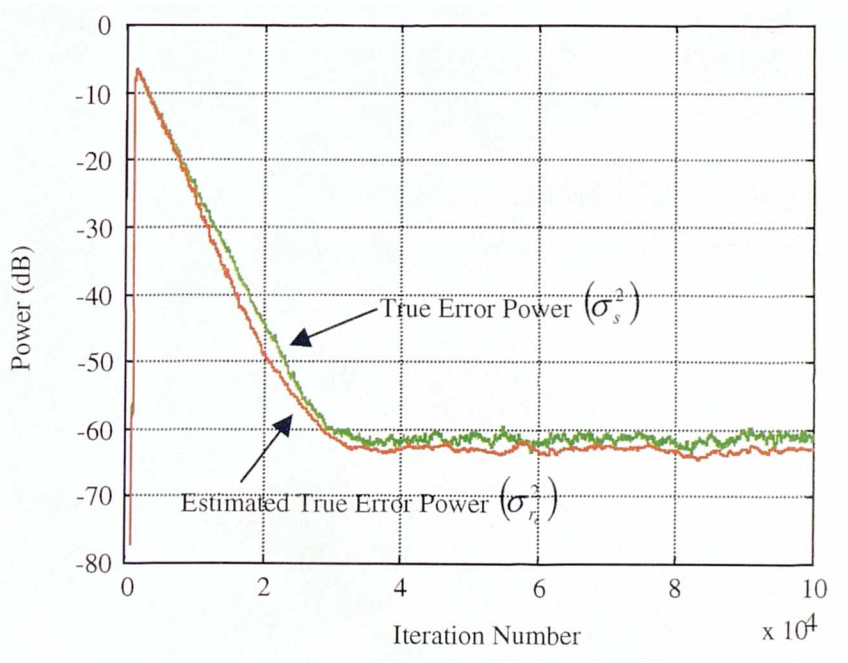
Comparison of the estimated and actual true error powers μ=0 15

Figure 4.11

During the initial training between iterations 1000 to 3000, the true error power is large. Over this time, the estimated power is approximately the same as the true error power. As adaptation proceeds, the estimated power lags behind the true value, but once in the steady state the estimate is an underestimate of the true value, as expected.

The lag is in part due to the effect of filtering and in part due to the estimate being updated less frequently as the adaptation proceeds because a smaller and smaller

proportion of $p[n]$ values lie outside the range of the quantised echo thresholds, L and U. The degree to which the estimated power lags the actual value, for a fixed power time constant α, is dependent upon the adaptation step size. Figure 4.12 below shows the learning curves when the step size is reduced from 0.15 to 0.015. Now the estimated power is always an underestimate of its true value, except during the initial stages of adaptation when, both powers are large.



Comparison of the estimated and actual true error powers μ=0.015

Figure 4.12

Once in the steady state the estimate will be updated most frequently when $c_1[n]$ lies in the lowest quantisation segments, as in the higher segments the filter will have converged sufficiently to allow most $p[n]$ values to lie inside the same quantisation level as $c_q[n]$. Once a steady state has been reached, $\sigma_{r_2}^2$ will be an underestimate of the true value because $r_2[n]$ is always less than $r_1[n]$. The following table shows the measured steady state powers for several different adaptation step sizes.

	μ=0.15	μ=0.015	μ=0.0015
Canceller Error Power	-43.55dB	-44dB	-44.25dB
True Error Power	-52dB	-62dB	-71dB
Estimated True Error Power	-54dB	-64dB	-73dB

Table 4.1

For step size values likely to be used in practice, the steady-state underestimation has been found to be approximately 2dB when estimating the true error produced using $r_2[n]$ in the estimation equation.

4.6 Calculation of The True Error Estimate

Now that the characteristics of the true error distribution and the estimation of its standard deviation have been discussed, the calculation of the median error estimate can be examined.

One way to compute the median value would be to generate a new Gaussian PDF with standard deviation σ_{r_e} for every iteration, find the total area between L and U, the median area value and hence the error estimate. However, this is not very efficient, especially for a real-time implementation. Here, the area is calculated using a look-up table that represents the normalised Cumulative Density Function (CDF).

4.6.1 The Gaussian Cumulative Distribution Function (CDF)

The median error estimate can be calculated more efficiently, by using a look-up table containing values of a Gaussian CDF with unity variance. The table entries are generated using a modified version of the error function, which is given by [PRES92]:

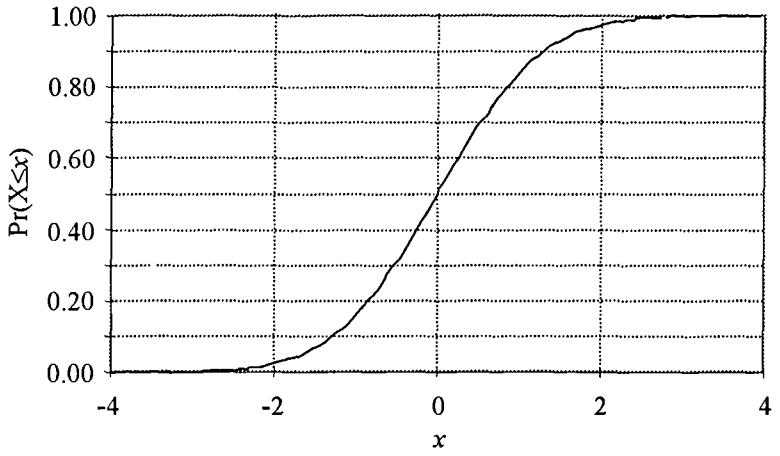
$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \qquad (4.10)$$

Equation (4.10) is modified to

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt \qquad (4.11)$$

in order to calculate the Gaussian CDF.

Evaluation of equation (4.11) for different values of x, yields the total probability that a zero mean, normally distributed random variable, will lie between $-\infty$ and x. A computer program to generate the error function [PRES94] was modified to calculate the Gaussian CDF by scaling with the change of variable. Figure 4.13 shows the CDF as described by equation (4.11).



The Gaussian Cumulative Density Function (CDF)

Figure 4.13

So that the table values need only be calculated once, the indices into the table are normalised by the estimated true error standard deviation. The table contains values for ±4 standard deviations since, for values greater than this, it can be assumed that the probability is either 1.0, if $x>4$, or 0.0 if $x<-4$.

4.6.2 Calculation of the Median Area and the Median Estimate

The median area position and the corresponding median estimate may be calculated very simply using the lookup table. First define,

$$x_1[n] = p[n] - L[n] \tag{4.12}$$

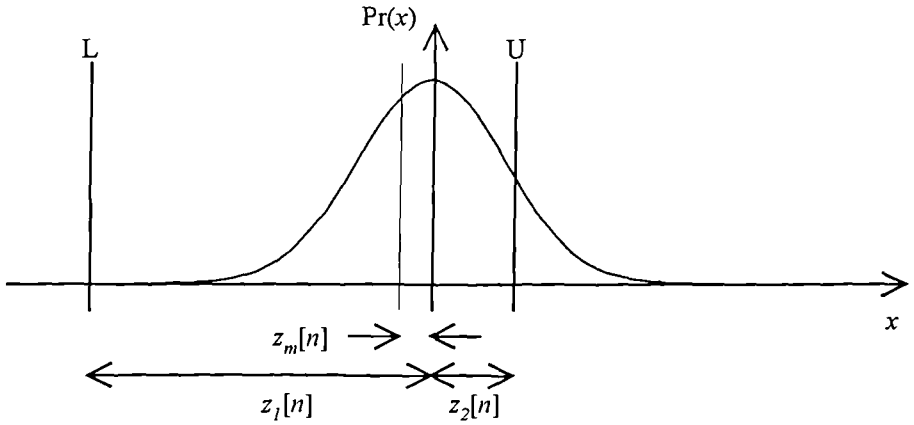
and $$x_2[n] = U[n] - p[n] \tag{4.13}$$

as the differences between $p[n]$ and the lower and upper thresholds respectively. Normalising by the standard deviation of the true error gives:

$$z_1[n] = \frac{x_1[n]}{\sigma_{r_e}[n]} \tag{4.14}$$

and $$z_2[n] = \frac{x_2[n]}{\sigma_{r_e}[n]} \tag{4.15}$$

These distances are shown in Figure 4.14 below.



Calculation of the median estimate

Figure 4.14

Now, half the total area between L and U is given by

$$AreaA = \frac{1}{2}[F(z_2[n]) - F(z_1[n])] \tag{4.16}$$

where F(x) is defined in equation (4.11) but, in practice, the values are obtained from the look-up table. Thus, the total area under the PDF between -∞ and the median position $z_m[n]$ is given by

$$\begin{aligned} F(z_m[n]) &= F(z_1[n]) + \frac{1}{2}[F(z_2[n]) - F(z_1[n])] \\ &= \frac{F(z_1[n]) + F(z_2[n])}{2} \end{aligned} \tag{4.17}$$

Hence, the median position z_m is obtained from the inverse of equation (4.17), i.e.:

$$z_m[n] = F^{-1}\left[\frac{F(z_1[n]) + F(z_2[n])}{2}\right] \tag{4.18}$$

Thus the estimated true echo $c_g[n]$ is given by

$$c_g[n] = p[n] + z_m[n]\sigma_{r_e}[n] \tag{4.19}$$

and the median estimated true error $r_g[n]$ is given by

$$r_g[n] = z_m[n]\sigma_{r_e}[n] \tag{4.20}$$

4.6.3 Special Conditions

Although the procedure described in the previous section is straightforward, three checks must be made to ensure that the values $z_1[n]$ and $z_2[n]$ do not exceed the bounds of the look-up table. These are:

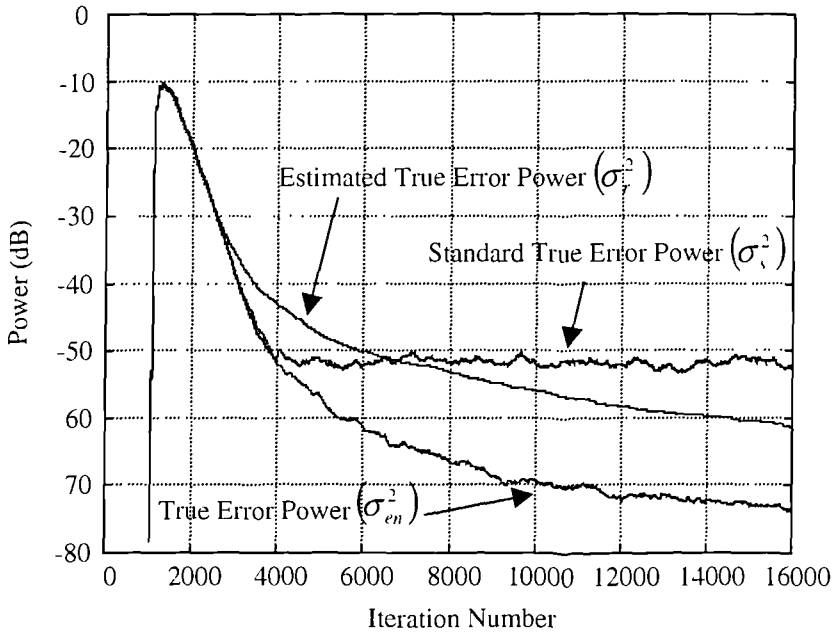
i) When $z_1[n]$ corresponds to a value in the CDF table between ±4 standard deviations, but $z_2[n]$ points to a value that is greater than +4 standard deviations, it is assumed the probability of $z[n]$ lying between -∞ and $z_2[n]$ is 1.0. Similarly, when $z_1[n]$ is less than -4 standard deviations, but $z_2[n]$ lies within the table limits, it is assumed that the probability of z lying between -∞ and $z_1[n]$ is 0.0.

ii) When $z_1[n]$ and $z_2[n]$ both correspond to values that are less than -4 standard deviations, it is assumed that the true echo is at the position given by the lower quantisation threshold. Similarly, when $z_1[n]$ and $z_2[n]$ both correspond to values that are greater than +4 standard deviations, the true echo is assumed to be at the upper quantisation threshold.

iii) When $z_1[n]$ is less than –4 standard deviations and $z_2[n]$ is greater than +4 standard deviations, it is assumed that the median area position is at the origin of the PDF, and so the estimated true error is set to zero.

4.7 Simulation Results

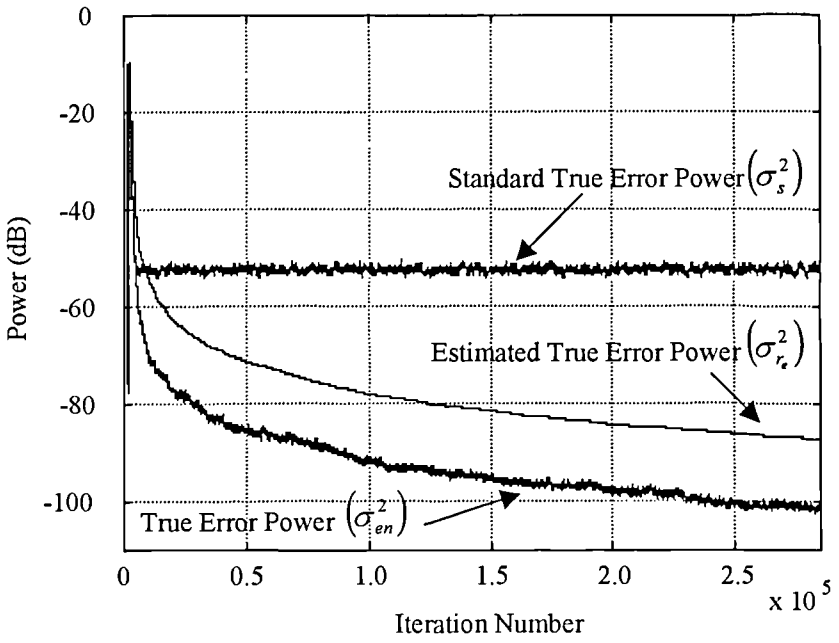
This section discusses the characteristics of the enhanced algorithm when the reference signal is white Gaussian noise. Figure 4.15 shows how the true error power, obtained when the estimated error $r_g[n]$ is used in place of $r_q[n]$, compares with the true error power obtained using the standard NLMS algorithm with a step size of 0.15. In this case, the FIR filter had 128 coefficients that were initially set to zero, and an ERL of 6dB was used.



Enhanced algorithm performance with μ=0.15

Figure 4.15

During initial adaptation, the true error power of the enhanced method (σ_{ε}^2) is approximately the same as that of the standard algorithm (σ_{ε}^2), until at iteration 4000 the quantisation noise limits σ_{ε}^2 to a steady state value of approximately 52dB However, σ_{en}^2 continues to decrease and after 16000 iterations has not reached a steady state. Figure 4.16 shows that after $2x10^5$ iterations (25 seconds) σ is approximately 50dB smaller than σ_{ε}^2, and is still decreasing.

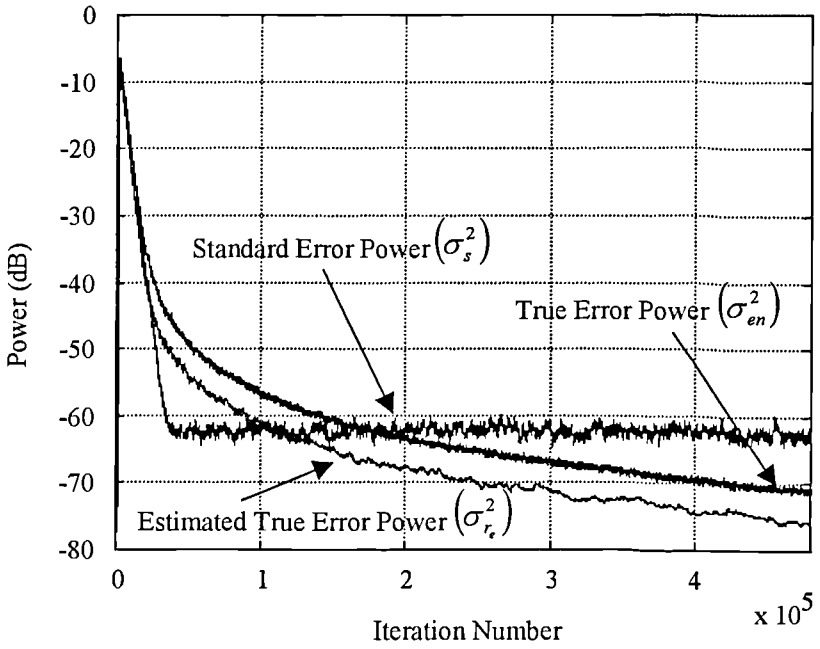


Enhanced algorithm performance with μ=0.15

Figure 4.16

As discussed in section 4.5.1 the relationship between the true error power and the estimated true error power is dependent upon the step size, because the learning rate and the misadjustment affect the power estimation algorithm. This still holds when $r_g[n]$ is used in the adaptation equation. For this particular step size, the estimated true error is an overestimate of the actual true error power.

For small step sizes giving slow adaptation, $\sigma_{r_e}^2$ is always an underestimate of σ_{en}^2. This is similar to the steady state case discussed earlier, but here the true error power is decreasing sufficiently slowly to compensate for times when the estimate is not updated. As the step size increases σ_e^2 is approximately the same as σ_{en}^2 during the initial stages of adaptation but after a number of iterations, the underestimation present with smaller step sizes disappears only to be replaced by lag/overestimation. For very large step sizes $\sigma_{r_e}^2$ will always lag σ_{en}^2 because the error has reduced so quickly that the lag effect dominates. This behaviour is shown in Figure 4.17 where the step size is 0.015.



Enhanced algorithm performance with μ=0.015

Figure 4.17

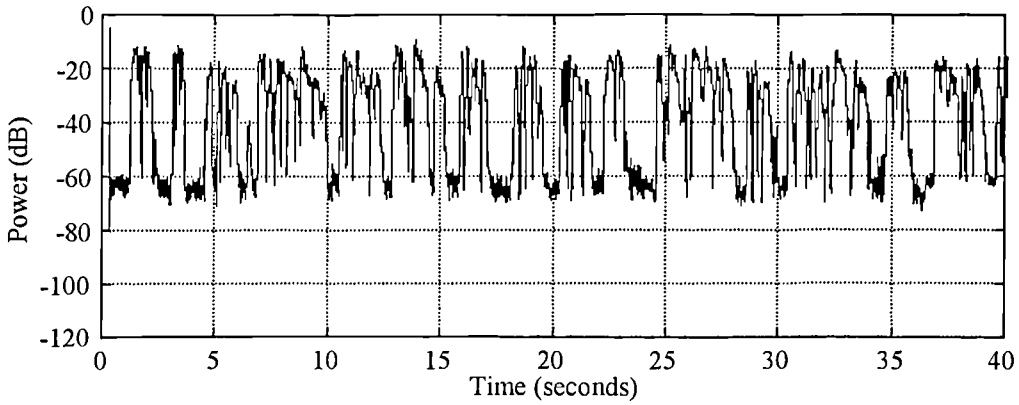
It can be seen that the largest improvement after a fixed period of time always results from using the largest step size, even though the estimated true error power used in the adaptation is too large. This is to be expected because as the step size is reduced, the misadjustment caused by the quantisation noise decreases and therefore the improvement over the standard algorithm decreases.

4.7.1 Learning Curves for Speech

This section discusses the behaviour of the algorithm when speech signals are used, rather than white Gaussian noise. As in the previous section, a 128 tap FIR filter was used whose coefficients were adapted from an initial untrained state using either the enhanced or the standard NLMS algorithms. When speech is used as the reference signal, the convergence speed is slower than when white noise is used. In order to eliminate the initial convergence period, the filter weights could have been initialised to their optimum values before commencing adaptation, and thus any resulting improvement or degradation would then be due to the characteristics of the enhanced

algorithm only. However, the algorithm must work from an initially untrained state and thus the results presented here are for this case.

When speech is used rather than random noise, learning curves of the form shown in the previous sections are slightly more difficult to interpret because the power level of the speech is constantly changing. The power variation of the speech test signal that was used in the simulations is shown in Figure 4.18.



The power variation of a typical speech waveform

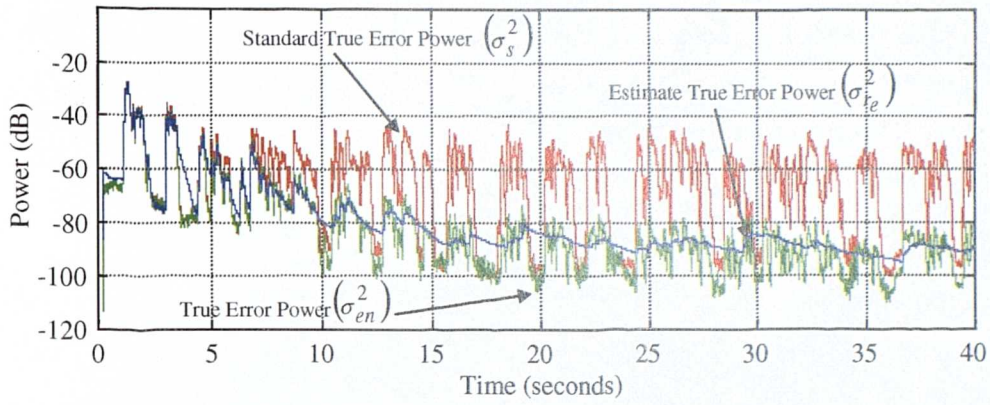
Figure 4.18

This figure shows that when speech is present the power level varies between approximately –65 and –10dB and that during the 'silence' periods between words, the power level has a mean value of approximately –65dB.

The mean square error performance of the enhanced technique could have been compared to that of the standard algorithm by calculating an improvement factor given by the ratio of true error powers obtained using the two techniques. However, this improvement factor will vary depending upon the power of the speech reference, and therefore it is beneficial to present the results in the same format as for the white noise simulations.

Figure 4.19 how the true error power, obtained when the estimated error $r_g[n]$ is used in place of $r_q[n]$, compares with the true error power obtained using the standard

NLMS algorithm, when speech is used as the reference. For this test a step size of 0.5 was used with an ERL of 6dB.



The true error power for the standard and enhanced methods, μ=0.5.

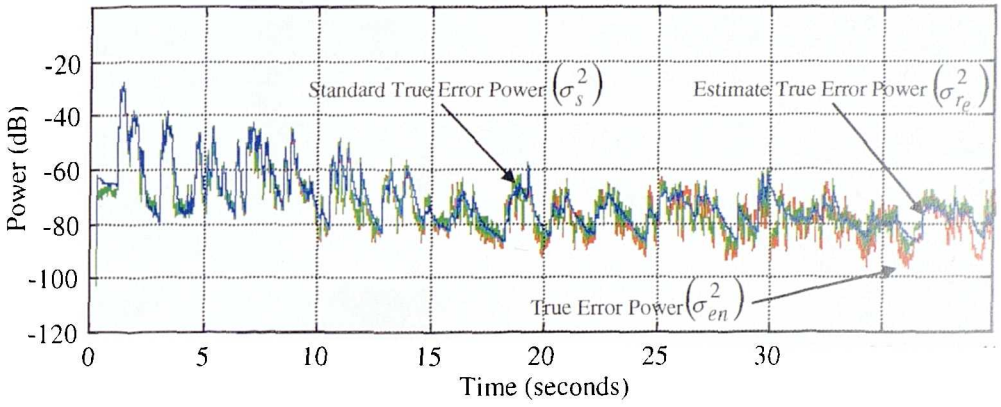
Figure 4.19

For this particular step size, the enhanced technique has a significant performance advantage over the standard NLMS algorithm when speech is used as the reference. During initial convergence, the true error powers of the enhanced and standard algorithms are approximately the same. However, after approximately 5 seconds, the true error power obtained using the standard NLMS algorithm, σ_s^2, reaches a steady state, whilst the true error power of the enhanced technique, σ_{en}^2, continues to decrease. It can also be seen that the magnitude of the improvement depends upon the power of the reference process. Figure 4.19 shows that the improvement is likely to be large when the power of the speech is also large, and small when the speech power is small. For a given number of iterations, the improvement obtained when speech is present is likely to be less than if white noise was used.

The maximum improvement in this 40 second segment is approximately 30dB. However, if adaptation was allowed to continue this figure would increase. In fact, after 1.5 minutes over 40dB of improvement has been observed when using speech as the reference signal.

Figure 4.19 also shows the behaviour of the power estimation algorithm with speech signals. It shows that when the error is large, the estimated power is approximately the same as the actual true error power. As the true error power decreases, σ^2_{en} is updated less frequently and this causes σ^2_{en} to track the peak level of the actual true error power. As expected σ^2_{en} is an underestimate of the true value.

Figure 4.20 shows a set of learning curves obtained under the same conditions as before, but using a step size of μ=0.05 that is ten times smaller than previously



The true error power for the standard and enhanced methods μ=0 (

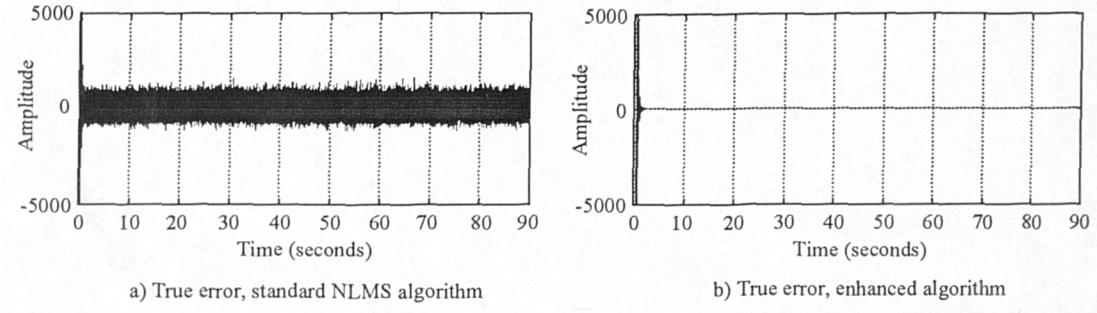
Figure 4.20

sing this step size, the performance of the enhanced technique is
fferent to that of the standard NLMS algorithm. This is because t ne
p size results in less steady state misadjustment of the adaptive il
naller step size is used in Figure 4.20, the steady state power obt
LMS method is approximately 20dB larger than that of the enhanced
larger step size. This suggests that to gain the full benefit of the enhan
e step size should be chosen to be as large as possible, within the limits of

4.7.2 Companding of the predicted echo

This section describes the characteristics of the residual error, $r_{qq}[n]$, and the true error, $r_t[n]$, that is obtained when the predicted echo, $p[n]$, is subject to the same non-linearity as found in the network. As in the previous sections, the non-linearity used here is that of an ideal A-Law compander. A 128 tap FIR filter was used, whose weights were adapted using a step size of 0.5.

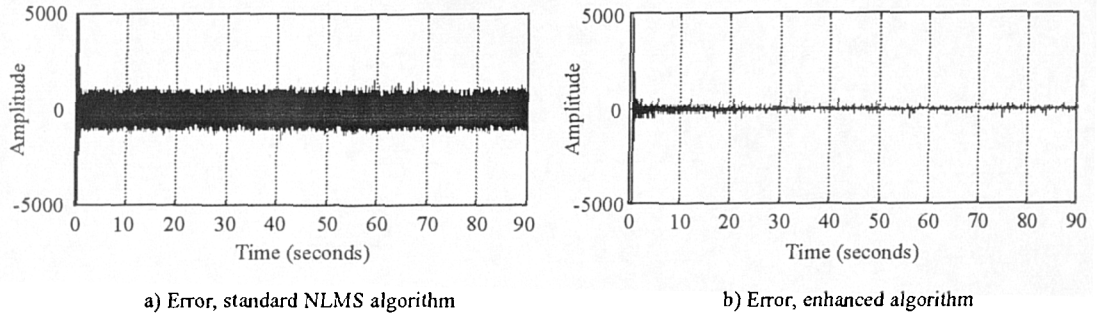
Figure 4.21 and Figure 4.22 show the effects of companding the predicted echo when the reference signal is white Gaussian noise.



a) True error, standard NLMS algorithm

b) True error, enhanced algorithm

Figure 4.21



a) Error, standard NLMS algorithm

b) Error, enhanced algorithm

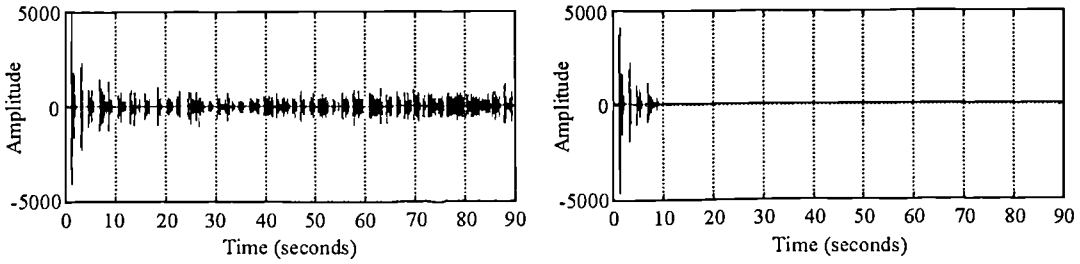
The residual error obtained by companding the predicted echo (for noise)

Figure 4.22

As noted before, the true error is significantly smaller when using the enhanced algorithm to adapt the filter coefficients - in this case by over 60dB after 90 seconds of adaptation. Figure 4.22 shows that when the predicted echo is companded, the error obtained for the NLMS algorithm does not change significantly, whilst that for the enhanced method is smaller by approximately 40dB. It can be seen that even with 40dB of improvement over the standard NLMS algorithm, the residual echo has not

been removed completely because quantisation level mismatches still occur. The residual consists of isolated 'spikes' whose amplitudes are determined by the allowed levels of the quantisation process.

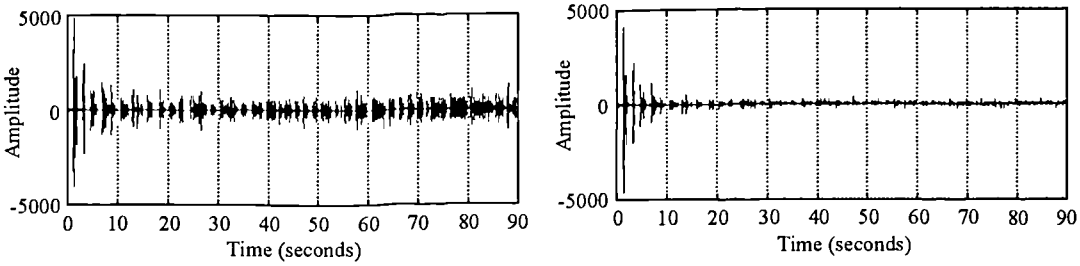
Figure 4.23 and Figure 4.24 show the effects of companding the predicted echo when the reference waveform is speech and a step size of μ=0.5 is used.



a) True error, standard NLMS algorithm

b) True error, enhanced algorithm

Figure 4.23



a) Error, standard NLMS algorithm

B) Error, enhanced algorithm

The residual error obtained by companding the predicted echo (for speech)

Figure 4.24

Compared to the standard NLMS algorithm, both error measures are smaller when using the enhanced technique, by approximately 50dB and 30dB respectively. Although companding of the predicted echo $c[n]$ reduces the power of the residual error, Figure 4.24b shows that quantisation level mismatches still occur but not with the same frequency as in Figure 4.24a. From this result, it might be expected that the error, which consists of a series of isolated 'spikes' whose amplitudes are limited to those permitted by the quantisation process, will still be perceived as distorted echo by the far-end subscriber. However, it has been found that when listening to the far-end speech and the residual error simultaneously, the 'crackles' do not have the same

characteristics as those of the quantisation noise alone. Although the crackles are only present during periods of echo, they appear to be independent of the echo amplitude and therefore they may not be perceived as distorted echo.

4.8 Algorithm Limitations

In the previous sections, it was assumed that the echo was corrupted by only the noise introduced by an ideal compander. In practice, this situation will seldom arise since the echo will be also corrupted by environmental noise from the near-end of the network and by noise from other sources. Examples of the background noise could be traffic noise from outside a telephone kiosk or the background sounds from an office.

To show the effects of the near-end background noise on the adaptation process, white noise was added to the echo before the compander in the send path (i.e. $d[n]\neq 0$). Simulations were then performed using a white Gaussian reference signal, a step size of 0.15 and an ERL of 6dB. After 60 seconds, the adaptation was frozen and the true error power obtained using the two algorithms was measured. Table 4.2 shows how the improvement factor, defined as the ratio of these two powers, varies as the near-end noise changes.

Echo Power (dB)	Quantisation Power (dB)	Noise Power (dB)	Quantisation to Interference Ratio (dB)	Improvement (dB)
-6	-44	$-\infty$	∞	53
-6	-44	-86	42	35
-6	-44	-76	32	25
-6	-44	-66	22	15
-6	-44	-56	12	5

Table 4.2

The table shows that the performance is dramatically affected by the presence of near-end noise at small power levels relative to the quantisation noise power. The reduction of the improvement factor arises because the additive noise corrupts both the true error power and the true error sample estimation processes.

As the noise power increases, the average difference between $c_q[n]$ and $c[n]$ increases because $c_q[n]$ is the quantised version of the actual echo plus additive noise, i.e. $c_q[n] = Q(c[n] + b[n])$ (where Q(•) means quantised version of). Thus, it is likely that $c[n]$ and $c_q[n]$ will lie in different quantisation levels more often. When $c[n]$ falls in a quantisation segment with a wide spacing, it is more likely that $c[n]$ will be adjacent to $c_q[n]$, whereas, if $c[n]$ lies in a segment with a small level spacing, the difference between $c_q[n]$ and $c[n]$ is more likely to be greater than the level spacing. The assumption that the distance from U to $p[n]$ (or $p[n]$ to L) is a good estimate of the true error becomes less valid as the power of the interfering noise increases, and thus the true error power estimate become less accurate. The estimation of the true error samples is also corrupted by a similar mechanism. As the noise power increases, the quantisation thresholds of $c_q[n]$, L and U, no longer correspond to the quantisation thresholds of $c[n]$ alone, and therefore the estimated true error is forced to lie within the wrong quantisation level.

As a result of the corruption of the two estimation processes that occurs when the near-end noise power increases, the performance obtained by using $r_g[n]$ in the adaptation equation will be no better than that obtained when using $r_q[n]$. It is difficult to see how the adaptation algorithm could be modified to reduce the effects of the interfering noise. In particular, it will be more difficult to improve the accuracy of the power estimation process than to improve the estimation of the true error samples. If the standard deviation of the true error were known then it should always be possible to use the resulting error distribution to obtain a more accurate true error sample estimate. One possible way in which this could be achieved is to centre the error distribution on $p[n]$ as before, but to use the median position between $c_q[n]$ and p[n] as the estimate of $c[n]$. However, if the standard deviation estimate is too large

then the use of the median estimate cannot be expected to yield a less noisy error term.

4.9 Conclusions

This chapter has discussed how the NLMS adaptation process is affected by the quantisation noise that is applied to the transmitted speech signals. Echo cancellers that use linear adaptive filters are unable to remove this quantisation noise, even if the filter weights are set to their optimum values. If the quantisation noise is transmitted through the network, it is likely to be perceived by the far-end subscriber as distorted echo and will thus make the connection more difficult to use. Most network echo cancellers use a non-linear processor, usually a centre-clipper, in order to remove this quantisation noise. In principle, the quantisation noise could be removed from the residual echo by introducing non-linearities, similar to those present in the echo path, into the adaptive filter output. When a fixed step size is used in the adaptation process, it must be large in order to give short convergence times but this gives a minimum error power that is too large to enable this technique to work.

It has been demonstrated that, under certain conditions, it is possible to estimate the value of each echo sample before it has been corrupted by the quantisation noise. This estimated echo sample may then be used to form a new error term, which when used in the NLMS adaptation equation with large step sizes, gives a similar convergence speed but results in a significantly smaller error power for both noise and speech waveforms. As the step size is reduced, the convergence speed decreases and the algorithm shows less of an improvement in 'steady-state' error because the misadjustment of the standard NLMS technique is reduced.

By companding the adaptive filter output in conjunction with using the enhanced adaptation technique, it is possible to remove the uncancellable quantisation noise to some extent. However, even for very large improvements, it appears that it is not possible to eliminate the residual echo completely. Moreover, it has been found that

the performance of the enhanced algorithm is considerably reduced when background sounds from the near-end of the network are present. The estimate of the true error power becomes less accurate as the background noise power increases because the assumptions that are used in this process become invalid. At this stage, it is difficult to see how the power estimation process can be modified to make it more robust in the presence of noise. Although the background noise level of some calls may have a sufficiently small power to enable the misadjustment to be reduced by using the new algorithm, it is doubtful whether this will lead to any significant improvement when the adaptive filter output is quantised. Thus, any further work relating to echo cancellers should assume that near-end noise is always present in addition to the quantisation noise.

This result means that other methods for maximising convergence time and reducing misadjustment, such as variable step size or other algorithms are likely to be more successful in achieving the smallest possible convergence time and minimum error power. However, it seems unlikely that these techniques will yield a steady state error that is small enough to enable complete eliminate of the residual echo by using the previously described techniques. In addition, when background sounds are present, companding the filter output is unlikely to be successful at removing the quantisation noise in the residual. This is because the quantisation noise introduced into the filter output will no longer correspond to that of the echo alone because the quantisation process does not obey superposition.

5. Residual Echo Control

5.1 Introduction

As was seen in the previous chapters, uncancelled echo may arise from the use of companding within the network. The maximum echo attenuation, or Echo Return Loss Enhancement (ERLE), for echo cancellers using linear adaptive filters is largely determined by the companding process. For connections that have a low Echo Return Loss (ERL), a large component of the uncancelled echo is quantisation distortion which if not removed, may be disturbing to the far-end talker. As described in chapter three, the uncancelled echo due to the companding process will still be present in the residual error even if the filter coefficients are perfectly adapted.

Uncancelled echo may also arise from sources other than the quantisation distortion. For example, echo cancellers that use FIR filters to generate the echo replica are unable to exactly model the echo because of the infinite duration of the echo path impulse response, although the filter length is usually chosen to minimise this effect. Uncancelled echo can also arise because a linear filter is unable to model non-linearities within the echo path. One source of non-linearity within the echo path is the hybrid, which is often constructed using iron cored transformers and will not be exactly linear. Apart from the hybrids, non-linearity may be introduced by active circuits in a telephone handset or may be caused by a codec mismatch, i.e. the

characteristics of the compressor and expander in the codecs may not be exactly complementary.

The quality of performance of the echo canceller must also be maintained in the presence of near-end background sounds, for example when a telephone is used in a noisy office or when a mobile telephone is used in a car. The presence of background sounds and quantisation noise will cause unwanted fluctuation of the adaptive filter coefficients and therefore reduce the effectiveness of the canceller. This deterioration could be minimised by controlling the adaptive filter step size or by using a different adaption algorithm that is more robust to noise. Uncancelled echo may be masked, to some extent, by the background sounds, but the effectiveness of such masking cannot be guaranteed.

The presence of uncancelled echo due to any of the above factors is unwanted and thus further attenuation of this echo may be needed in order to maintain the quality of the connection. The purpose of this chapter is to examine techniques that may be used to attenuate uncancelled echo. Network echo cancellers that are designed to meet the standards of G.165 [ITUT94] and G.168 [ITUT97] must introduce less than 125μs of delay into the receive path and less than 1ms into the send path. This limits the scope for processing of the residual echo to remove the uncancelled echo.

Residual echo control in network echo cancellers is performed using a Non Linear Processor (NLP), which generally takes the form of a time-domain centre clipper. In this chapter, both single and twin threshold centre clippers are discussed, along with their multi-band equivalents. It is shown that although echo can be suppressed using this type of residual echo control, an undesirable effect known as 'noise modulation' or 'noise pumping' is introduced. Some echo cancellers attempt to mask noise modulation by injecting so called 'comfort noise' into the centre-clipper output when it is operating. The characteristics of noise modulation are described and the use of comfort noise is discussed. Finally, to complete the review, a frequency domain technique called residual echo shaping is also discussed. It will be seen that noise

modulation also occurs when using this method, although its characteristics are slightly different from those of centre-clipping.

Before discussing the different techniques, it is important to consider the characteristics of a 'perfect' residual echo control device. Ideally, the send path of an echo canceller should be transparent to all signal components except for those that are due to the uncancelled echo. In other words, only the near-end speech and near-end background noise should be heard at the far-end. This situation would arise if the adaptive filter were able to completely remove the echo, the quantisation noise and other distortions. This ideal suggests that in addition to assessing the attenuation of the residual echo, the effect that the echo control method has on the near-end speech and background noise should be considered.

5.2 Single Threshold Centre-Clipper

Network echo cancellers that are designed to meet the requirements of G.165 or G.168 generally employ some kind of centre-clipping in order to remove the residual echo. There are two basic types of single threshold clipper that may be used, and their transfer functions are shown in Figure 5.1.

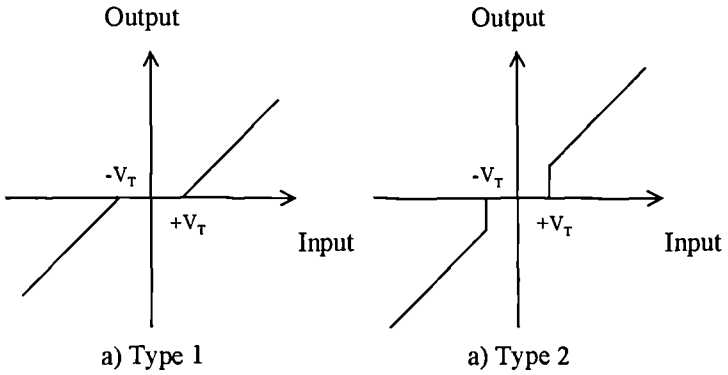


Figure 5.1

The transfer functions of the type 1 and type 2 clippers are defined by:

$$\text{Type 1 : } Output = \begin{cases} 0 & if\ |Input| \le V_T \\ Input - \text{sgn}(Input)V_T & if\ |Input| > V_T \end{cases} \qquad (5.1)$$

and

$$\text{Type 2 : Output} = \begin{cases} 0 & if\ |\text{Input}| \le V_T \\ \text{Input} & if\ |\text{Input}| > V_T \end{cases} \qquad (5.2)$$

The clipper described by equation (5.1) (type 1) was initially used for residual echo suppression in analogue echo cancellers, because it is simple to implement using analogue electronics. The second type of clipper described by (5.2) is known as a minimum distortion centre-clipper [MITC70].

Suppression of the uncancelled echo by both of these clippers is achieved by setting the threshold, V_T, so that it is larger than the peak uncancelled echo amplitude. This allows the near-end speech to pass whilst blocking the residual echo. Although near-end speech is passed, it will of course be distorted by the centre-clipping action, with the characteristics of the distortion being different for the two clippers.

5.2.1 Type 1 Clipper Distortion Characterisation

The distortion generated by both clippers may be usefully characterised by examining the harmonics that are created when a sinusoid is applied to the input of the system. Assuming that the input sinusoid has an amplitude of ±1, the Fourier coefficients of the output from a type 1 clipper are given by:

$$a_0 - b_n = 0 \qquad (5.3)$$

$$a_1 = \frac{1}{2\pi}[\sin(2A) - \sin(2B)] + \frac{2V_T}{\pi}[\sin(-B) - \sin(A)] + 1 + \frac{A}{\pi} - \frac{B}{\pi} \qquad (5.4)$$

$$a_n = \frac{1}{\pi(n+1)}[\sin((n+1)A)+\sin(-(n+1)B)]+ \quad (5.5)$$
$$\frac{1}{\pi(n-1)}[\sin((n-1)A)+\sin(-(n-1)B)]+$$
$$\frac{2V_T}{n\pi}[\sin(-nA)-\sin(-nB)]$$

where

$$A = \cos^{-1}(V_T),\ B = \cos^{-1}(-V_T) \quad (5.6)$$

V_T = the clipping threshold

Note that for this input waveform, $f(t) = -f\left(t+\frac{T_p}{2}\right)$, and hence the distorted output will only contain odd harmonics in addition to the fundamental. In addition, the output will have a smaller amplitude fundamental than the original input. The distorted output is a sum of the input waveform plus a distortion waveform, which may be calculated by subtracting the input from the output. In view of this, a distortion ratio for a sinusoidal input is defined here to be:

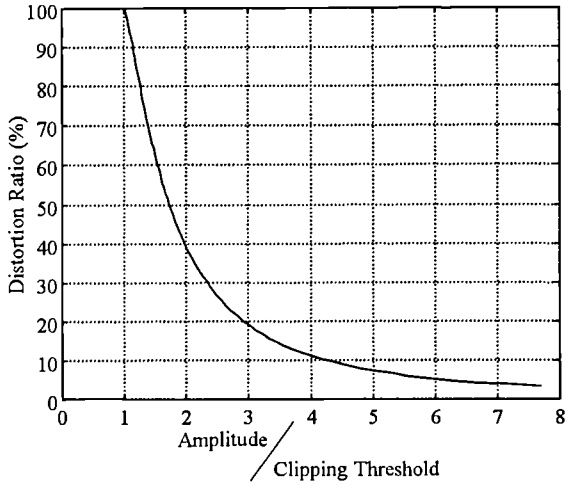
$$\text{Distortion Ratio} = \frac{\text{Power of Distortion Waveform}}{\text{Power of Input Signal}} \quad (5.7)$$

$$= \frac{[P-a_1]^2+\sum_{i=2}^{N} a_i^2}{P^2} \quad (5.8)$$

where P is the amplitude of the input sinusoid, which was assumed to be unity in the Fourier analysis, and N is the index of the highest harmonic that is within the system bandwidth. This definition of the distortion ratio allows the theoretical ratios obtained using a sinusoidal input to be compared with the distortion ratios calculated using speech, or other waveforms. It is important to note that as the amplitude to clipping threshold ratio is decreased, the distortion is more likely to be perceived as distinct periods of silence rather than 'continuous' distortion. For the distortion ratio

defined by (5.7), a value of 100% indicates that the clipper output is zero and therefore the distortion will be inaudible!

The exact value of the distortion ratio depends upon the threshold, V_T, and the amplitude of the input waveform, P. For example, the distortion ratio for a 1kHz sinusoid of amplitude ±1 that is clipped using a type 1 clipper having a threshold of T=0.5, is approximately 40%. Figure 5.2 shows more generally how the distortion ratio changes with the ratio of input amplitude to clipping threshold.



Distortion Ratio for a Type 1 Centre-Clipper

Figure 5.2

In practice, the threshold will be held constant for extended periods within which the amplitude of the waveform varies substantially. It can be seen that as the waveform amplitude increases, the distortion decreases for a fixed threshold. For the distortion to be negligible (less than 1% say), then the waveform amplitude must be approximately 13 times larger than the clipping threshold.

Note also that the power of the unwanted harmonics increases as the frequency of the fundamental decreases. This is because more harmonics are then present within the system bandwidth. In addition to the harmonic distortion, this clipper also causes another undesirable effect. If the clipper is switched out when echo is absent or when

near-end speech is present, to avoid distortion, then there will by a change in loudness of small amplitude near-end speech.

5.2.2 Type 2 Clipper Distortion Characterisation

For the second clipper (type 2), the Fourier coefficients of the output when a unity amplitude sinusoid is present at the input are given by:

$$a_0 = b_n = 0 \tag{5.9}$$

$$a_1 = \frac{1}{2\pi}[\sin(-2B) + \sin(2A)] + 1 + \frac{A}{\pi} - \frac{B}{\pi} \tag{5.10}$$

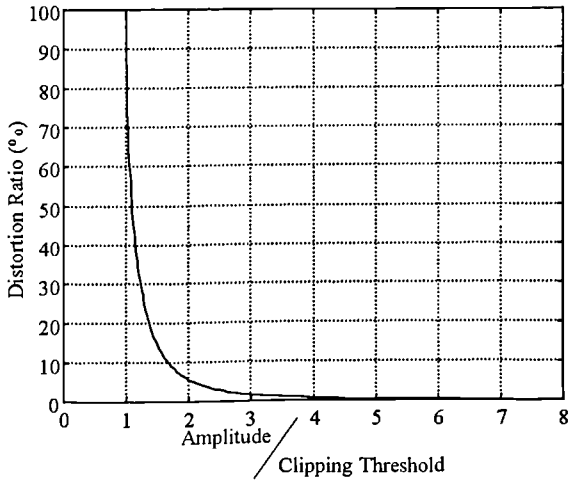
$$a_n = \frac{1}{\pi(n+1)}[\sin((n+1)A) + \sin(-(n+1)B)] + \frac{1}{\pi(n-1)}[\sin((n-1)A) + \sin(-(n-1)B)] \tag{5.11}$$

where

$$A = \cos^{-1}(V_T),\ B = \cos^{-1}(-V_T) \tag{5.12}$$

V_T = clipping threshold

The distortion generated by this clipper can be quantified using the ratio defined by equation (5.8). Figure 5.3 shows how this distortion measure changes with the ratio of input amplitude to clipping threshold, when a 1kHz sinusoid of unit amplitude is present at the input. The variation of distortion is similar to that obtained for the type 1 clipper except that it is somewhat smaller for any given amplitude to clipping threshold ratio.



Distortion Ratio for a Type 2 Centre-Clipper

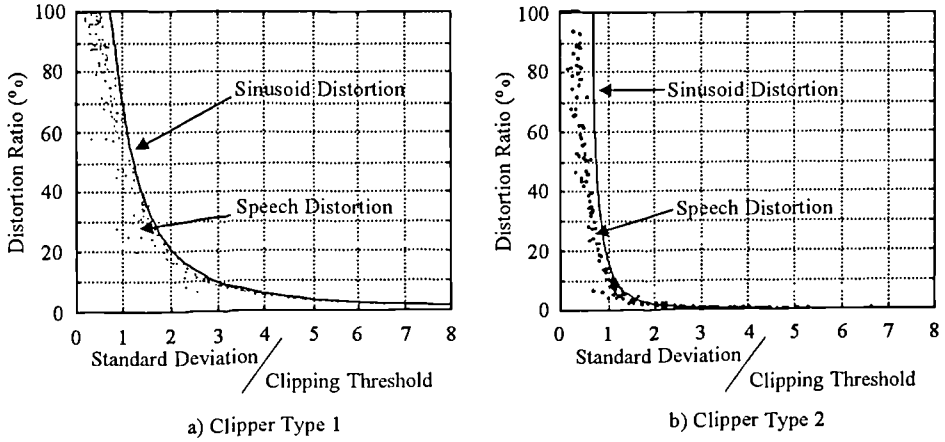
Figure 5.3

Figure 5.3 shows that the distortion ratio is approximately 5% when the threshold is half the input amplitude, whereas using the type 1 clipper it was 40%. Although the amplitude of the harmonics are similar when using the two clippers, the distorted fundamental in the output is much smaller when using the type 1 clipper, and therefore it generates more distortion. For the type 2 clipper, a distortion ratio of less than 1% is obtained when the input amplitude is approximately 3.5 times larger than the clipping threshold. In comparison to the type 1 clipper, negligible distortion is obtained when the input amplitude to threshold ratio is approximately 4 times smaller. Hence, for a fixed threshold the near-end signal will be less distorted by a type 2 clipper than by the type 1 clipper. However, it is a disadvantage of the type 2 clipper that if V_T is too small, then some echo residue will be transmitted and this will be less attenuated than for the type 1 clipper.

Although the unwanted distortion of the near-end waveform is reduced by the use of a type 2 clipper, ITU-T standard G.165 recommends that it is desirable to disable the clipper (by setting the threshold to zero) during periods of near-end talker to avoid the introduction of distortion. This means that during double-talk the uncancelled echo will be passed without attenuation, but will be masked to some extent by the near-end talker. Informal listening tests by the author support the above arguments for the superiority of the type 2 clipper.

It is important to note that the distortion ratios plotted in Figure 5.2 and Figure 5.3 do not exactly correspond to the distortion ratios that are obtained using speech. However, the general principle of the distortion varying with the ratio of waveform amplitude to clipper threshold is still valid. To illustrate this point, the distortion ratio was also calculated using ten seconds of continuous speech, in which the pauses between words had been removed. In order to measure the variation of distortion with amplitude/threshold ratio, the speech signal was processed using a fixed threshold and the distortion calculated by subtracting the original speech waveform from the output waveform. The distortion waveform was then segmented into 20ms blocks and the distortion ratio was calculated for each block, using equation (5.7). The block length was chosen as a compromise between being sufficiently large to give a smoothed estimate of the distortion power, but short enough so that the changes in distortion with speech amplitude can be measured.

The two scatter plots in Figure 5.4 show how the distortion ratios for the two clippers compare when a speech waveform is used at the input. The distortion ratio obtained using a sinusoidal input is also shown for comparison. Note that because the distortion ratio is calculated over 20ms blocks, the figures show how the ratio varies with standard deviation/threshold rather than amplitude/threshold.



Distortion Ratios Calculated Using Speech

Figure 5.4

For a given standard deviation to threshold ratio, the distortion obtained with speech is smaller than for the corresponding sinusoid. This is because speech has a different probability density function to that of a sinusoid. For a given standard deviation, speech has a higher proportion of samples at large amplitude, which are greater than the threshold. Thus, the distortion ratio is likely to be smaller than predicted by the Fourier analysis of a similarly distorted sinusoid. It would be interesting to find out if the distortion ratios obtained when Gaussian noise (or Laplacian/Gamma noise) is present at the clipper input correspond with those obtained for speech. It is expected that this will be the case.

5.2.2.1 Noise Modulation and Comfort Noise

In addition to the distortion of the near-end speech generated when the NLP is active, the clipping process causes what is known as 'noise modulation'. When the clipper is operating and the threshold is set to remove the residual echo, the far-end subscriber will hear silence or the idle channel noise of the digital long distance circuit. When switched out, the far-end subscriber will hear background sounds from the near-end. This switching between different noise levels is known as 'noise modulation' or 'noise-pumping' and is undesirable. Additionally, if the switching decisions are made in error, then disturbing bursts of background plus quantisation noise will be heard at the far-end.

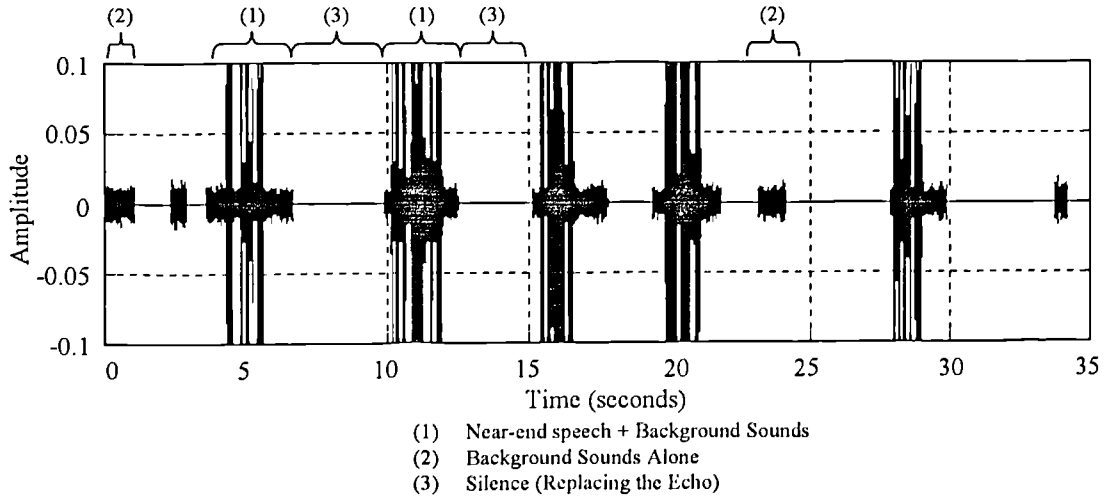


Figure 5.5

Figure 5.5 illustrates the process of noise modulation. In this example it can be seen that, when the clipper is operating, the output is zero, whilst when switched out, the near-end speech plus background sound is passed. At other times, bursts of background sounds can also be seen when the clipper is switched off in the absence of both echo and near-end talker. It is the switching between silence and background (and between silence and speech plus background) that is audible as noise modulation.

When echo cancellers were first used in analogue networks the circuit noise was more significant than the companding distortion or background sounds from the near-end of the network. The operation of any residual echo control device was then likely to be partially masked by the circuit noise. However, in today's digital networks the circuit noise can be much smaller than the acoustic noise, and hence the deficiencies of the residual echo control devices, such as noise modulation, have become more important.

Some echo cancellers attempt to mask noise modulation by using a comfort noise injection system that is similar to noise matching in DCME [ITUT91]. In such comfort noise systems, white pseudo-random noise whose power is matched to the average power of the background is injected during periods of centre-clipper operation. The characteristics of this process are illustrated in Figure 5.6, which shows time and frequency domain representations of the comfort noise.

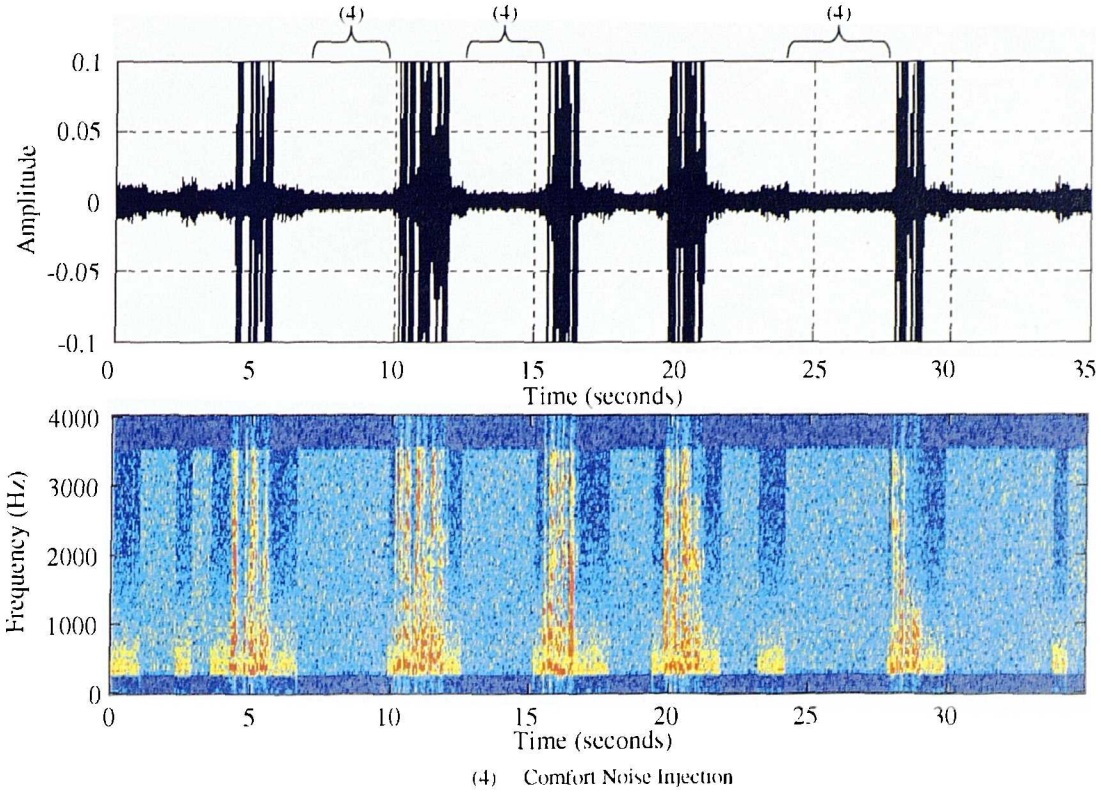


Figure 5.6

For this simulation, the power of the comfort noise was matched to that of the background sounds. However, Figure 5.6b shows that it is still obvious where comfort noise is added because its spectral shape is different from that of the actual background sounds. Informal listening tests by the author using high level language simulations to generate audio files, suggest that although much less disturbing than noise modulation, the switching between such comfort noise and the true background sounds is still audible, and does not sound very realistic. A number of schemes have been proposed to improve canceller behaviour in relation to the transmission of background sounds, and some of these will be described in the following sections

5.3 A Dual Threshold Centre-Clipper

An alternative to employing comfort noise to mask noise modulation is to use a clipper transfer function that attempts to pass some of the background sounds [BART91]. Figure 5.7 shows the transfer function of a dual threshold centre-clipper that seeks to do just this.

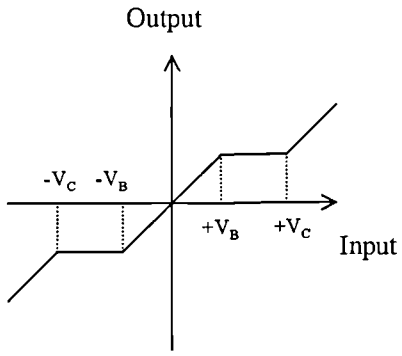


Figure 5.7

where V_B is the background transparency threshold and V_C is the echo clipping threshold. The transfer function of this clipper is described by equation (5.13).

$$\text{Type 3: Output} = \begin{cases} Input & if\ |\text{Input}| \le V_B \\ \text{sgn}(\text{Input}) \times V_B & if\ V_B < |\text{Input}| < V_C \\ Input - \text{sgn}(\text{Input}) \times (V_C - V_B) & if\ |\text{Input}| \ge V_C \end{cases} \quad (5.13)$$

The use of two thresholds permits small amplitude signals with peak amplitude less than the lower threshold V_B, to pass without distortion. Thus, the background sounds could be passed without modification by setting V_B to the peak amplitude of the background. Suppression of the uncancelled echo is achieved by limiting the clipper output to $\pm V_B$ when the residual echo, $r_q[n]$, is in the range $V_B < |r_q[n]| < V_C$. The upper threshold V_C should, as before, be set to the peak amplitude of the uncancelled echo. Therefore, the setting of V_B is critical in controlling the conflicting requirements of echo attenuation and background sound transmission. It will be argued later that to achieve acceptable echo suppression, V_B must be set to a level that is significantly smaller than the peak background amplitude, and this can result in audible mutilation of the background sounds.

As with the previous clippers, the dual threshold clipper will also distort any near-end speech, and should therefore be disabled during periods of near-end speech activity.

5.3.1 Type 3 Clipper Distortion Characterisation

As for the type 1 and type 2 clippers discussed previously, some appreciation of the distortion introduced by the type 3 clipper may be gained by Fourier analysing the response of the clipper to an input sinusoid. In this analysis, only the distortion of the background sounds by the lower threshold of the clipper will be considered.

It may be shown that the Fourier coefficients of a sinusoid distorted by the lower threshold of the clipping process are given by:

$$a_0 = b_0 = 0 \tag{5.14}$$

$$a_1 = \frac{1}{2\pi}[\sin(2B) + \sin(-2A)] + \frac{2V_B}{\pi}[\sin(B) + \sin(A)] + \frac{B}{\pi} + \frac{A}{\pi} \tag{5.15}$$

$$\begin{aligned} a_n = {} & \frac{1}{\pi(n+1)}[\sin((n+1)B) + \sin(-(n+1)A)] + \\ & \frac{1}{\pi(n-1)}[\sin((n-1)B) + \sin(-(n-1)A)] + \\ & \frac{2V_B}{n\pi}[\sin(nB) + \sin(nA)] \end{aligned} \tag{5.16}$$

where

$$A = \cos^{1}(V_B),\ B = \cos^{-1}(-V_B) \tag{5.17}$$

Figure 5.8 shows how the distortion ratios for both sinusoidal and speech inputs vary with standard deviation to clipping threshold ratio. The distortion ratios for speech were calculated using the method described in the preceding sections.

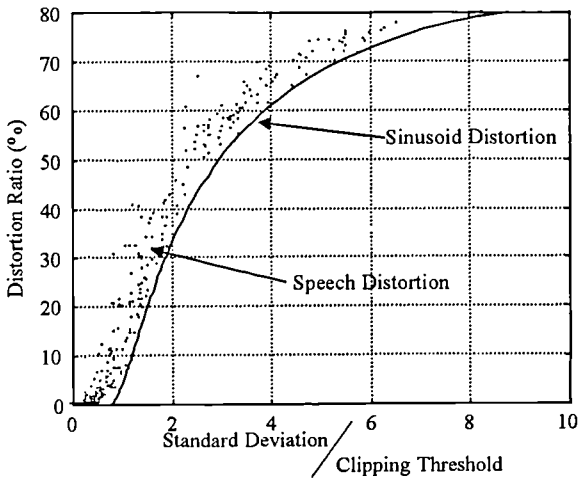


Figure 5.8

The distortion generated by 'peak-clipping' of the input waveform increases with the standard deviation/threshold ratio - this is expected as making V_B smaller increases the degree of clipping applied to the waveform. As previously, the distortion ratio calculated from Fourier analysis of a clipped sinusoid, does not exactly correspond with the measured values for a speech waveform. Here, the distortion obtained using speech is generally larger than the corresponding value obtained by Fourier analysis. As before, this is due to the different probability distributions of speech and sinusoids. For a given standard deviation to threshold ratio there will be a larger proportion of samples that are substantially greater than V_B for speech than for a sinusoid and thus, the distortion ratio for speech will be larger. The sinusoid distortion ratios should therefore be regarded as being an indication of the smallest distortion obtainable when using speech.

For peak clipping, the distortion ratio indicates that the difference between the clipper input and output waveforms increases as the standard deviation/threshold ratio increases. This difference manifests itself as not only corruption of the background sounds, but also as reduction of the transmitted power. Therefore in this case, the distortion ratio is helpful as a measure of the noise modulation introduced by the clipper.

Suppose that the samples of the background sounds have normally distributed amplitudes with unity standard deviation ($\sigma_b = 1$). Now, if V_B is chosen to be $4\sigma_b$ so that the standard deviation/threshold ratio is 0.25, then 99.99% of all the background samples will lie below the threshold. Figure 5.8 shows that after clipping the distortion will be insignificant and thus, the background sounds will be passed with little alteration. The power of the waveform at the clipper output will, essentially, equal σ_b^2.

The upper threshold should be set to a value such that it is seldom exceeded by the residual echo. The choice of V_B is then critical in determining the degree of both echo suppression and noise modulation. Consider the clipper output when $V_B = 4\sigma_b$ and uncancelled echo plus background sounds are present at the input. If the uncancelled echo is very much larger than the background then the clipper output is like a variable frequency square wave, which has a standard deviation of V_B. At the onset of echo, the clipper output power changes from σ_b^2 to $16\sigma_b^2$, i.e. there is an increase of approximately 12dB. This suggests that although the output is limited to $\pm V_B$ there will not be sufficient residual echo attenuation. However, there will be no noise modulation because the component of the input due to the background sounds has amplitudes less than V_B. Hence, the far-end subscriber is likely to hear background sounds plus distorted uncancelled echo. If the power of the uncancelled echo is now reduced, a smaller proportion of samples will be larger than V_B and hence the attenuation provided by the clipper decreases. This will lead to recognisable uncancelled echo being transmitted to the far-end. The uncancelled echo will be distinguishable even if it has no more than the same power as the background. Of course at the extreme, if the ERL is such that residual echo is sufficiently small to be masked by the near-end background sounds then no extra echo attenuation would be required.

It is clear that to obtain satisfactory suppression of the residual echo, V_B must be set to less than $4\sigma_b$ and thus, when the clipper is active the background sounds cannot be passed without some distortion. Setting $V_B = \sigma_b$ would appear to be a good

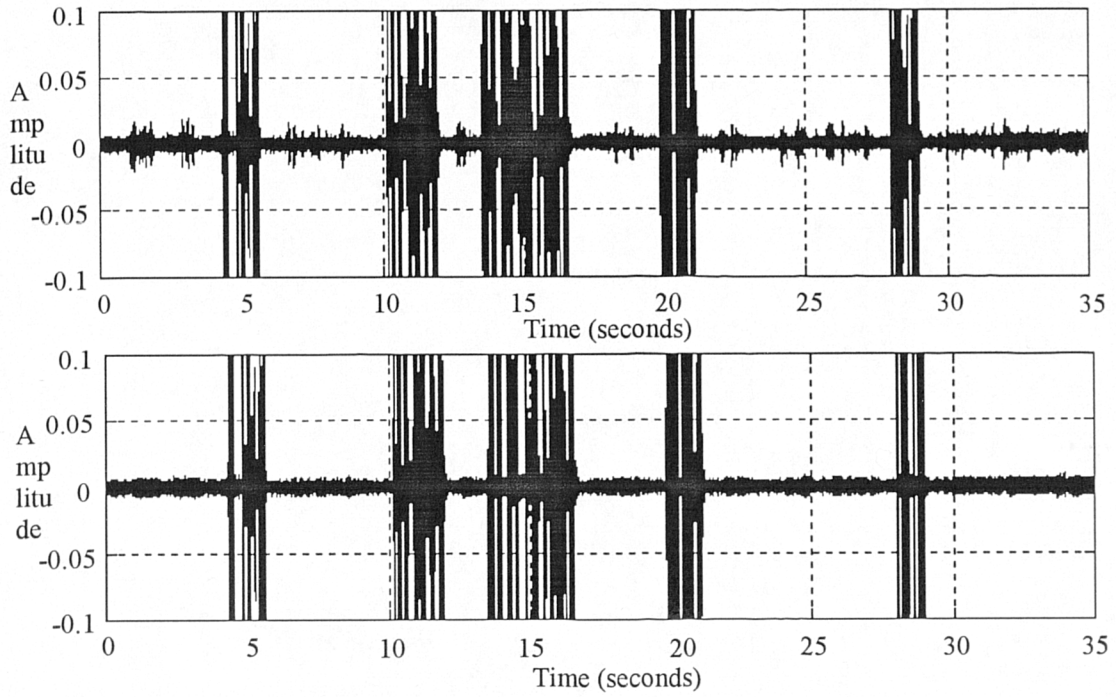
compromise because when uncancelled echo plus background sound is present, the power of the clipper output waveform is approximately the same as that of the background alone.

5.3.2 Type 3 Clipper Tests

The dual threshold clipper was tested using a high-level language simulation, which enabled the thresholds to be determined from prior knowledge of the background noise and the uncancelled echo. The lower threshold V_B was set in proportion to the envelope of the background sounds, which was estimated using a 1st order fast attack slow decay (FASD) filter (this was used as an estimate of the standard deviation of the background sounds). The upper threshold V_C was set in proportion to the output of another FASD filter that estimated the envelope of the uncancelled echo. Two different types of background noise, car noise and multispeaker noise, were used in the tests. These noise types may be classified as stationary and non-stationary respectively, and are believed to be representative of the kinds of noise that may be encountered in the network. Here, stationary is used in the sense that, the power of the waveform calculated with an averaging time of, say, 100 ms has little variation with time. Non-stationary, on the other hand, implies substantial variations of such a measure of power.

5.3.2.1 Clipper Characteristics Using Car Noise

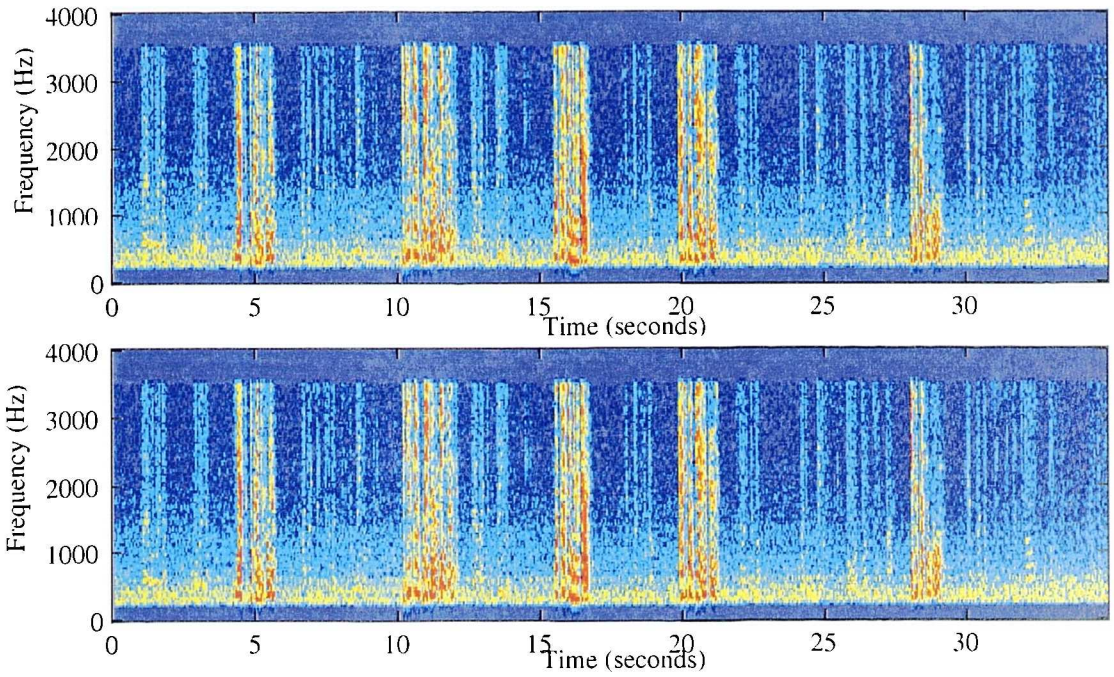
Figure 5.9 shows the unprocessed and processed residual echo, when ERL=6dB, ERLE=30dB (i.e. the linear echo component is reduced by ERLE dB) and using car noise whose power is 40dB less than that of the far-end speech. Note that at this level the background sound to quantisation noise ratio for A-Law is 25dB. In this example, V_B was set to $3\sigma_b$. Notice that the amplitude axis has been set to ± 0.1, from a full range of ± 1.0, in order to see the effects of the clipper more clearly.



a) before clipping, b) after clipping

Figure 5.9

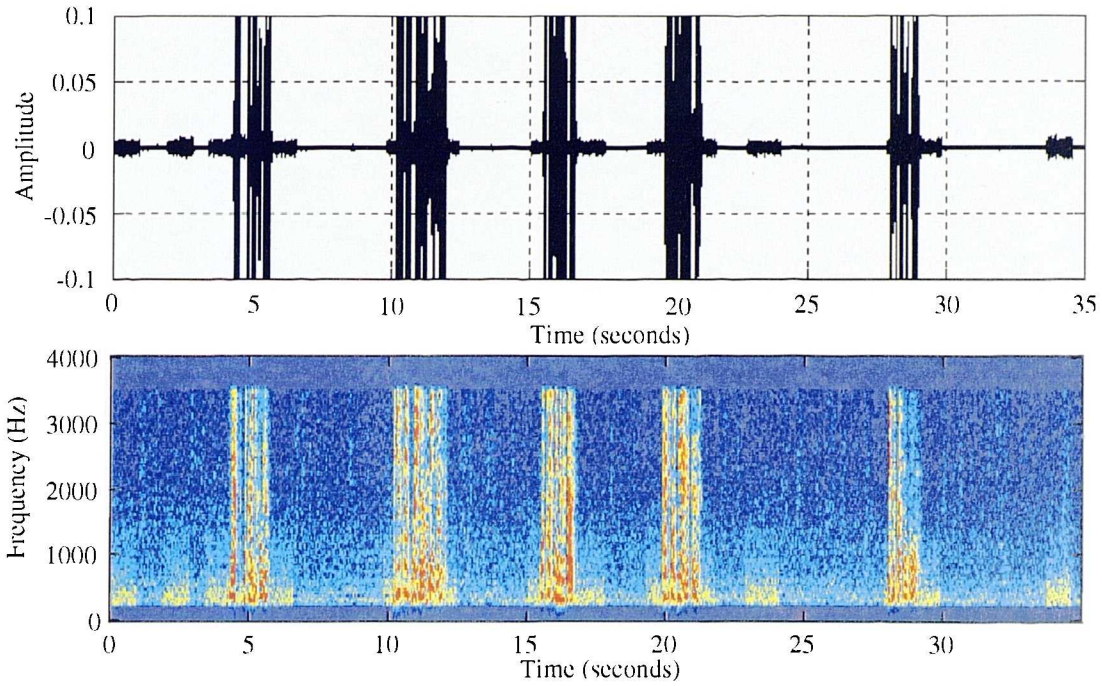
These graphs suggest that using these thresholds, the 'peaks' of the uncancelled echo are indeed removed. However, examination of the above signals in the frequency domain gives a better understanding of the characteristics. Figure 5.10 shows the spectrograms that were obtained from the unprocessed and processed waveforms of Figure 5.9.



a) before clipping, b) after clipping

Figure 5.10

The spectrograms suggest that, the power of the uncancelled echo is hardly reduced and this is confirmed by informal listening tests. To obtain satisfactory suppression of the uncancelled echo, it was necessary to set V_B to less than the standard deviation of the background sounds. Using a background threshold of approximately σ l to greater suppression whilst leaving the background sounds perceptually unal Although subscribers may find this level of uncancelled echo unobjectionable compared to the previous level, it is still audible. To reduce the uncancelled e further, it is necessary to set V_B to less than σ_b. Figure 5.11 shows the time d and frequency domain versions of the processed residual when $V_B=0.5\sigma$



a) clipped time domain waveform, b) spectral content of clipped waveform

Figure 5.11

It can be seen that not only is the power of the residual echo smaller than before, but that the power of the background sounds has also been reduced. This power variation is similar to that generated by the single threshold clipper, except not as severe Although preferable to the presence of uncancelled echo, this variation in background level is not ideal. The only way to obtain complete removal of the uncancelled echo is to use a threshold of V_B=0, i.e. to use a single threshold clipper

5.3.2.2 Clipper Characteristics using Multi-speaker Noise

Figure 5.12 shows the unprocessed residual echo, when ERL=6dB, ERLE=30dB (i e the linear echo component is reduced by ERLE) and using multi-speaker noise whose power is 40dB less than that of the far-end speech.

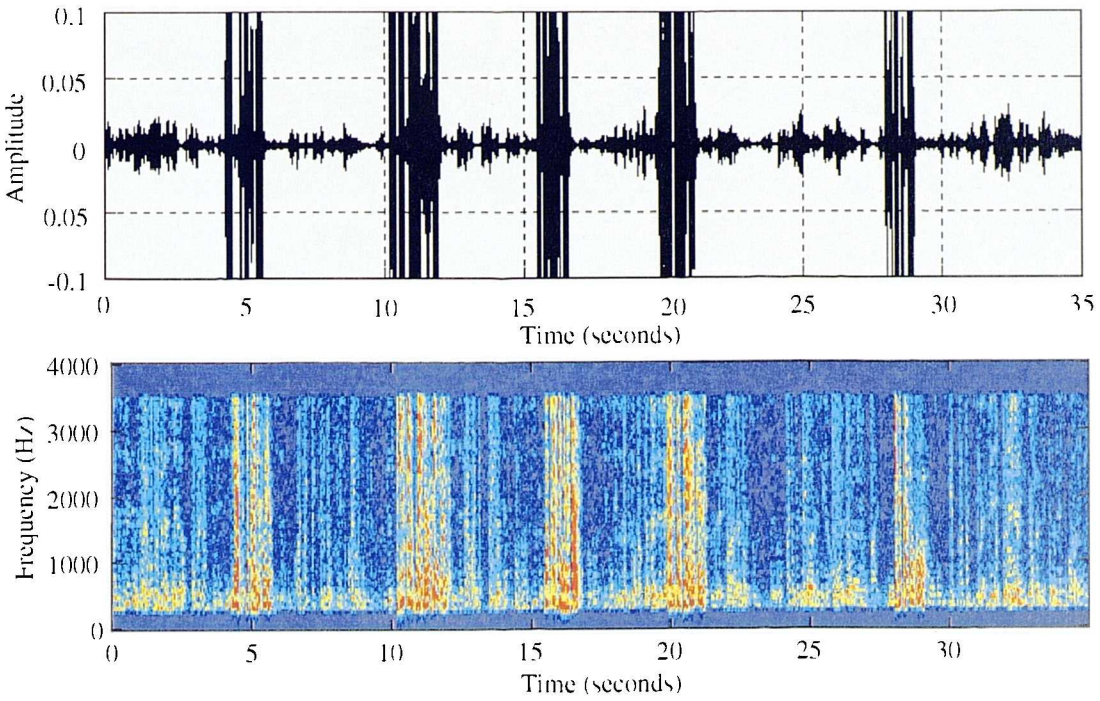


Figure 5.12

Again, in order to obtain any echo suppression it was necessary to set V_B to less than σ_h. Figure 5.13 shows the clipper output when V_B=0.5σ_l.

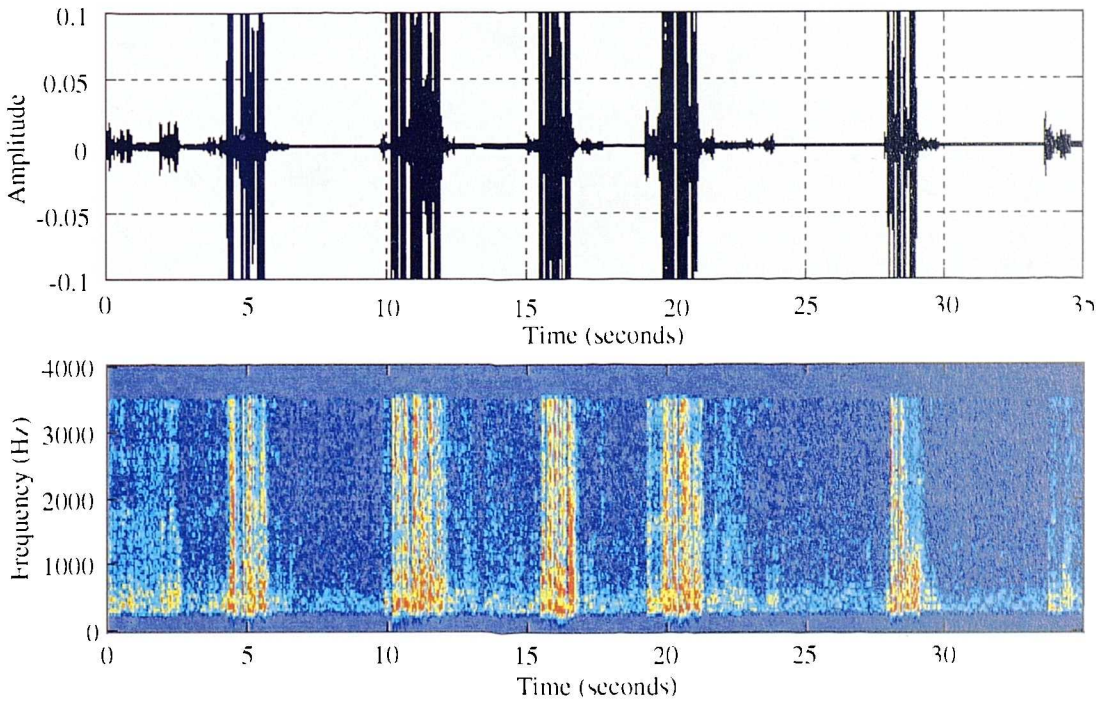


Figure 5.13

Once again, the attenuation of the uncancelled echo can be seen clearly. However what is less obvious, is the attenuation of the background sounds because the power and spectral shape of the background is continuously changing. This means that the noise modulation is now much less audible compared to when car noise was present. However, the presence of non-stationary background noise can also affect the audibility of the uncancelled echo. It was noticed that the uncancelled echo is more audible whenever the background sounds were loud prior to the onset of uncancelled echo, and quiet during this echo period. During these times, it is likely that V_B will be set too high and hence uncancelled echo is passed.

In comparison to the single threshold clipper, the dual threshold clipper does not suffer from noise modulation to the same degree, but in both cases the noise modulation will be worse when the background noise is stationary. If V_B is set to minimise the noise modulation ($V_B \approx \sigma_b$) the uncancelled echo will generally still be audible although at a greatly reduced level. However the uncancelled echo can only be eliminated completely if V_B is zero, i.e. the dual threshold clipper is now acting as a single threshold device, and this negates the advantages of using the two clipping thresholds. As with the single threshold clipper, the dual threshold clipper must also be disabled during near-end talker to prevent unwanted distortion.

5.4 Multi-band Centre-Clipping

The previously described centre-clippers are time domain devices that process the full bandwidth of the residual echo on a sample by sample basis. It was demonstrated however, that these clippers have two undesirable side-effects. Firstly, any near-end speech that is present during double-talking will be distorted, and secondly, the background sounds will suffer unwanted 'modulation' that depends upon the clipper type, threshold settings and the near-end background sounds.

It has been reported [MITC71] that the distortion of the near-end speech created during double-talk periods may be reduced by using a multi-band clipping process. This technique was originally proposed for use in echo suppression, where the

clipping applied in each band was responsible for removing the full echo signal. The multi-band technique has therefore been investigated here to discover if it can reduce the near-end talker distortion and modulation of the background sounds in an echo cancellation environment.

In the multi-band technique, the residual echo signal is split into a number of different frequency bands, using a bank of bandpass filters with different centre frequencies. Each separate filtered waveform is passed through a centre-clipper and the resulting output is again filtered using an appropriate bandpass filter. Finally, the clipped and filtered signals are added together to reconstruct the full-band waveform. The block diagram of a multi-band centre-clipper that has three frequency bands is shown below in Figure 5.14.

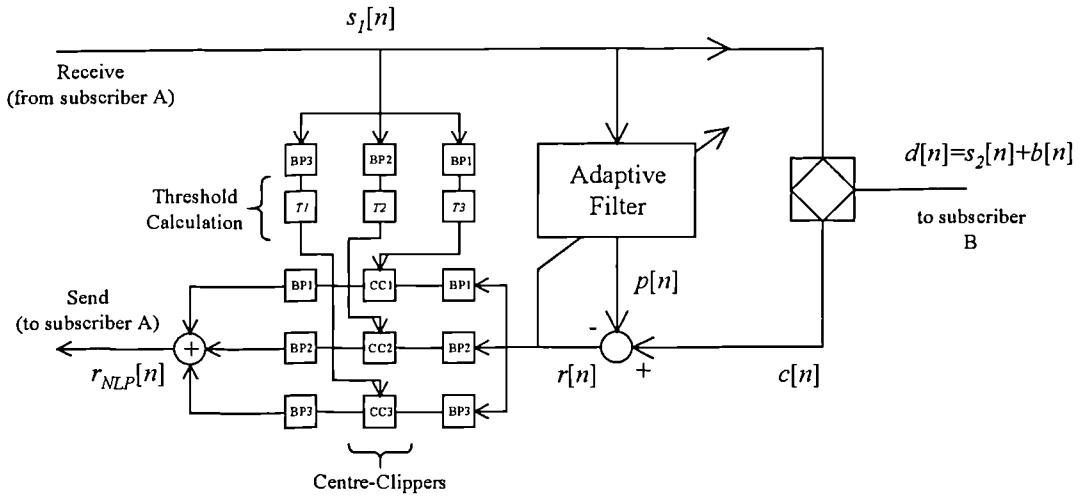
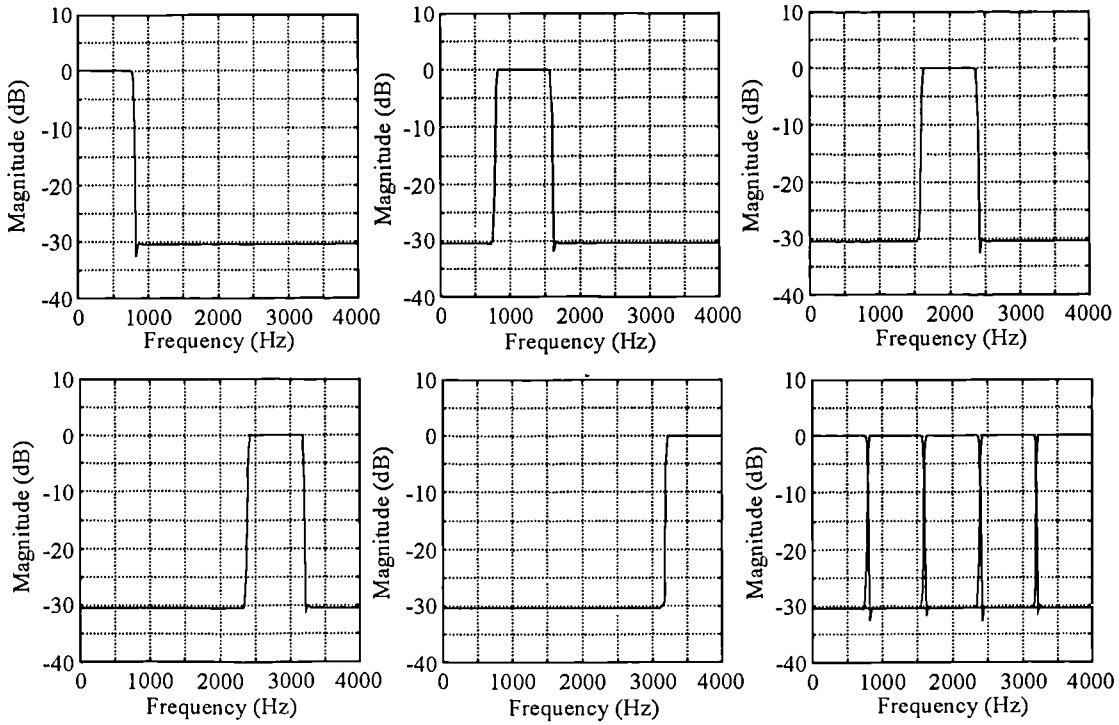


Figure 5.14

In the diagram, the bandpass filters are denoted by BPn, and the clippers by CCn. The original echo suppressor system made use of single threshold type 1 clippers. However, the system investigated here uses the dual threshold (type 3) clippers, so that the effects of background sound transmission in the multi-band case can be compared with the single-band system.

The motivation for using the band splitting process is as follows. When the signal in each sub-band is clipped any resulting distortion will occupy the full Nyquist bandwidth (4kHz in this case), not just the bandwidth of that sub-band. Thus when the clipped signal is passed through the corresponding output bandpass filter, any distortion that is present outside the designated passband will be attenuated. Distortion that is present within the passband, and is therefore transmitted, may be masked by the signal itself. If there is sufficient stopband attenuation, the attenuated distortion that lies outside the bandwidth of a particular sub-band, is masked by the signal present in the adjacent sub-bands. Therefore provided there are a sufficient number of bands, there should be less distortion of the near-end talker if the clippers remain enabled during periods of both near-end single and double talk. In addition, use of the band splitting process enables the thresholds to be adjusted separately for each band and this should help to reduce the distortion and noise modulation.

In order to test the effectiveness of this technique, a multi-band system using twin threshold clippers has been implemented as a high-level language simulation. The clipper was implemented with 5 sub-bands of width 800Hz that are uniformly spaced across the Nyquist bandwidth, i.e. 0 to 4kHz. The band splitting is achieved using 512th order FIR filters that were designed using the MATLAB fir2 procedure, and have the frequency responses shown below in Figure 5.15.

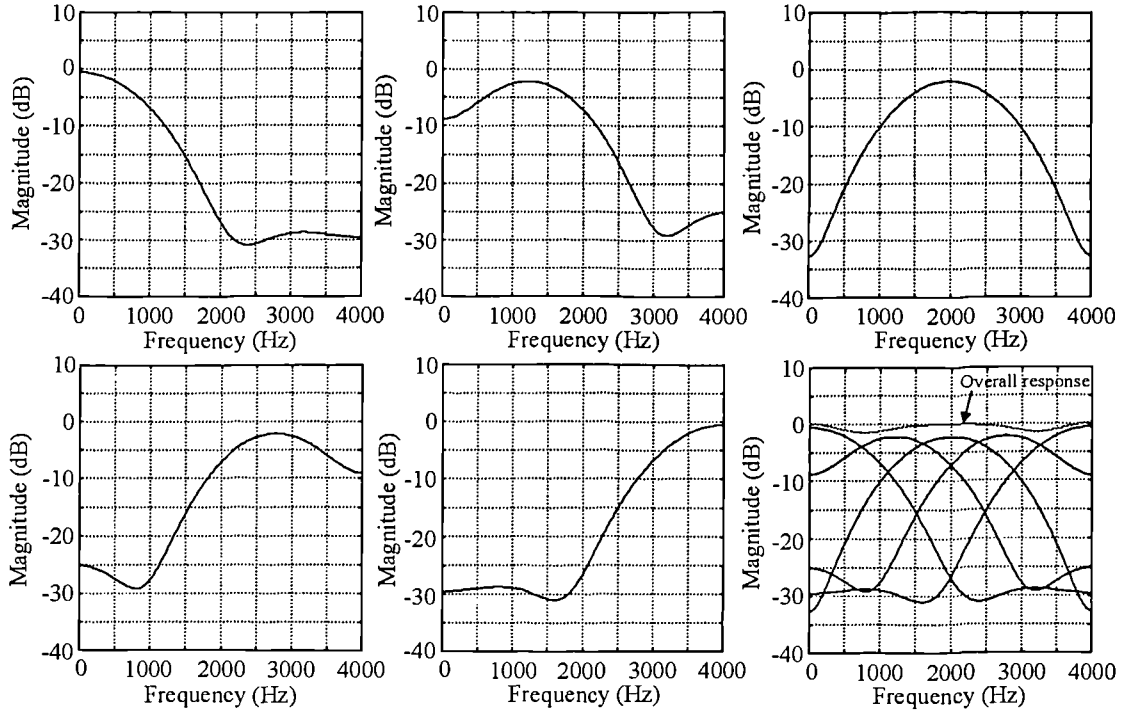


a) to e) frequency response of each filter, f) combined frequency response

Figure 5.15

The five filters cover the full Nyquist bandwidth but without overlapping. It may be seen that the roll-off at the passband edges is very sharp because of the large order used, and that there is a stop-band attenuation of approximately 30dB (set arbitrarily).

In a network echo canceller no more than 1ms of group delay must be introduced into the send path [ITUT94], which corresponds to 8 samples at a sampling frequency of 8kHz. Thus, the use of high order filters to perform the band splitting would not be permitted. The largest order linear phase FIR filter permitted is 8 because each band requires the use of two filters. Therefore, the following set of 8^{th} order filter characteristics, that result from using the same design parameters, were also used.



a) to e) frequency response of each filter, f) combined frequency response

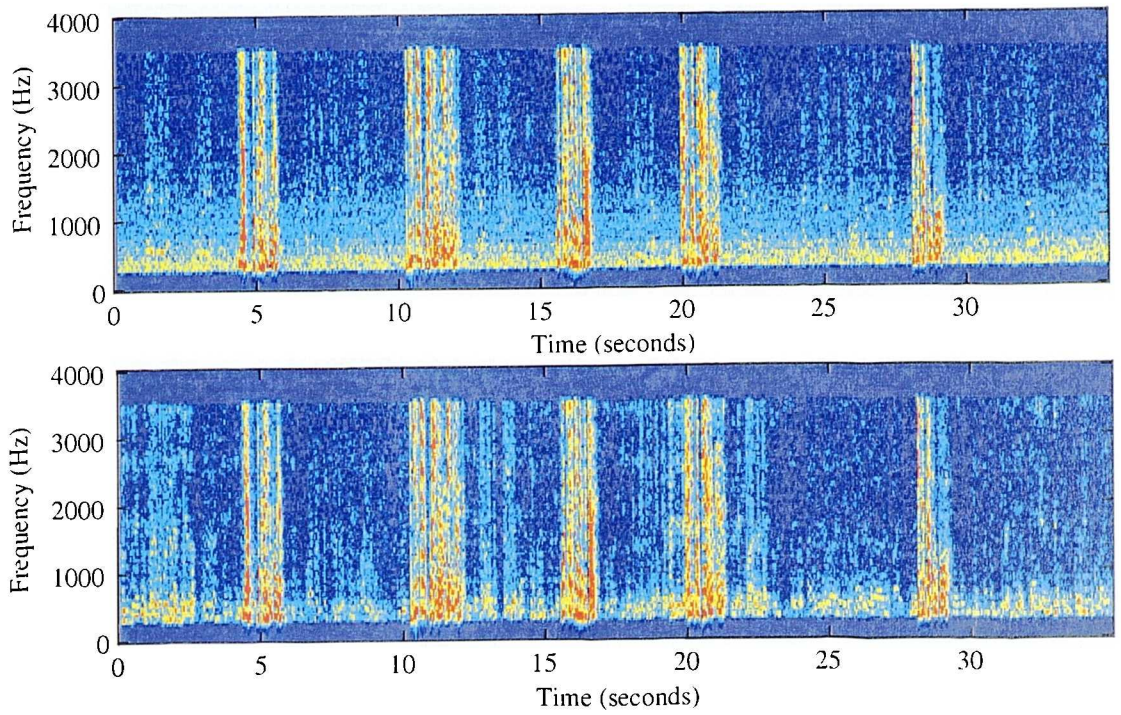
Figure 5.16

The desired passband regions are now defined with much less precision than for the 512 tap filter, with the roll-off from the point of smallest attenuation in each band being very slow. Note that gain responses similar to those of the 512 tap FIR filter could be achieved with much lower order IIR filters. Although it is generally considered that the ear is insensitive to non-linear phase, network echo cancellers are required to operate, at times, whilst non-speech signals are present, for example digitally modulated waveforms. If IIR filters are to be used, they must be designed carefully so as not to infringe the permitted group delay requirements, and hence ensure that they do not interfere with any non-voice traffic that is present while the canceller is active.

5.4.1 Multi-band characteristics using high order filtering

As with the testing of previous clippers, the high-level language implementation allows the thresholds to be set from prior knowledge of the test signals. The upper and lower thresholds in each band were set by the method described in section 5.3.2.

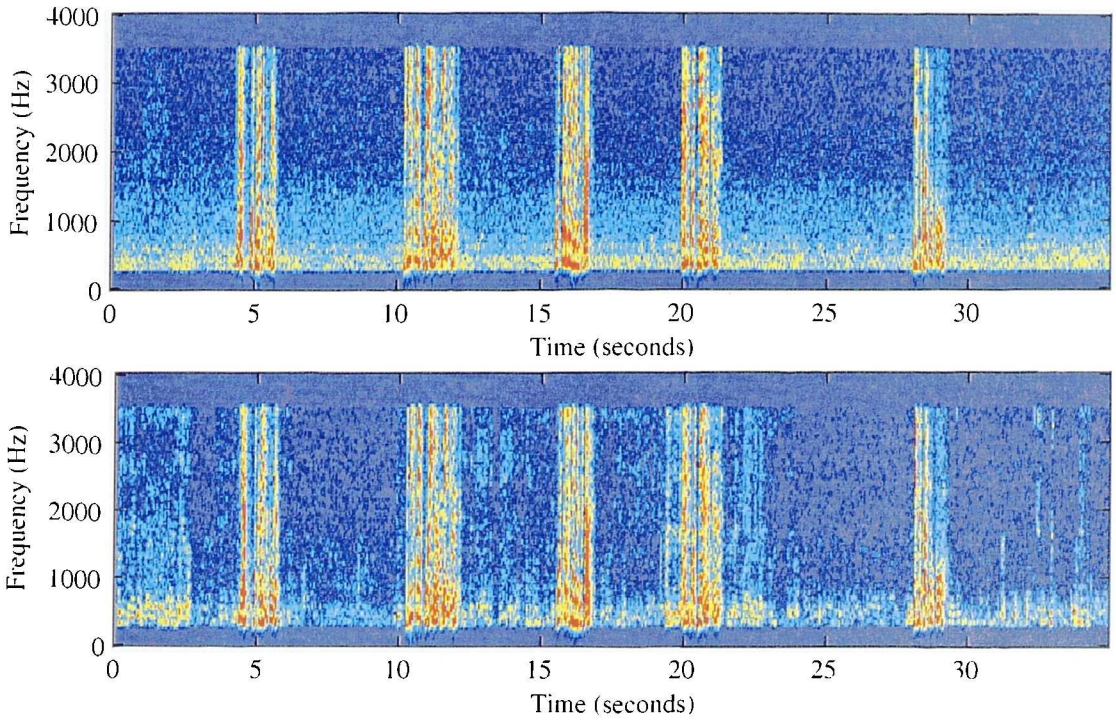
Figure 5.17 shows two spectrograms, calculated when car and multi-speaker noise are present at the near-end and have powers that are 40dB less than the far-end talker. In this case ERL=6dB, ERLE=30dB (i.e. linear component of the echo is attenuated by 30dB), and V_B is approximately $3\sigma_b$.



Clipped waveforms for a) car noise and b) multi-speaker noise @ -40dB

Figure 5.17

As with the single-band type 3 clipper, there is very little echo suppression for both the car noise and multi-speaker noise cases, and it is necessary to set the background transparency threshold V_B to less than σ_b, in order to obtain satisfactory suppression. Figure 5.18 shows similar spectrograms obtained under identical conditions except that V_B now equals σ_b.



Clipped waveforms for a) car noise and b) multi-speaker noise @ -40dB

Figure 5.18

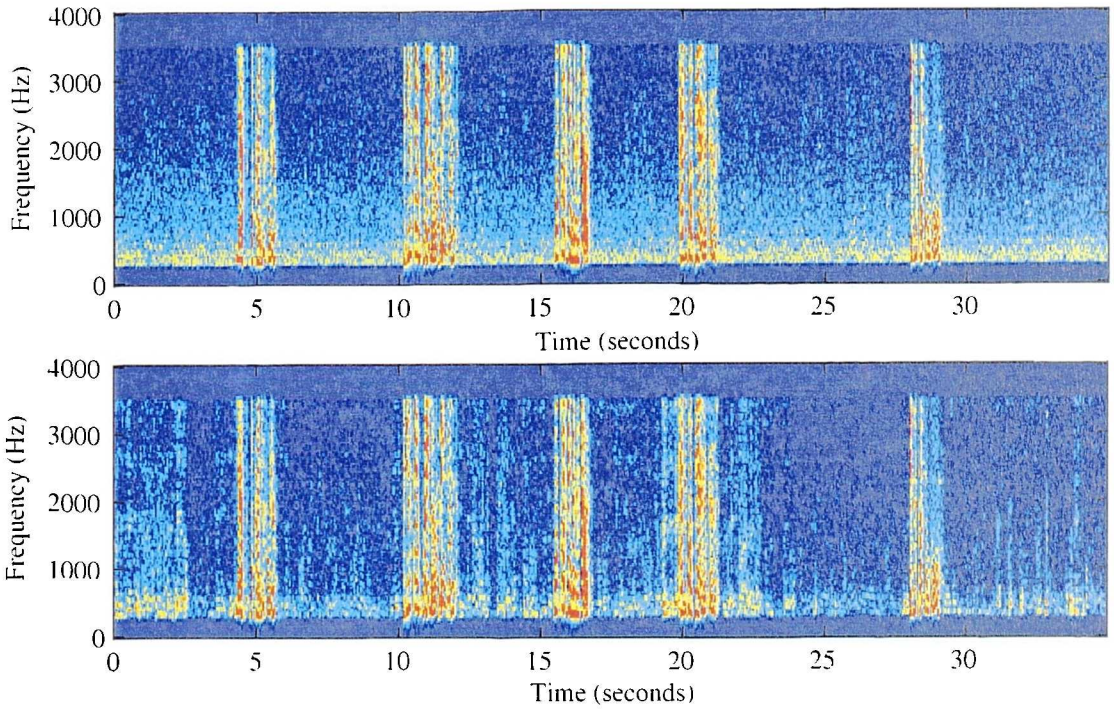
It may be seen that for car noise, the uncancelled echo appears to have been completely removed, and that during these periods, the background sounds have only been altered slightly. However, listening tests show that the uncancelled echo is still audible, although at a greatly reduced level. The tests also suggest that the background sound is indeed passed without perceivable alteration, i.e there is no audible noise modulation. Compared to a single-band type 3 clipper with $V_B=\sigma_l$, the residual echo is less audible so there is no need to use threshold settings of less than this value.

When multi-speaker noise is present at the near end, the uncancelled echo is removed in a similar way, whenever the lower thresholds are good estimates of the background sound standard deviation in each sub-band. However, as with the single band equivalent, the variable nature of multi-speaker noise results in audible uncancelled echo being passed if the background noise was loud prior to the onset of echo, and quiet during the echo. This effect may be clearly seen by comparing Figure 5.12b to Figure 5.18b. Thus when the background noise is stationary, the use of the

multi-band configuration with twin threshold clippers has the potential to reduce the uncancelled echo power without obvious noise modulation. However, when the background noise is non-stationary and the transparency threshold is set as described, it is likely that uncancelled echo will frequently be 'unmasked' and louder than the background sounds. Informal listening tests confirm that during double-talk periods, the uncancelled echo is inaudible, because it is masked by the near-end speech, and that the near-end speech although not identical to the original, is less distorted than with the single band clipper.

5.4.2 Multi-band Characteristics using low order filtering

The following pair of spectrograms were calculated using a transparency threshold $V_B=\sigma_l$, and using the 8th order FIR bandpass filters whose frequency responses are shown in Figure 5.16.



Clipped waveforms for a) car noise and b) multispeaker noise @ -40dB

Figure 5.19

All the comments that were made in the previous section generally apply when using low order filtering, but with the following exceptions. When the car noise is present at the near-end, noise modulation is again virtually inaudible, but the uncancelled echo is not suppressed quite as well. This can by seen by comparing Figure 5.19a and Figure 5.18a. Although the cancelled echo is not suppressed to the same degree, it is still less audible than the suppressed echo obtained for the equivalent full-band type 2 clipper. It can therefore be concluded that very little attenuation in the stopbands is required in order for the multi-band technique to show an improvement over the single band implementation. During double-talk, the distortion of the near-end speech is more evident than when the higher order filters were used. Again, this distortion is smaller than the near-end speech distortion created by the single-band implementation.

5.4.3 Comparison of multi-band and single band twin-threshold clippers.

The multi-band technique is superior to the single band method in that the distortion suffered by the near-end talker during double talk is much less severe, even when low order FIR filters are used to perform the band splitting. When the background noise is stationary, the use of the multi-band system with twin threshold centre clippers leads to superior suppression of the uncancelled echo (although echo is still audible), which may be achieved with very little noise modulation. However, as with the single-band implementation, when the background sounds are non-stationary it is probable that the uncancelled echo, although attenuated, will be unmasked due to the time varying nature of the background sounds. This means that if complete removal of the residual echo is to be guaranteed in all circumstances, a single threshold clipper must be used and noise modulation is then likely to be a problem.

5.5 Frequency Domain 'Echo Shaping'

The previously described techniques rely on time domain methods to suppress the uncancelled echo. Although the multi-band centre-clipping method splits the residual echo signal into different sub-bands, time domain clipping is still used to remove the uncancelled echo. An alternative is to process the residual echo signal in the

frequency domain, either by using linear filtering (implemented in either the time or frequency domain), or by using non-linear techniques to process the spectrum.

The ultimate goal of processing the canceller residual in the frequency domain is to modify its spectrum such that, after processing, it has a similar spectral shape and level to any background noise that may be present. In this way, the uncancelled echo may be masked by the remaining background sounds. The ideas described in this section were first proposed in [MART95] and [MART96], to reduce the level of uncancelled echo that arises in an acoustic echo environment. This section briefly examines the principles involved in this technique, but from the viewpoint of use in network echo cancellation.

In an acoustic echo canceller, a typical impulse response of a room might require the use of several thousand taps of an FIR filter. In the original technique [MART95], it is assumed that the length of the adaptive filter delay line is set such that the echo arising from the path of lowest attenuation can be cancelled. Any further echoes will lie outside the span of the filter and will contribute to the uncancelled echo. However, it is assumed here that the only source of echo is the hybrid and that the impulse response of the echo path fits completely within the span of the adaptive filter. The uncancelled echo will then be due to companding noise, non-linearities and filter misadjustment.

In the echo shaping technique proposed by [MART95], an additional time varying FIR filter, of order N is placed in the send path of the canceller, as shown in Figure 5.20.

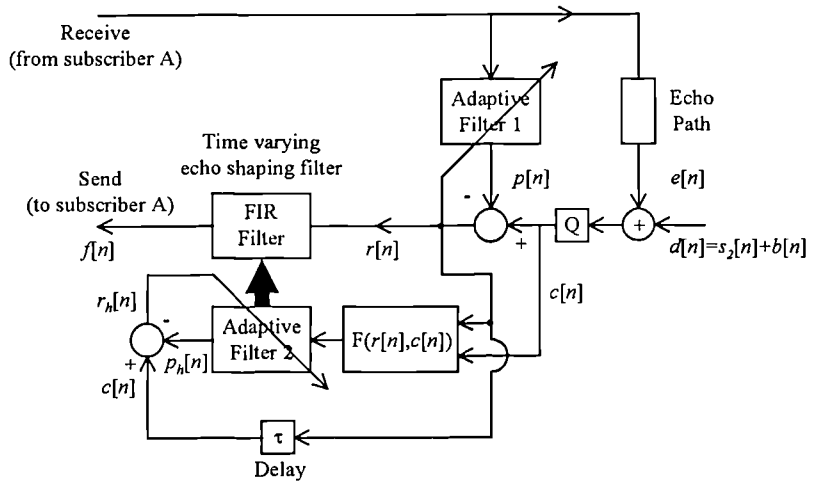


Figure 5.20

In a 'perfect' system, the additional FIR filter would be able to transform the uncancelled echo plus near-end signal so that it is spectrally similar to the near-end signal alone. This might be achieved by varying the frequency response of the filter such that its attenuation is high at frequencies where the uncancelled echo is large and small when the uncancelled echo is small. Thus the processed signal, $f[n]$, may have a spectral shape that is more similar to the background sounds and hence the uncancelled echo may be indistinguishable from the background sounds. In this system an adaptive filter, denoted as 'Adaptive Filter 2' in Figure 5.20 is used to identify the required frequency response. The resulting set of coefficients is transferred to the send path FIR filter at every iteration, and this filter then performs the echo shaping function.

It should be noted that the filtering operation and identification of the frequency response could also be implemented in the frequency domain. However as mentioned previously, a network echo canceller is only allowed to introduce a maximum of 1ms delay in the send path. In general, this rules out the possibility of using a frequency domain implementation, but the use of an adaptive time domain system, as shown in Figure 5.20, allows the delay criteria to be satisfied by using a 16 tap FIR filter. The operation of the adaptive filter is discussed in the following sections.

5.5.1 Echo Shaping: The 'Ideal' Filter Response

It has been proposed that if the background noise and near-end talker signals were known, the adaptive system shown in Figure 5.21 could be used to identify the desired characteristic [MART96].

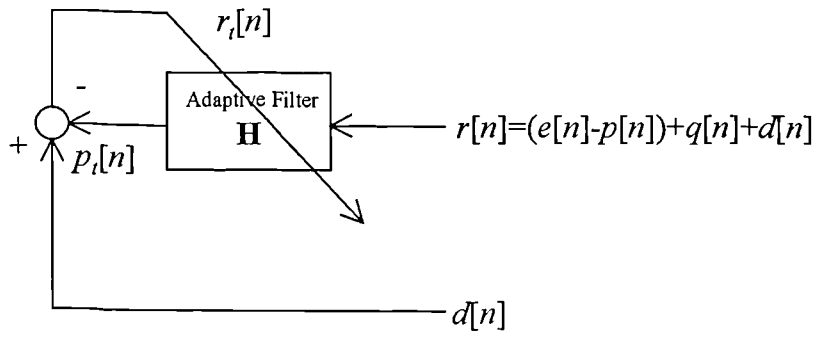


Figure 5.21

In this configuration, the adaptation process attempts to minimise the error between the near-end signal $d[n]$ and the filter output $p_t[n]$, and hence identify a transfer function that converts the uncancelled echo plus near-end signal into the near-end signal alone. During periods of near-end talker, $r[n]=d[n]+q[n]$, the filter would adapt to become 'transparent' and therefore pass all signals without alteration. Of course, in a practical network echo canceller $d[n]$ (or $q[n]$) will not be available and a slightly different arrangement must be used. This will be discussed in the next section.

It may be shown [LATH89] that the frequency response of the optimum Wiener filter that minimises the mean squared difference between $r[n]$ and $p_t[n]$ can be written in terms of the cross power spectral density (PSD) between $d[n]$ and $r[n]$, $P_{dr}(\omega)$, and the auto-PSD of $r[n]$, $P_{rr}(\omega)$:

$$H_{opt}^{(I)}(\omega) = \frac{P_{dr}(\omega)}{P_{rr}(\omega)} \qquad (5.18)$$

where (I) denotes that this is the transfer function of the optimum filter for this 'ideal' filter. Since $(e[n]-p[n])$ is uncorrelated with $d[n]$ we may write:

$$P_{dr}(\omega) = P_{dd}(\omega) \qquad (5.19)$$

and $$P_{rr}(\omega) = P_{dd}(\omega) + P_{qq}(\omega) + P_{(e-p)(e-p)}(\omega) \quad (5.20)$$

where $P_{dd}(\omega)$ is the PSD of the near-end signal, $P_{qq}(\omega)$ is the PSD of the quantisation noise, and $P_{(e-p)(e-p)}(\omega)$ is the PSD of the uncancelled echo, the frequency response may be expressed as:

$$H_{opt}^{(I)}(\omega) = \frac{P_{dd}(\omega)}{P_{dd}(\omega) + P_{qq}(\omega) + P_{(e-p)(e-p)}(\omega)} \quad (5.21)$$

This frequency response consists purely of real values because the PSD's are Fourier Transforms of even autocorrelation functions. Therefore because the frequency response has zero phase, $H_{opt}^{(I)}(\omega)$ is equal to the gain response $G_{opt}^{(I)}(\omega)$. Equation (5.21) may be simplified by writing the PSD of the uncancelled echo, $P_{(e-p)(e-p)}(\omega)$, as a scaled version of the echo PSD, $P_{ee}(\omega)$:

$$P_{(e-p)(e-p)}(\omega) = \delta_{ERLE}{}^{2}(\omega) P_{ee}(\omega) \quad (5.22)$$

where $\delta_{ERLE}(\omega)$ is the scale factor that determines the echo attenuation at frequency ω. The scale factor is given by:

$$\delta_{ERLE}(\omega) = \sqrt{10^{-ERLE(\omega)/10}} \quad (5.23)$$

where ERLE(ω) is the echo attenuation in dBs at frequency ω. Equation (5.21) may now be written as:

$$H_{opt}^{(I)}(\omega) = \frac{P_{dd}(\omega)}{P_{dd}(\omega) + P_{qq}(\omega) + \delta_{ERLE}^{2}(\omega) \times P_{ee}(\omega)} \tag{5.24}$$

$$= \frac{1}{1 + \frac{P_{qq}(\omega)}{P_{dd}(\omega)} + \delta_{ERLE}^{2}(\omega) \times \frac{P_{ee}(\omega)}{P_{dd}(\omega)}}$$

$$= \frac{1}{1 + SNR_q(\omega) + \delta_{ERLE}^{2}(\omega) \times SNR(\omega)} \tag{5.25}$$

where $SNR(\omega)$ is the echo to near-end signal power ratio at frequency ω.

$$SNR(\omega) = \frac{P_{ee}(\omega)}{P_{dd}(\omega)} \tag{5.26}$$

and $SNR_q(\omega)$ is the quantisation noise to near-end signal power ratio at frequency ω.

$$SNR_q(\omega) = \frac{P_{qq}(\omega)}{P_{dd}(\omega)} \tag{5.27}$$

Equation (5.25) states that the magnitude of the response of the optimum Wiener filter is determined by the frequency dependent echo attenuation provided by the canceller, the frequency dependent quantisation noise to background noise power ratio, and the frequency dependent echo power to background noise power ratio. Figure 5.22 shows how the attenuation at frequency ω varies with the ERLE(ω) and $SNR(\omega)$, if the optimum filter given by equation (5.25) was used with $SNR_q(\omega)$=0.

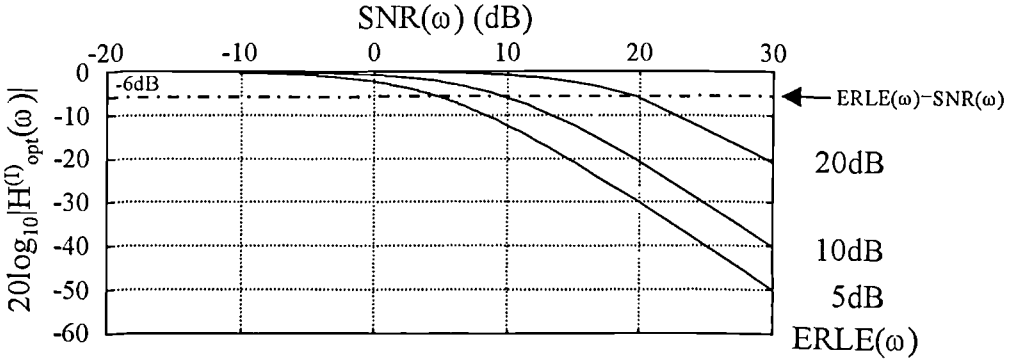


Figure 5.22

Figure 5.22 shows that the attenuation increases with $SNR(\omega)$, and is inversely related to the ERLE(ω) provided by the adaptive filter in the echo canceller. In other words, as the power of the uncancelled echo increases, the attenuation of the ideal filter characteristic also increases. When $SNR(\omega)$=ERLE(ω), the residual echo has the same power as the background noise and the filter attenuation is 6dB. In this case however, only 3dB of attenuation is required to reduce the canceller residual PSD, $P_{rr}(\omega)$, to the same level as $P_{dd}(\omega)$. Thus if the optimum filter was used, the frequencies that contain residual echo are likely to be attenuated to a level below that of the background noise, and therefore the filtered signal will not have a shape that is similar to the background.

By using the relationship

$$P_{ff}(\omega) = \left|H_{opt}^{(I)}(\omega)\right|^2 P_{rr}(\omega) \qquad (5.28)$$

that describes the power spectrum at the output of a linear filter in terms of the input power spectrum and filter frequency response, it is straightforward to show that when using the 'ideal' optimum filter, the PSD of the 'shaped' residual, $P_{ff}(\omega)$, is given by:

$$P_{ff}(\omega)dB = P_{dd}(\omega)dB + \frac{20\log_{10}\left|H_{opt}^{(I)}(\omega)\right|}{2} \qquad (5.29)$$

This equation shows that $P_{ff}(\omega)$ will always be smaller than $P_{dd}(\omega)$ if the attenuation at frequency ω is non-zero. The use of the 'ideal' optimum filter will never result in the processed signal having the same shape as the background noise, especially if either ERL or ERLE(ω) is small. Under these conditions, the use of the optimum filter is likely to create a 'hole' in the spectrum at frequencies where the echo is larger than the background sounds. It is therefore likely that the use of the 'ideal' echo shaping as proposed by [MART96] will cause noise modulation.

The introduction of quantisation noise, by setting $SNR_q(\omega)\neq 0$, does not alter the generality of the above conclusion. Equation (5.25) shows that provided $b[n]\neq 0$, the presence of quantisation noise increases the attenuation provided by the ideal filter. If there was no near-end background noise, i.e. $b[n]=0$, the filter would provide infinite attenuation at all frequencies and hence any uncancelled echo would be removed completely.

5.5.2 Echo Shaping: A Practical Filter Response

The previously described system cannot be used in a real canceller because the input $d[n]$, that consists of the background noise plus near-end talker, is unknown. It has been proposed [MART96] that a different set of inputs could be used in practice.

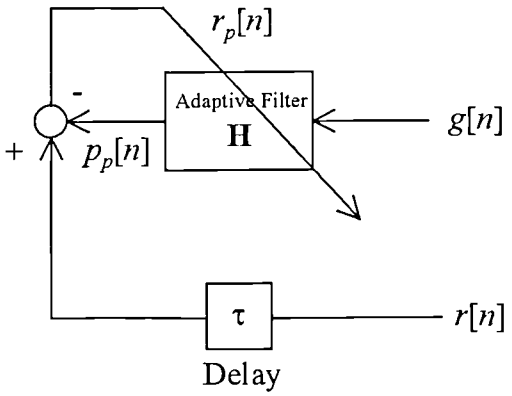


Figure 5.23

Now, the reference input to the adaptive filter, g[n], is given by:

$$g[n]=\alpha[n]c[n]+(1-\alpha[n])r[n] \qquad (5.30)$$

where $\alpha[n]$ is a time varying mixing factor ($0\leq\alpha[n]<1$) whose value is made dependent upon the presence or absence of double-talk. During periods of far-end echo (single-talk), the system should set $\alpha[n]=1$ so that the adaptive filter will attempt to transform the near-end signal $c[n]$, that consists of the full echo plus background plus quantisation noise, into the canceller output signal $r[n]$, i.e. uncancelled echo plus near-end background and quantisation noise. With these

inputs, the adaptive filter identifies the impulse response of the path through the echo canceller subtractor, which is governed by the adaptive filter in the echo canceller. Therefore, the overall attenuation applied to the uncancelled echo by the send path filter depends upon the echo attenuation of the canceller.

During periods of near-end single-talk, when the system should again set $\alpha[n]=1$, $g[n]\approx c[n]$ and $r[n]\approx c[n]$, and therefore the filter should adapt to become transparent and hence allow the near-end speech to pass without distortion. The echo shaping filter should also be transparent during periods of double-talk, when the residual echo is masked by the near-end speech. In this case transparency is achieved by setting $\alpha[n]=0$ so that the adaptive filter has $r[n]$ as both reference and desired signals.

It has been shown by [MART96] that the magnitude response of the optimum filter, for the system shown in Figure 5.23, may be written as:

$$H_{opt}^{(P)}(\omega)=\frac{1+\Phi[n](\omega)\times SNR_t(\omega)\times \delta_{ERLE}(\omega)}{1+\Phi[n](\omega)^2\times SNR_t(\omega)} \tag{5.31}$$

where

$$\Phi[n](\omega)=\delta_{ERLE}(\omega)+\alpha[n](1-\delta_{ERLE}(\omega)) \tag{5.32}$$

$$SNR_t(\omega)=\frac{P_{ee}(\omega)}{P_{dd}(\omega)+P_{qq}(\omega)} \tag{5.33}$$

and all other symbols have their previously defined meanings. For an infinite echo power to background plus quantisation noise power ratio, ($b[n]=0$, $q[n]=0$), the transfer function is given by:

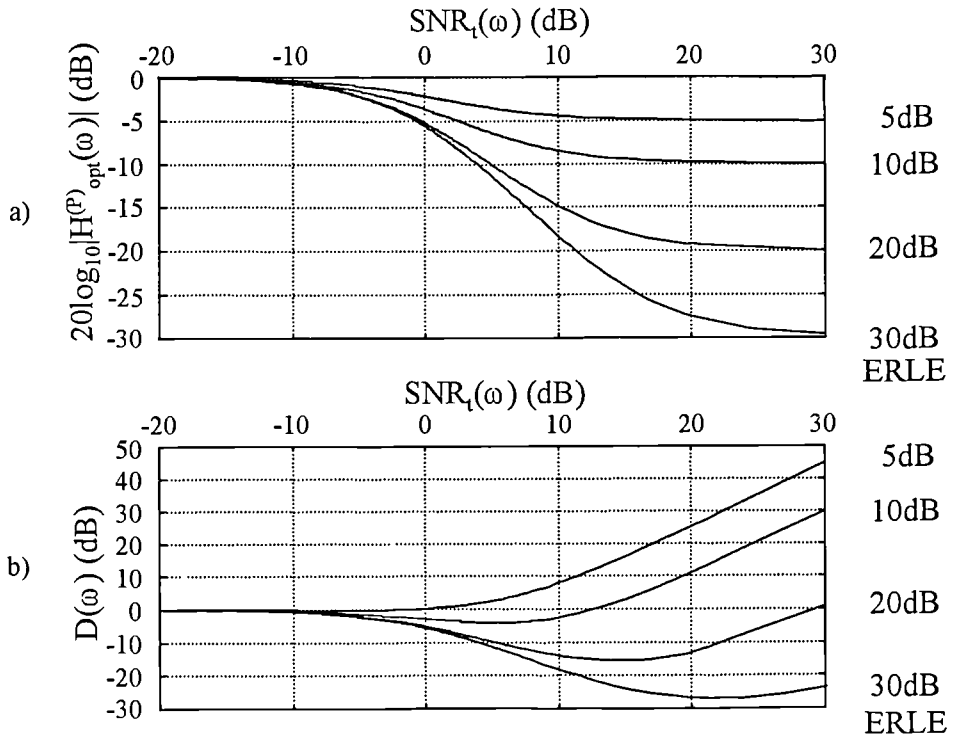
$$H_{opt}^{(P)}(\omega)=\frac{\delta_{ERLE}(\omega)}{\delta_{ERLE}(\omega)+\alpha[n](1-\delta_{ERLE}(\omega))} \tag{5.34}$$

This equation shows that in single talker mode ($\alpha[n]=1$) the attenuation provided by the optimum filter will be equal to the ERLE provided by the echo canceller. During

double talk mode ($\alpha[n]$=0) the frequency response of the optimum filter has a gain of 0dB, for all ω.

$$H_{opt}^{(P)}(\omega) = \begin{cases} \delta_{ERLE} & \text{when } \alpha[n]=1 \\ 1 & \text{when } \alpha[n]=0 \end{cases} \quad SNR(\omega) = \infty dB \qquad (5.35)$$

The magnitude response of the practical filter is determined by the ERLE provided by the echo canceller and the frequency dependent echo power to background plus quantisation power ratio. Figure 5.24a illustrates how the attenuation in single talker mode varies with $SNR_t(\omega)$.



a) Practical attenuation curves, b) difference between practical and ideal curves

Figure 5.24

As with the 'ideal' filter, the attenuation increases with $SNR_t(\omega)$, but now the maximum attenuation is limited to a maximum of ERLE(ω)dB, as described by equation (5.35). Figure 5.24b plots the difference between the practical and 'ideal' filter attenuations, i.e.

$$D(\omega) = 20\log_{10}\left|H_{opt}^{(P)}(\omega)\right| - 20\log_{10}\left|H_{opt}^{(I)}(\omega)\right| \qquad (5.36)$$

where a negative value indicates that the attenuation of the practical filter is greater than that of the ideal filter, and a positive value indicates the opposite. It can be seen that the optimum practical and ideal filters only provide the same attenuation when $SNR(\omega) < -10dB$. As ERLE(ω) increases, the attenuation provided by the optimum practical filter exceeds that provided by the optimum ideal filter, over an increasing $SNR_i(\omega)$ range. Figure 5.24b suggests that for any significant ERLE(ω) (>10-15dB), the attenuation provided by the optimum practical filter will exceed that provided by the optimum 'ideal' filter.

Equation (5.29) showed that for the optimum ideal filter, the PSD of the 'shaped' echo, $P_{ff}(\omega)$, will always be less than that of the near-end signal, $P_{dd}(\omega)$. A similar expression for the optimum practical filter could be obtained by using its frequency response and equation (5.28). However, the resulting expression is not as simple as (5.29). A simpler expression may be obtained by comparing the attenuation produced by the ideal and practical filters. It may be seen from Figure 5.22 that for ERLE(ω)=10dB and $SNR_i(\omega)$=20dB, the optimum 'ideal' filter response is $H_{opt}^{(I)}(\omega) = -20dB$. The use of equation (5.29) shows that at frequency ω, the shaped echo is 10dB less than the near-end signal. For the same conditions, the optimum 'practical' filter has a gain of $H_{opt}^{(P)}(\omega) = -10dB$. Now, because the practical filter is providing 10dB less attenuation than the ideal filter, it can be seen that the resulting shaped echo must have the same power as the near-end signal at frequency ω. Thus when the practical filter is used, an equation for the PSD of the shaped echo, $P_{ff}(\omega)$, in terms of the ideal and practical responses may be determined:

$$P_{ff}(\omega) = P_{dd}(\omega) + \left(\frac{-20\log_{10}\left|G_{opt}^{(I)}(\omega)\right|}{2} \right) - \left(-20\log_{10}\left|G_{opt}^{(P)}(\omega)\right|\right) \qquad (5.37)$$
$$= P_{dd}(\omega) + L(\omega)$$

This function is plotted in Figure 5.25 below.

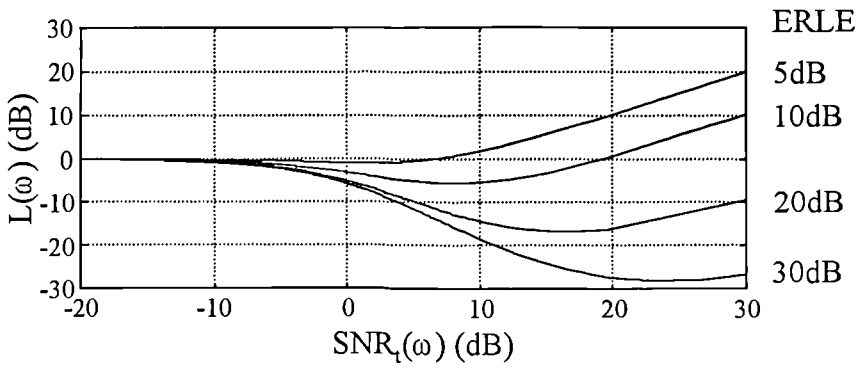


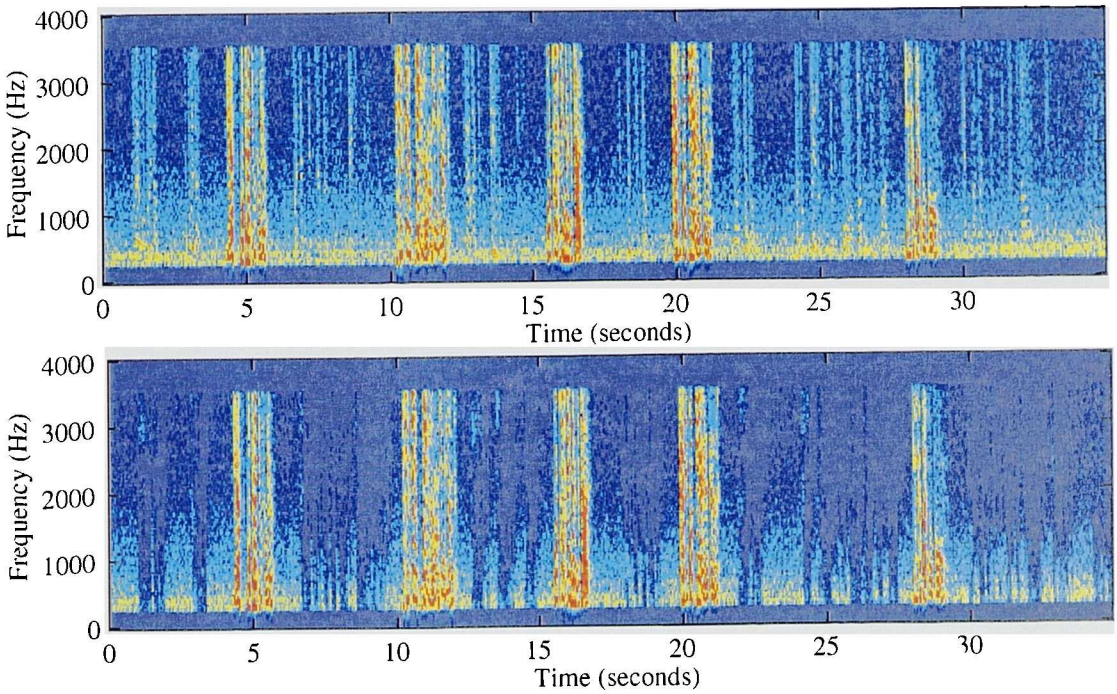
Figure 5.25

This shows that for the practical filter, the attenuation will be such that the PSD of the 'shaped' echo is less than that of the background sounds, because the attenuation is too large. As with the ideal optimum filter, this is likely to lead to noise modulation problems.

5.5.3 Testing of the Echo Shaping Technique

The echo shaping system described above has been implemented as a high-level language simulation, in order to evaluate its performance in terms of echo suppression, noise modulation and near-end talker distortion, when used in a network echo canceller.

Figure 5.26 shows how the unprocessed residual echo compares with the residual echo after 'shaping', when car noise is present at the near-end. In this example, ERL=6dB, ERLE=30dB (i.e. the linear component of the echo is attenuated by 30dB) and the average echo to background noise ratio is approximately 30dB.

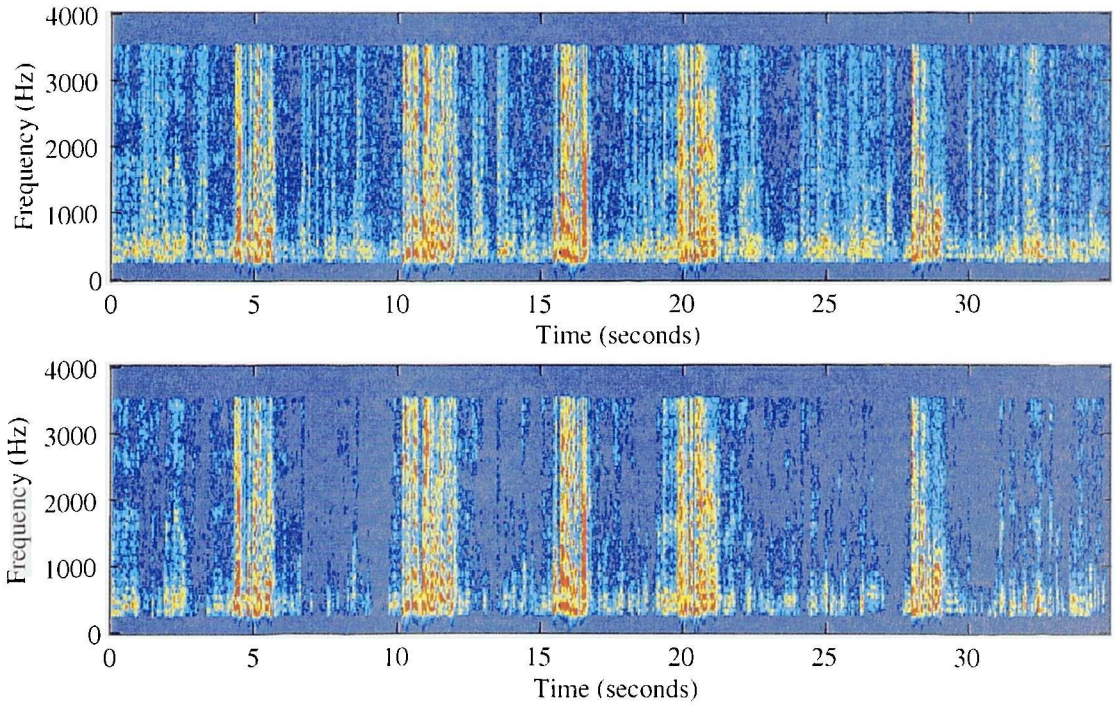


a) before 'echo-shaping', b) after 'echo-shaping'

Figure 5.26

Figure 5.26b shows that after the filtering process, the uncancelled echo is indeed being suppressed, but as expected the background sounds are also being attenuated Measurements with ERL=6dB suggest that, even if the average echo to background ratio is decreased to approximately 0dB, the noise modulation is still evident from the spectrograms. Again, informal listening tests suggest that the noise modulation introduced by this residual control system is audible, although preferable to the presence of unattenuated residual echo.

Figure 5.27 shows how the uncancelled echo is suppressed when multi-speaker noise is present at the near-end, with an average SNR of approximately 30dB.



a) before 'echo-shaping', b) after 'echo-shaping'

Figure 5.27

The level of uncancelled echo is indeed reduced by an amount that depends on the ERLE, but modulation of the background sounds may still be audible. For ERL=20dB and ERLE=10dB the system sounded transparent to background sound but uncancelled echo could still be heard. For ERLE=20dB no echo was audible but modulation of the background was more prominent than before although this is less unpleasant to listen to than unattenuated echo. Although the effects of this type of modulation may not interfere with communication, the talker may still be aware that something is happening because of this 'modulation', which is undesirable.

5.6 Conclusions

This chapter has examined several techniques that may be used to reduce the effects of uncancelled echo. The simplest of these is the single-band, single-threshold centre-clipper. In this clipper, the uncancelled echo may be completely removed by appropriate choice of the clipping threshold. If the input is larger than the threshold, near-end speech for example, then the signal will be passed in distorted form.

This characteristics of the distortion have been examined in terms of the theoretical harmonic distortion obtained using a sinusoidal input, and has been compared to the distortion that was measured when speech was present at the clipper input. It was found that although the distortion ratios calculated for sinusoids and speech do not exactly correspond, due to their different probability distributions, the distortion ratio obtained for sinusoidal inputs may be regarded as the worst case that might be expected for speech inputs. Of the two single-threshold transfer functions considered, the minimum distortion variant causes significantly less corruption of the input signals. However any distortion is unwanted, and therefore the clippers should be disabled during the presence of double-talk and near-end single-talk.

Another disadvantage of the single threshold transfer function, is that its use introduces unwanted noise modulation, which is caused when near-end background sounds present at the clipper input is suppressed, in addition to uncancelled echo. The noise modulation is likely to be less severe when the background noise is non-stationary. Instead of completely removing all signals below the echo suppression threshold, a dual threshold clipper attempts to pass some of the background and hence reduce the effects of noise modulation.

It was found that in order to obtain sufficient echo attenuation, it was necessary to set the background transparency threshold to a value that is less than the standard deviation of the background sounds. However under these conditions, noise modulation of the background sound is introduced, although it is less severe than for the single threshold clipper. While the dual threshold clipper may be successfully used when the background sounds are stationary, for example when certain types of

car noise are present, measurements suggest that when the background sound is time varying suppression of the uncancelled echo cannot be guaranteed. For example, if the background sounds are loud prior to the onset of echo, the background transparency threshold should be increased to follow this variation. If the background sounds are then quiet during the following period of echo (when the threshold is frozen), it is likely that the uncancelled echo will be transmitted without attenuation. Obviously, the degree to which this occurs is dependent upon how the thresholds are controlled, but no matter what method is used, there will always be a trade off between noise modulation and residual echo suppression.

The dual threshold centre-clipping technique has also been investigated in a multi-band configuration. This configuration has the potential to reduce both the noise modulation problems and distortion suffered by near-end talker during double-talk, at the expense of introducing extra delay into the canceller send path. It was found that by using twin-threshold clippers in each of the sub-bands, the uncancelled echo could be suppressed so that it was virtually inaudible without suffering noise modulation. However, the previously mentioned problems that arise when non-stationary noise is present still occur. Compared with the single band configuration, the distortion introduced during double-talk is greatly reduced, even when the band splitting was performed using 8^{th} order FIR filters. It is possible that other types of filter could have been used to perform the band splitting, for example Quadrature Mirror Filters (QMF), which are designed specifically for this type of operation.

The final residual echo control method that was tested is the echo shaping technique. This technique has the advantage that it is very simple to implement (in adaptive filter form) and can effectively suppress the echo, so long as the canceller is providing a minimum ERLE of approximately 10dB. However, the system described here does not shape the echo so that it has the same spectral shape as the near-end signal – in fact the PSD of the shaped echo will always be less than the PSD of the background sounds. Thus, 'holes' are created in the background spectrum at frequencies where the echo has significantly more power than the background noise, and again, noise modulation is likely to be a problem.

The echo shaping technique involves processing the residual echo in the frequency domain using linear techniques. An alternative to both the echo shaping and to non-linear processing in the time domain, is non-linear processing in the frequency domain. For example, the magnitude spectrum of the residual echo can be obtained using an FFT, and each frequency bin could then be processed separately using a centre-clipping function. A similar scheme has been successfully used for enhancement of noisy speech signals [MUND88]. It is possible that using a 256 point DFT/FFT say, would result in suppression of the residual echo whilst giving inaudible distortion during periods of near-end talker. One drawback however, is that the signal delay incurred in performing such a DFT/FFT would considerably exceed the ITU-T limit of 1ms.

For all the residual echo control systems tested here, noise modulation is likely to be a problem. Some echo cancellers attempt to mask the operation of their residual echo control devices by using comfort noise, which is the subject of the next chapter.

6. Comfort Noise

6.1 Introduction

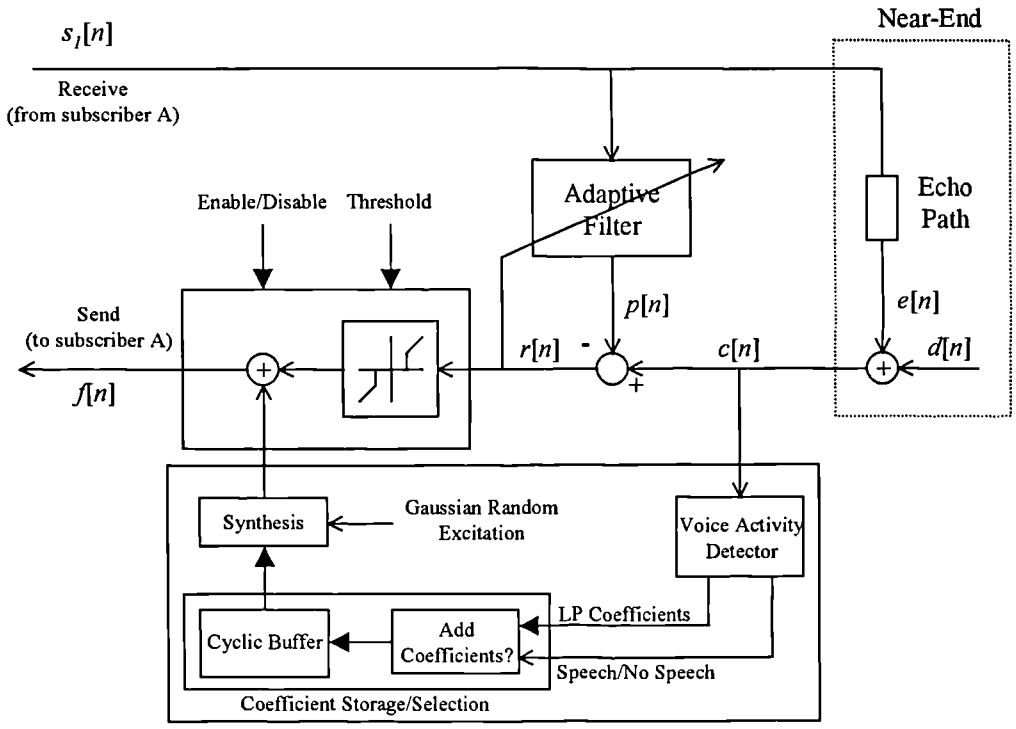
It was seen in the previous chapter that noise modulation is one of the problems associated with residual echo control techniques. It is caused by the removal of circuit and near-end background sounds when a centre-clipper operates to remove uncancelled echo. Noise modulation is a problem in low-noise digital networks because the operation of the clipper is easily noticeable by the far-end talker thus reducing the overall 'naturalness' of the call.

Some echo cancellers attempt to mask noise modulation by injecting white comfort noise into the output of the centre-clipper. Although this does indeed reduce the effects of noise modulation, the comfort noise does not sound very realistic. This chapter describes an enhanced system for use with a single threshold clipper. The enhanced system attempts to mask noise modulation by generating comfort noise that sounds similar to the actual near-end background sounds. This comfort noise has two important characteristics. Firstly, its spectral shape is matched to that of the actual background sounds and secondly, the temporal variation of the spectral shape is similar to the actual background sounds.

The performance of the system has been evaluated in the presence of several different types of background noise and compared to that of the standard centre-clipper, both with and without white comfort noise injection.

6.2 The Dynamic Comfort Noise Injection (DCNI) System

Figure 6.1 shows the block diagram of the combined adaptive filtering and comfort noise injection system.



The Dynamic Comfort Noise Injection (DCNI) System

Figure 6.1

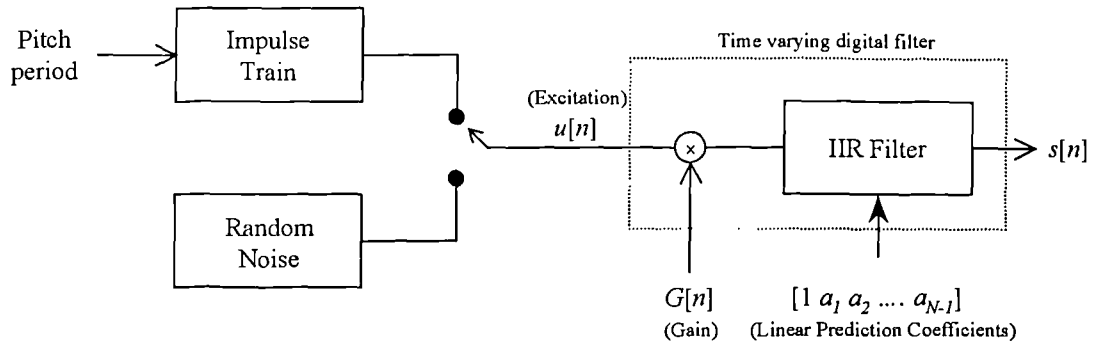
In Figure 6.1 $f[n]$ is the residual error after the centre-clipper and comfort noise injection. All other signal definitions have their previous meanings. In addition to the usual adaptive filter, centre-clipper and control units, the modified echo canceller also makes use of a Voice Activity Detector (VAD) and a comfort noise control/generation unit. The VAD is used to detect periods of time when only background sounds from the near-end are present. During these times, the background spectral characteristics are modelled using the discrete-time all pole model, with linear prediction being used to obtain sets of parameters. Since the

comfort noise is required to mimic only the background sounds and not the echo or near-end speech, the VAD should be sufficiently accurate to discriminate the echo and near-end speech from the background sounds. The GSM Voice Activity Detector [FREE89] is ideally suited for this application because it has been shown to give accurate detection of speech signals [FREE89], [WATS97], in noisy environments where the signal power to background noise power ratio can be very small. Moreover, the linear prediction coefficients of the background sounds may be obtained, at no extra computational cost, because they are calculated as part of the VAD operation. A modified version of this VAD is employed here, and its operation and performance are discussed in more detail in section 6.5.

Once a set of linear prediction coefficients has been obtained, a synthesis filter excited by a white Gaussian random input is used to generate the comfort noise. This is then power scaled to match the average power of the background. The use of this excitation means that the resulting comfort noise does not have any of the periodicity that might have been present in the original near-end background. However, informal listening tests indicate that the comfort noise is sufficiently similar to the background for the far-end talker to be less aware of when it is being added, than when unshaped comfort noise is used.

In addition to having a similar spectral shape to the background sounds, it is also desirable for the comfort noise to have a temporal variation that resembles that of the original background. This is achieved by storing several sets of LP coefficients that represent the spectral shape of the background sounds. The coefficient storage/selection unit, shown in Figure 6.1, is responsible for maintaining these coefficients and deciding which ones should be used in the generation of the comfort noise.

Care must be taken when deciding if a particular set of coefficients that represent the background noise should be added to the buffer. For example, it has been found that it is undesirable to add 'transient' sounds (short duration sounds with strong periodic components), such as ringing telephones, to the coefficient buffer, as this will result



The all-pole model

Figure 6.2

The model consists of a time varying all-pole digital filter that is excited either by a quasi-periodic train of impulses or by a random noise source.

When the model is used for speech production, it is useful to relate the elements of the model to the physical processes that occur during speech production. One type of speech sound, called voiced speech, is generated when pressure waves, caused by air passing through vibrating vocal cords, travel through the vocal tract. The shape of the vocal tract influences the resonances and harmonics that are produced in the final speech. In the all-pole model, this is analogous to a quasi-periodic impulse train, which has similar properties to the pressure waves, passing through a filter that represents the vocal tract. Another kind of speech sound, unvoiced speech, is produced when air passes through a constriction in the vocal tract without causing the vocal chords to vibrate. Again, the shape of the vocal tract influences the sound of the final speech. This is analogous to using random noise as the input to the 'vocal tract' filter in the all-pole model.

In both cases, the shape of the vocal tract is responsible for transforming the excitation into the speech sounds. In the all-pole model the characteristics of the vocal tract are modelled by the all-pole filter, whose transfer function is given by:

$$H(z) = \frac{G}{1 - \sum_{i=1}^{P} a_i z^{-i}} = \frac{G}{A(z)} \qquad (6.1)$$

where

$H(z)$ = all pole transfer function

$A(z) = 1 - \sum_{i=1}^{P} a_i z^{-i}$

G = gain parameter

a_i = linear prediction coefficients

P = the prediction order

The filter which implements $H(z)$ is known as the synthesis filter and the filter which implements $A(z)$ is known as the inverse or analysis filter. Once the linear prediction coefficients have been calculated for a short segment of waveform $s[n]$, the corresponding excitation can be obtained by filtering the segment using the resulting analysis filter. Conversely, the speech may be reconstructed by passing the excitation through the synthesis filter $H(z)$, as shown in Figure 6.2.

Equation (6.1) for $H(z)$ may be applied to an excitation $U(z)$ and inverse transformed to yield an output waveform $s[n]$ given by:

$$s[n] = G.u[n] + \sum_{i=1}^{P} a_i s[n-i] \qquad (6.2)$$

where

$s[n]$ = the current sample

$s[n\text{-}i]$ = the previous samples

$u[n]$ = the current excitation sample

This equation shows that the current sample $s[n]$ may be obtained from a linear combination of the past samples $s[n\text{-}i]$ and the excitation $u[n]$, i.e. the current sample is, to some extent, predictable from its past values. The relative weightings, a_i, given to the past speech samples are hence known as the Linear Prediction Coefficients, and when applied with the gain, G, to speech production, they form a recursive filter,

whose frequency response is the spectral envelope of the speech. For this reason, it is expected that when $s[n]$ is a non-speech signal, for example car noise, the use of linear prediction will result in a set of prediction coefficients that represent the spectral envelope of the input.

In the echo canceller comfort noise application, both the frequency response of the synthesis filter and the excitation, $u[n]$, are of interest. By attempting to match the frequency response of the synthesis filter to that of the actual background sounds and by using an appropriate excitation it is possible to generate synthetic background sounds. The frequency response of the synthesis filter and the excitation are discussed in sections 6.3.4 and 6.3.5 respectively, but before this, a technique that is commonly used to compute the predictor coefficients is presented.

6.3.2 Computation of the Linear Prediction Coefficients

The linear prediction coefficients are calculated by minimising the squared error between the actual waveform $s[n]$ and a predicted waveform $\hat{s}[n]$. This error is defined by:

$$e[n] = s[n] - \hat{s}[n] = Gu[n] \tag{6.3}$$

where

$$\hat{s}[n] = \sum_{k=1}^{P} a_i s[n-k] \tag{6.4}$$

Thus the total squared error, over the time interval $t_1<n<t_2$, is given by

$$E = \sum_{n\ t_1}^{t_2} e^2[n] = \sum_{n=t_1}^{t_2} \left[s[n] - \sum_{k\ 1}^{P} a_k s[n-k] \right]^2 \tag{6.5}$$

The minimum squared error is obtained by setting

$$\frac{\partial E}{\partial a_i} = 0 \qquad \text{for } 1 \le i \le P \tag{6.6}$$

in equation (6.5) and this yields the following set of equations:

$$\sum_{k=1}^{P} a_k \sum_{n=t_1}^{t_2} s[n-i]s[n-k] = \sum_{n=t_1}^{t_2} s[n]s[n-i] \qquad \text{For } 1 \le i \le P \tag{6.7}$$

This set of P linear equations is known as the 'normal' equations and solving them yields the linear prediction coefficients (a_i).

6.3.3 The Autocorrelation Method

The normal equations given in (6.7) are often solved using either the autocorrelation method, or the covariance method. Only the autocorrelation method is considered here because it is particularly amenable to economical solution and always yields a stable result. The covariance technique, on the other hand, does not always guarantee a stable synthesis filter [MAKH76].

In the autocorrelation technique, the total error E is minimised over the interval $-\infty < n < \infty$. The normal equations may now be simplified to:

$$\sum_{k-1}^{P} a_k R[k-i] = R[i] \qquad \text{For } 1 \le i \le P \tag{6.8}$$

where $$R[i] = \sum_{n=-\infty}^{\infty} s[n]s[n+i] \tag{6.9}$$

Equation (6.9) is recognised as the autocorrelation function of $s[n]$. In practice, for speech, $s[n]$ is usually analysed over a small time interval because a representation of the changing properties of the waveform is needed. With speech this interval is usually in the range 10-20ms as, over this range, the speech may be assumed to be quasi-stationary. Thus over this interval, the signal $s[n]$ is effectively multiplied by a

'window' function $w[n]$, with the samples lying outside the window being set to zero. The windowed speech samples may be written as:

$$s_w[n] = s[n] \cdot w[n] \tag{6.10}$$

Now, the autocorrelation function is given by:

$$r[i] = \sum_{n=0}^{W-1-i} s_w[n] s_w[n+i] \tag{6.11}$$

where the window has length W. The normal equations may now be written as:

$$\begin{aligned} a_1 r[0] + a_2 r[1] + \cdots\cdots + a_P r[P-1] &= r[1] \\ a_1 r[1] + a_2 r[0] + \cdots\cdots + a_P r[P-2] &= r[2] \\ &\vdots \\ a_1 r[P-1] + a_2 r[P-2] + \cdots + a_P r[0] &= r[P] \end{aligned} \tag{6.12}$$

$$\begin{bmatrix} r[0] & r[1] & \cdot & \cdot & r[P-1] \\ r[1] & r[0] & & & r[P-2] \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ r[P-1] & r[P-2] & & & r[0] \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ a_P \end{bmatrix} = \begin{bmatrix} r[1] \\ r[2] \\ \cdot \\ \cdot \\ r[P] \end{bmatrix} \tag{6.13}$$

or, in matrix form:

$$\Phi.\mathbf{A} = \mathbf{r} \tag{6.14}$$

where

Φ = the time-averaged autocorrelation matrix,

$\mathbf{A}$ = the prediction coefficient vector,

$\mathbf{R}$ = the autocorrelation vector.

This system can be solved using various techniques, but one of the most common methods is Durbin's recursive algorithm [MAKH76], [DURB60], which exploits the

Toeplitz and symmetrical properties of the autocorrelation matrix Φ to compute the prediction coefficients. The algorithm is described by the following equations:

$$E_0 - r[0] \tag{6.15}$$

$$k_i = \frac{\left[r[i] - \sum_{j=1}^{i-1} a_j^{(i-1)} r[i-j] \right]}{E_{i-1}} \tag{6.16}$$

$$a_i^{(i)} = k_i \tag{6.17}$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)} \qquad 1 \le j \le i-1 \tag{6.18}$$

$$E_i = (1 - k_i^2) E_{i-1} \tag{6.19}$$

Equations (6.16) to (6.19) are iterated to find **A,** where $1 \le i \le P$ and P is the desired order of the synthesis filter. The terms denoted as k_i in equation (6.16) are known as the reflection, or partial correlation coefficients, and are a useful 'by-product' of Durbin's algorithm. Note that k_i is equal to a_i for an i^{th} order linear prediction analysis. Although the reflection coefficients are not used directly in the synthesis of the comfort noise, they are used to control the storage of coefficient sets. This will be discussed further in chapter seven.

6.3.4 Frequency Response of the Synthesis Filter

Once the linear prediction coefficients have been calculated by analysing a short segment of $s[n]$, they may be used in the synthesis filter. This section briefly discusses how the frequency response of the resulting filter compares with the discrete Fourier transform (DFT) of the analysed speech segment. The relationship between the two is important because the comfort noise is to have a similar spectral shape to the original background sounds.

It may be shown that by re-writing equation (6.2) in terms of the error $e[n]$ and then taking the z-transform, the power spectrum of the original signal, $P(\omega)$, is given by:

$$P(\omega) = \frac{\left|E\left(e^{j\omega}\right)\right|^2}{\left|A\left(e^{j\omega}\right)\right|^2} \qquad (6.20)$$

where

$\left|A\left(e^{j\omega}\right)\right|^2$ = the power spectrum of the inverse filter characteristic

$\left|E\left(e^{j\omega}\right)\right|^2$ = the power spectrum of the error

It may also be shown that, by using equation (6.1), the power spectrum of the signal estimated by using linear predication, $\hat{P}(\omega)$, is given by

$$\hat{P}(\omega) = \left|H\left(e^{jw}\right)\right|^2 = \frac{G^2}{\left|A\left(e^{jw}\right)\right|^2} \qquad (6.21)$$

By comparing (6.20) and (6.21) it can be seen that the error power spectrum is being modelled by a flat spectrum with power equal to G^2, i.e. the actual error $e[n]$ is approximated by another signal which is white. By using Parseval's theorem the error spectrum can be related to the time domain mean square error:

$$E = \sum_{n=t_1}^{t_2} e[n]^2$$

$$= \frac{1}{2\pi}\int_{\pi}^{\pi}\left|E\left(e^{j\omega}\right)\right|^2 d\omega \qquad (6.22)$$

$$= \frac{G^2}{2\pi}\int_{-\pi}^{\pi}\frac{P(\omega)}{\hat{P}(\omega)}d\omega \qquad (6.23)$$

It may be shown that the minimum mean squared error is equal to G^2 [MAKH76] and therefore

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{P(\omega)}{\hat{P}(\omega)}d\omega = 1 \qquad (6.24)$$

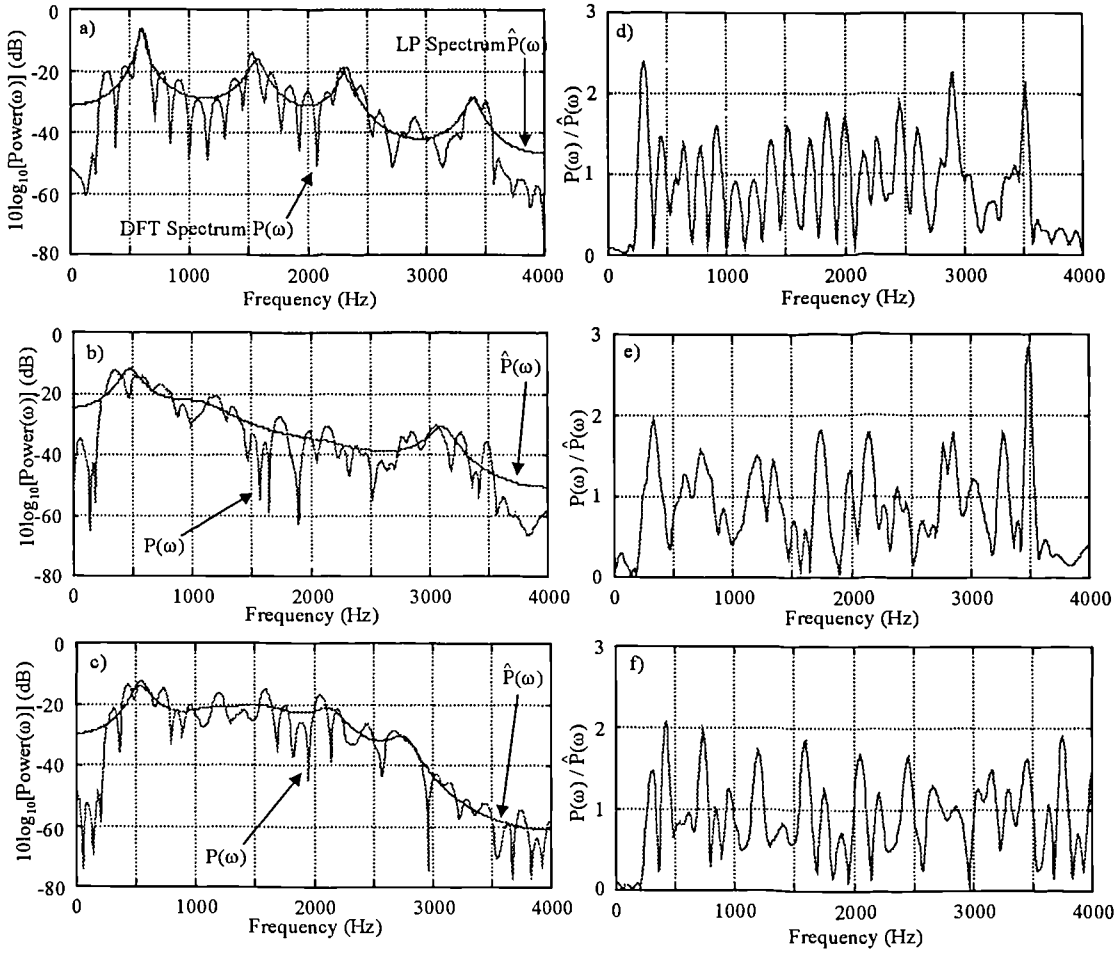
$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|E\left(e^{j\omega}\right)\right|d\omega = G^2 \qquad (6.25)$$

Equation (6.23) suggests that the minimisation of the mean squared error is equivalent to the minimisation of the integrated ratio of the actual and predicted spectra. From equation (6.24) it follows that when the mean square error is minimised the mean value of the ratio of the actual spectrum to the predicted spectrum is unity, and thus the average value of the error spectrum is equal to G^2. In other words, at some frequencies the predicted spectrum may be larger than the actual spectrum, and at other frequencies may be smaller than the actual spectrum. However, minimisation of the mean square error ensures that the mean value of $P(\omega)/\hat{P}(\omega)$ over all ω is unity, and therefore on average, the spectral matching should be uniform across the whole frequency range. This property is important because the spectrum of the comfort noise is required to be similar to that of the background sounds at all frequencies.

The general relationship between the spectrum obtained using linear prediction and that given by the discrete Fourier transform is illustrated in Figure 6.3 using three different sound types, which are believed to be representative of the types of background that are most likely to be encountered. These are i) multispeaker noise that consists of several people talking simultaneously (non-stationary), ii) car noise which was recorded in a car that was travelling at a constant speed (stationary), and iii) babble noise that is a mixture of stationary and non-stationary sounds. The properties of these noise types will be discussed in more depth later. In all cases, the sounds are limited to the bandwidth of 300 to 3.5 kHz.

For each of these sounds, the LP spectrum was obtained from a 160 sample segment that was Hann windowed, before analysis using a 10^{th} order model. The DFT

spectrum was computed from the same segment, but was first zero padded to 1024 samples.



a) to c) $P(\omega)$ and $\hat{P}(\omega)$ for voiced speech, car noise and babble noise,

d) to f) $P(\omega)/\hat{P}(\omega)$ for voiced speech, car noise and babble noise.

Figure 6.3

It may be seen that for these three background sounds (and this is, in fact, more generally observed to be the case using LP), $\hat{P}(\omega)$ is more accurate, in predicting the DFT spectrum, at the peaks of $P(\omega)$ than at the 'troughs' between the peaks. When 10^{th} order analysis is used to model these sounds, $\hat{P}(\omega)$ is a good estimate of the spectral envelope of $P(\omega)$ between 300 and 3.5kHz. The relationship between $P(\omega)$ and $\hat{P}(\omega)$ is illustrated in Figure 6.3d to Figure 6.3f and show that across the whole frequency range $P(\omega)/\hat{P}(\omega)$ varies in accordance with equation (6.24).

It should also be noted that the 10^{th} order model used here, is generally unable to accurately estimate $P(\omega)$ in the stop bands below 300 Hz and above 3500 Hz – Figure 6.3 shows that $\hat{P}(\omega)$ is much too large. This is because there is very little energy contained within the stop bands and therefore the contribution to the total error is very small. It may be demonstrated that by increasing the model order, $\hat{P}(\omega)$ becomes a more accurate estimate of $P(\omega)$, i.e. the spectrum given by linear prediction tends towards the DFT spectrum. The implications of this will be discussed in section 6.3.5, when considering the synthesis filter excitation signal.

6.3.5 The Excitation Signal u[n]

Once the prediction coefficients have been computed, the original speech may be applied to an inverse filter and this yields the excitation $u[n]$. In the all-pole model of Figure 6.2, the excitation is impulse-like for voiced signals and random for unvoiced signals. Prior to the generation of comfort noise the properties of the excitation must be considered, as the actual background sounds will contain components that are similar to the voiced and unvoiced components of speech. Three possibilities may be considered:

i) To record the excitation obtained by inverse filtering,

ii) To calculate pitch period of the excitation and re-synthesise the background sounds using impulses for voiced parts and random excitation for unvoiced parts (i.e. implement a full linear predictive speech coder), and,

iii) To use random excitation only.

The first option is inappropriate in this application – the actual environmental sounds could be recorded directly to create the same effect. It is probable that the far-end talker would be aware of the switching between the recorded and actual background sounds when using this method, because the comfort noise would be identical to previous environmental sounds. In effect, the system would be generating listener

echo of the background sounds. These comments also apply to the second option, because the 'speech coder' would be replicating the background sounds in a similar fashion. It is therefore not desirable for the comfort noise to sound identical to the original.

In option three the excitation is white Gaussian noise. Informal listening tests suggest that 'speech' synthesised using this excitation still sounds speech-like, even though all the periodic components have been removed. In fact, the resulting 'speech' sounds like whispering because only the spectral envelope remains from the original. When applied to background sounds with periodic components the effect is the same - the comfort noise will sound similar, but not identical to the original background. For background sounds, such as car noise, where the power of the periodic components is small compared with the power of the non-periodic components, the synthetic background is almost indistinguishable from the original.

As mentioned in the previous section, the estimated spectrum $\hat{P}(\omega)$ is much larger than $P(\omega)$ between 0 to 330Hz and 3.5kHz to 4kHz, because the near-end signal contains very little energy at these frequencies. Generally, we expect that the energy in these low and high frequency bands will be removed from the source signal by the local loop circuit and by the codec ADC. The use of white excitation will therefore result in the comfort noise having too much energy in these ranges. It cannot be certain that the far-end local loop and codec DAC will remove this energy and this will result in the background sounds being perceived, by the far-end user, as unrealistic. For the background sounds used here, the unwanted energy is more audible at low frequencies, because the background sounds have more energy at low frequencies than at high frequencies. Clearly, this is a problem because it will be obvious to the far-end subscriber when the comfort noise is being injected.

Ideally, a spectral estimation method should be used that is able to model the stop-band regions, because the effectiveness of the comfort noise system would then be independent of the channel frequency response. This could be accomplished by increasing the LP model order but would lead to all of the spectral peaks being

modelled including those resulting from periodic processes. In speech coding the use of a high order analysis is undesirable because the vocal tract/synthesis filter will begin to model the excitation, rather than that of the vocal tract alone. It was initially thought that for the comfort noise application, increasing the analysis order dramatically would be beneficial for two reasons. Now, not only would the stop-band regions be modelled more accurately, but also, the synthesis filter would model the periodic components of the actual excitation that are not present in the noise excitation. It was hoped that this would result in more realistic comfort noise, however experimentation showed that this was not the case.

One way in which the estimation procedure could be improved is to derive the prediction coefficients for the comfort noise from a spectral envelope obtained directly from the DFT power spectrum. This would involve picking the spectral peaks and then calculating a best-fit envelope using a technique such as cubic-spline interpolation. Taking the inverse DFT of this envelope gives the autocorrelation coefficients of the power spectrum, which can be processed using Durbin's algorithm to calculate a set of high-order predictor coefficients. The resulting spectrum would then be a better approximation to the spectral envelope over the whole frequency range, including the stop-bands. The disadvantage of this procedure is that it is more complex from the processing point of view – the background LP coefficients are no longer obtained 'for free' from the VAD calculations.

For the simulation results presented in this chapter, the excitation has been bandlimited using a filter whose frequency response is identical to the filter which bandlimited the near-end signal. The simulations are therefore not entirely realistic since this information might not be known in practice.

6.4 Coefficient Storage & Selection

In the dynamic comfort noise injection system proposed in this thesis, the linear prediction coefficients of frames that are classified as background sounds are added to a cyclic buffer every 10ms and these coefficients are used to shape the spectral

envelope of the comfort noise. The coefficients are obtained from the VAD, and their calculation will be discussed in section 6.5. As new sets of coefficients are added the oldest in the buffer are overwritten, so that the spectrum of the comfort noise can always approximate the characteristics of the latest background sounds.

6.4.1 Adding Frames To The Buffer

Coefficients are added to the buffer that can store a total of N sets, at a position indicated by a 'fill pointer'. This pointer is incremented every time a new set is added, as shown in Figure 6.4 below.

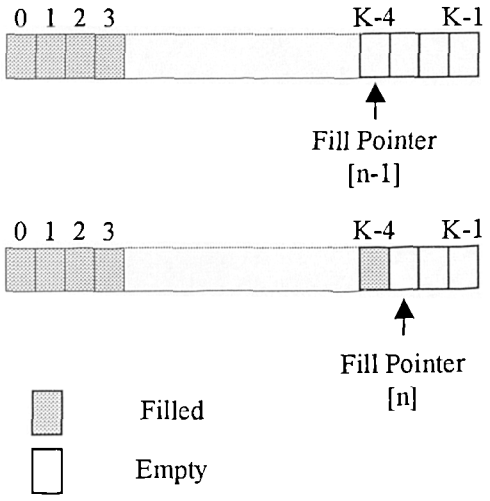


Figure 6.4

When a set of coefficients is stored at position (K-1) the buffer is full, so the pointer wraps-around to the beginning and the next set is stored at position 0. Thus the buffer is cyclic, and this ensures that the comfort noise can approximate the background by storing coefficients that represent the latest spectral estimates. Currently the maximum number of coefficient sets is set to 200, which corresponds to two seconds of comfort noise generated in 10ms blocks and at a sampling frequency of 8kHz. The storage space required for 200 sets of coefficients is modest. Assuming that each of the ten coefficients is stored as a four byte floating point number the total memory required is only 200×10×4=8000 bytes (=7.8125kB).

6.4.2 Coefficient Selection for Playback

Suppose the last set of coefficients were added in at position M-1, where positions M to N-1 are as yet empty. If comfort noise is required, the sets of coefficients are used in reverse order starting from the last set at M-1.

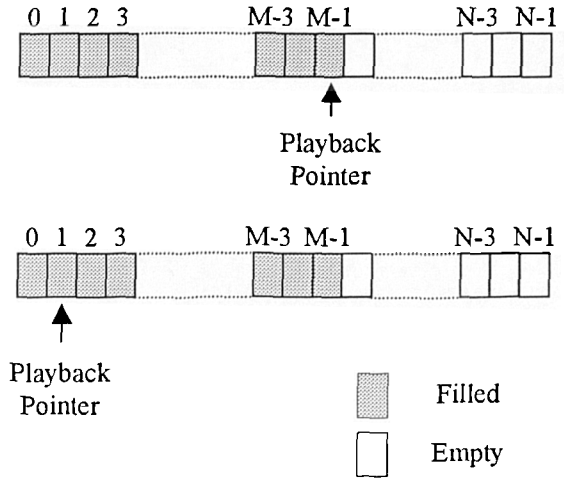


Figure 6.5

As playback continues, the pointer moves from the fill pointer position M-1 towards position 0, and on reaching this point, wraps around to position M-1. This process continues until comfort noise is no longer required. Note that if all N-1 buffer elements are full, the 'playback' of all N-1 blocks may occur in a similar manner.

The number of sets to be stored is a trade off between several factors. From an implementation point of view it helps to have as small a buffer as possible because this will reduce the required storage space. In addition, it means that the stored characteristics will be of background sounds that will have recently occurred. If the buffer is too long, then the stored characteristics may no longer be appropriate because the background sounds could have changed. However, using a large buffer reduces the frequency at which the same set of coefficients must be used – undesirable 'looping' of the comfort noise is highly audible for short buffer lengths.

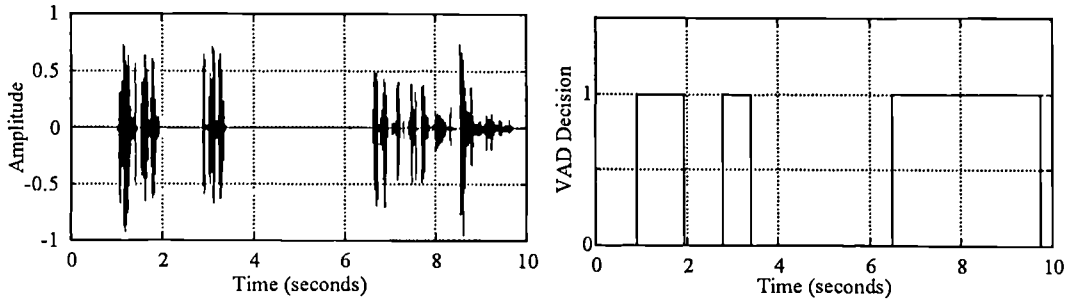
Informal listening tests have led to the conclusion that a length of 200 is a satisfactory compromise between the above factors. If the comfort noise is listened-to in isolation, looping is sometimes audible but depends on the type of background noise. For example looping in car noise is much less noticeable than in multi-speaker noise because the spectral envelope changes much less than in car noise. Reverse playback was chosen to reduce the possibility of the comfort noise being perceived as a distorted replay of the true background sounds.

6.4.3 Comfort Noise Synthesis

Comfort noise is initially generated using equation (6.2) with unity gain, and a white Gaussian excitation that has been bandlimited to 330-3.5kHz. After generation, the comfort noise is scaled so that its rms power is similar to that of the original environmental sounds, which is estimated during periods when the VAD indicates that only background sounds are present. When comfort noise is required, it is faded in, whilst the near-end signal is faded out over a duration of 40 samples (5ms). This is done to reduce the possibility of the far-end talker hearing clicks caused by waveform discontinuities between the synthetic and original background. This process is performed in reverse when comfort noise is no longer required. During periods of continuous comfort noise injection, the spectral shape of the comfort noise is altered by using a different set of prediction coefficients every 10ms, as described in the previous section.

6.5 Voice Activity Detection

The times during which echo and/or near-end speech are present must be detected, so that only the background sounds are used to shape the spectrum of the comfort noise. This may be achieved by using a Voice Activity Detector (VAD), which is a device that indicates when speech signals are present at its input. Figure 6.6 shows the behaviour of an ideal VAD when the background is silence.



a) Input waveform, b) VAD Decision where Speech Detected = 1

Figure 6.6

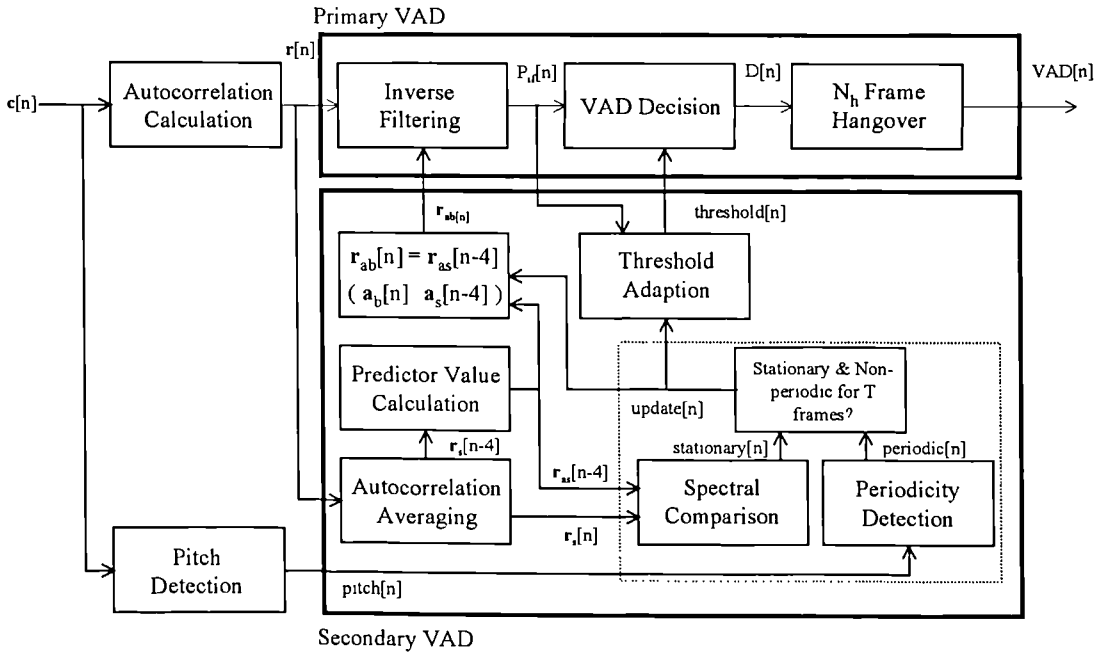
In this ideal VAD, the output is one when speech is detected and zero at all other times.

In the absence of background sounds energy thresholding may be used to make an activity decision. If the energy of the current frame is greater than some threshold then speech is deemed to be present, otherwise there is silence. However, the accuracy of this technique deteriorates when background sounds are present, as there is an increasing likelihood that the background noise may be classified as speech, or that low amplitude speech classified as background noise. For very high power background signals, all such sounds would be classified as speech. Clearly, a simple thresholding process is inadequate for applications where accurate speech detection is required even in the presence of loud background sounds.

In GSM mobile telephony [FREE89], a VAD is used to detect the presence of speech so that the total number of bits used to encode the microphone signal may be reduced. In order to prevent any speech from being mutilated, the VAD used in GSM must be able to cope with the significant levels of background noise that are often encountered in a mobile environment. The GSM VAD improves on the elementary energy thresholding technique by emphasising components of the speech that are more significant that those of the background. The energy of the resulting signal is then compared to an adaptive threshold to make the activity decision. Once the VAD has correctly indicated the presence of speech, it is likely that the energy of the emphasised speech will fluctuate, and for short periods will fall below the threshold

level. During these periods, the VAD will incorrectly indicate the absence of speech. To prevent this incorrect classification, a hangover is added to the VAD decision. It is only if the VAD decision is silence for several consecutive frames that the final VAD decision is silence.

The block diagram of the VAD is shown below in Figure 6.7 [WATS98]. In this implementation, the VAD operates on input frames of length 160 samples, which overlap by 80 samples. In other words, a voiced/unvoiced decision is made every 10ms.



Voice Activity Detector used for Comfort Noise Injection

Figure 6.7

Figure 6.3 shows that the overall operation of the VAD is split between two functional units – the primary and secondary VADs. As described previously, the primary VAD makes the overall activity decision, whilst the secondary VAD is responsible for controlling emphasis of the speech relative to the background and adaptation of the decision threshold level. The following sections discuss the overall operation and performance of the VAD in terms of the primary and secondary detectors.

6.5.1 The Primary VAD

The primary VAD consists of three functional parts that are used to make the overall activity detection decision. In Figure 6.3 these are labelled as the *Inverse Filtering* process which emphasises the speech components relative to the background sounds, the *VAD Decision* which makes the activity decision by comparing the energy of the emphasised signal to a threshold, and the N_h *Frame Hangover* which prevents premature release of the voiced detection condition.

6.5.1.1 Inverse Filtering

One way in which the components of the speech that are more significant than the background sounds may be emphasised is to use a linear prediction inverse filter whose frequency response is the inverse of the background noise spectrum. To understand how this works, consider the inverse filtering of a speech waveform, using linear prediction coefficients calculated using Durbin's algorithm. Now, H(z) represents the spectral envelope of the speech, and hence, 1/H(z) = A(z) represents its inverse spectral envelope. Thus, inverse filtering attempts to flatten the spectrum by removing the spectral envelope from the speech.

Now, suppose that the speech signal contains some background noise and that A(z) now represents the inverse spectrum of the background. After inverse filtering, frequencies where the background sounds have greater power will be attenuated more than frequencies that have less power. The inverse filter attempts to emphasise components of the input speech that are more significant than the background sounds. Alternatively, the filtering process can be viewed as attempting to improve the SNR by attenuating frequencies where the power of the background sounds is large. Thus a larger proportion of the energy for each frame of the filtered signal ($P_{if}[n]$, in Figure 6.3) will be due to speech rather than background sounds.

Although the energy, $P_{if}[n]$, could be calculated from a frame of input speech that has been passed through the inverse filter, there is a more efficient way to compute its value within the VAD. It can be shown [WATS98] that the energy may be calculated

by using the autocorrelation coefficients of the current frame, $\mathbf{r}[n]$, and the autocorrelated LP coefficients, $\mathbf{r}_{ab}[n]$, of the background sounds which are already available in the secondary VAD. The energy is now calculated using the following equation

$$P_{if}[n] = r_{ab}[n][0] \cdot r[n][0] + 2\sum_{i=1}^{P} r_{ab}[n][i] \cdot r[n][i] \tag{6.26}$$

where

$$r[n][i] = \sum_{k\ 0}^{W-1-i} s_w[n][k] s_w[n][k+i] \tag{6.27}$$

$$s_w[n][k] = s[nW+k]w[k]$$

are the autocorrelation coefficients of the current frame n, $s[k]$ is the input speech, $w[k]$ is a window function, and

$$r_{ab}[n][i] = \sum_{k=0}^{P-i} a_b[n][k] a_b[n][k+i] \tag{6.28}$$

is the autocorrelation function of the background sound LP coefficients, $\mathbf{a}_b[n]$, which is calculated by the block labelled *Predictor Values Calculation* in Figure 6.7.

6.5.1.2 VAD Decision & Hangover

Once the frame energy $P_{if}[n]$ has been calculated an initial VAD decision is made by comparing $P_{if}[n]$ with a threshold that is determined by the secondary VAD, i.e.

$$\begin{aligned} &\text{if} \quad (P_{if}[n] > \text{threshold}[n]) \quad D[n]=\text{TRUE} \\ &\text{else} \qquad\qquad\qquad\qquad\quad D[n]=\text{FALSE} \end{aligned} \tag{6.29}$$

If the initial decision D[n] indicates the presence of speech, then the final decision should also indicate this. This is because, in both the GSM and comfort noise

applications, it is better to misclassify background noise as speech rather than speech as noise.

It is likely that the energy, $P_{if}[n]$, will fluctuate and for short periods fall below the threshold level whilst speech is still present. In the GSM application, these short sections of low energy speech would not be transmitted if D[n] was used alone to make the activity decision, this would lead to mutilation of the speech at the receiver. To minimise the frequency of this incorrect classification, a hangover is added to the VAD decision. Then, it is only if the VAD decision is background for N_h consecutive frames that the final VAD decision is background.

6.5.2 The Secondary VAD

The secondary VAD is responsible for setting the energy thresholding level, threshold[n], and the background noise model, $\mathbf{r}_{ab}[n]$, used in the inverse filtering process. These are both updated when it is known that only background sounds are present. The operation of the secondary VAD is dependent upon the detection of such background sounds and makes use of the assumptions that the background sound is both stationary and non-periodic. In reality, these assumptions are more or less reasonable depending upon the type of background sound present. For example, car noise is likely to be more stationary and non-periodic than multi-speaker noise. The presence of background sounds is detected here by using a *Spectral Comparison* test and a *Periodicity Detection* test to identify when these assumptions are satisfied.

It is very important that the secondary VAD should only indicate the presence of background noise alone when this is true, otherwise the inverse filtering process in the primary VAD will be corrupted by the use of an incorrect background model. The secondary VAD declares 'background' only after several strict conditions have been met, and this means that only a small proportion of background sound frames will be detected – many frames are falsely identified as 'speech'. In the GSM system, the VAD is used to decide whether to transmit the current frame (if it is speech) or not (if it is noise) and so accurate detection of large portions of silence is required. This

cannot occur using the secondary VAD alone. Superior performance is obtained using the combined primary and secondary VAD system. Similarly in the comfort noise application, as many background coefficients sets as possible need to be captured and for this reason the secondary VAD, again, cannot be used on its own.

The following sections explain the operation of the spectral comparison and periodicity tests that are used to detect the presence of background sounds. The threshold adaptation algorithm is also discussed.

6.5.2.1 Spectral Comparison

The spectral comparison test uses an inverse filtering process, that is similar to the one used in the primary VAD, to estimate whether the spectrum of the input is changing significantly from frame to frame. In this case, the filter attempts to whiten the average spectrum of $c[n]$ by using a set of coefficients derived from $c[n]$ several frames earlier. The ratio of the filter input energy to output energy is then used as an estimate of spectral stationarity by examining its value over several frames. As in the primary VAD, the energy of the filtered signal is calculated directly from the autocorrelations, with equation (6.30) being used to calculate the desired energy ratio, $dm[n]$:

$$dm[n] = \frac{\text{Energy of filtered signal[n]}}{\text{Energy of original signal[n]}} \qquad (6.30)$$

$$= \frac{r_{as}[n-4][0]\cdot r_s[n][0] + 2\sum_{k=1}^{N-1} r_{as}[n-4][k]\cdot r_s[n][k]}{r_s[n][0]}$$

where 0<dm[n]<1. In equation 6.30 $\mathbf{r}_s$[n] is a measure of the average autocorrelation function calculated over the previous 4 frames:

$$r_s[n][i] = \sum_{k=0}^{3} r[n-k][i] \qquad \text{for } 0 \le i \le 10 \qquad (6.31)$$

and where $\mathbf{r}_{as}$[n] is given by

$$r_{as}[n][i] = \sum_{k=0}^{P-i} a_s[n][k] a_s[n][k+i] \qquad (6.32)$$

with

$$\mathbf{a}_s[n] = \Phi_s[n]^{-1} \cdot \mathbf{r}_s[n] \qquad (6.33)$$

When stationary background sound has been present for several frames, the frequency response of the inverse filter given by $\mathbf{a}_s[n\text{-}4]$ is approximately the same as that given by $\mathbf{a}_s[n]$. Therefore successive values of the energy ratio, dm, will be approximately the same because $\mathbf{a}_s[n\text{-}4]$ is used to whiten the current input. However, when a non-stationary signal such as speech is present, successive values of dm will be different because $\mathbf{a}_s[n\text{-}4]$ and $\mathbf{a}_s[n]$ yield inverse filters with different frequency responses.

The input is deemed stationary if two successive energy ratios are approximately equal, i.e.

$$\begin{array}{ll} \text{if} \quad |dm[n] - dm[n-1]| < \Delta_s & \text{stationary}[n] = \text{TRUE} \\ \text{else} & \text{stationary}[n] = \text{FALSE} \end{array} \qquad (6.34)$$

where Δ_s is a constant, equal to 0.05 in this implementation.

6.5.2.2 Periodicity Detection

When the input contains a quasi-periodic signal it is likely that speech rather than background sound is present. The periodicity test uses an estimate of the fundamental frequency (pitch) of the input signal to decide whether the input contains strong periodic components.

Pitch estimation is often accomplished by examining the peaks that are produced in the input autocorrelation function when periodic signals are present. The pitch estimation scheme employed here is slightly different in that the autocorrelation

function of the linear prediction residual that is obtained after using a 1kHz low-pass filter on the input signal is calculated [CHOI97]. The fundamental frequency of human voiced speech is almost invariably less than 1kHz, and so the filtering helps to reduce the effects of background sound that has frequencies above 1kHz. It is believed that this makes the VAD more accurate as a speech detector because background sounds that contain periodic components above 1kHz are less likely to be falsely detected as speech.

When voiced speech is present the pitch values vary with time, but over consecutive frames, the difference in pitch will be small. When non-periodic background sounds, or unvoiced-speech are present, the autocorrelation function will not contain any significant peaks and the pitch value will vary randomly from one frame to the next. Therefore, if the difference in pitch of several successive frames is small, the input is more likely to be periodic than non-periodic. The following equations describe an algorithm that is used to compare the pitch values from consecutive frames:

$$P_{\max} = \max(pitch[n] \quad pitch[n-1]) \tag{6.35}$$

$$P_{\min} = \min(pitch[n] \quad pitch[n-1]) \tag{6.36}$$

$$diff = \min\left[rem\left(\frac{P_{\max}}{P_{\min}}\right) \quad P_{\min} - rem\left(\frac{P_{\max}}{P_{\min}}\right)\right] \tag{6.37}$$

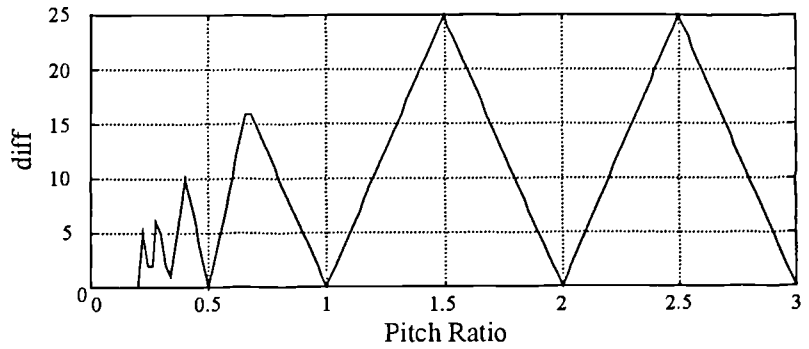
$$\begin{aligned} &if \quad (diff < \Delta_p) \quad periodic'[n] = TRUE \\ &else \qquad\qquad periodic'[n] = FALSE \end{aligned} \tag{6.38}$$

$$\begin{aligned} &\text{if} \quad (\text{periodic}'[n] = \text{TRUE}) \,\&\, (\text{periodic}'[n-1] = \text{TRUE}) \\ &\qquad\qquad \text{peridoic}[n] = \text{TRUE} \\ &\text{else} \qquad\quad \text{periodic}[n] = \text{FALSE} \end{aligned} \tag{6.39}$$

Equation (6.37) uses modulo division to compare the current and previous pitch values, where P_{max} is the larger and P_{min} is the smaller of the two values. Modulo division is used to overcome the problem that the autocorrelation function sometimes has its largest peak value at a multiple of the pitch period rather than at the true pitch

period. To demonstrate this algorithm, Figure 6.8 shows how varies with the ratio of the two pitch values that are being compared. In this example a constant pitch value of 50 (160Hz) is compared with pitch values in the range 10 to 150 (50 to 800Hz), i.e. the pitch ratio is:

$$\text{Pitch Ratio} = \frac{(10 \rightarrow 150)Hz}{50Hz} \tag{6.40}$$



Behaviour of the pitch detection variable 'diff'

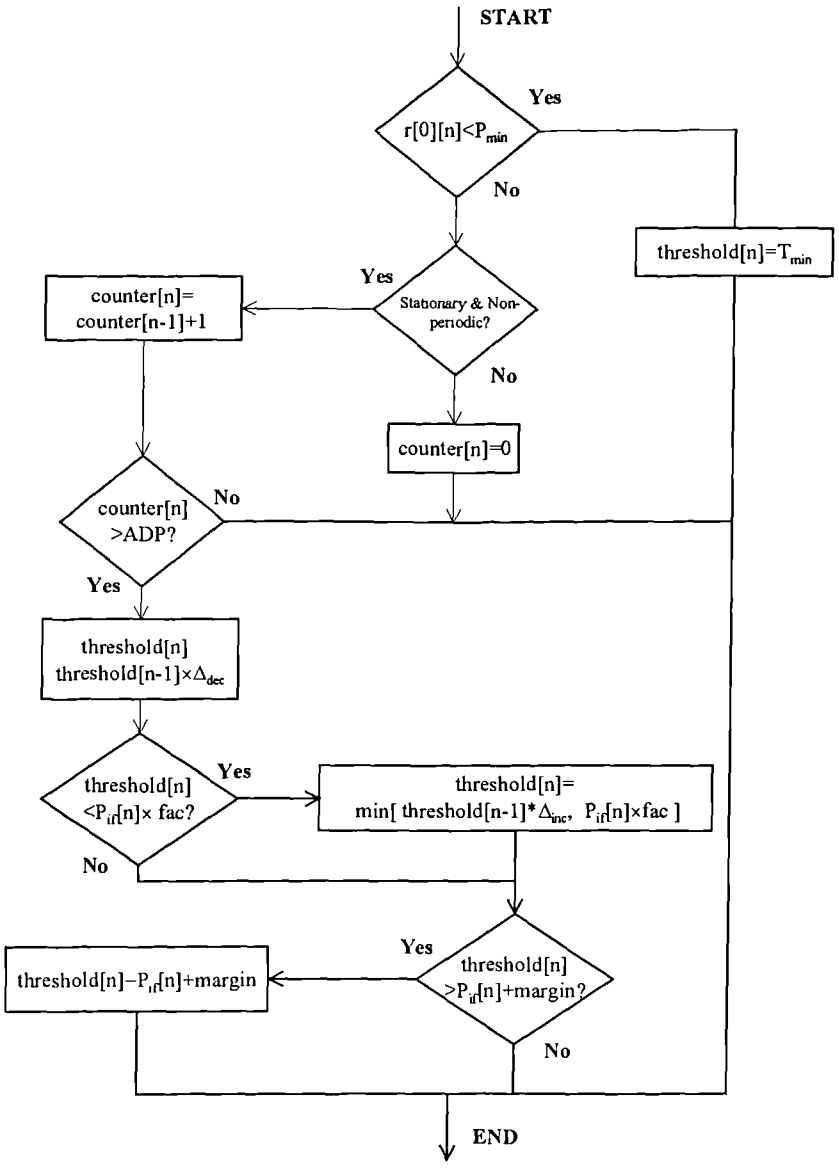
Figure 6.8

When neighbouring pitch values are integer multiples (i.e. pitch ratio=4,3,2,1,0.5,0.25 ....) it is seen that diff=0. This enables the signal to be deemed periodic when consecutive frames have the same pitch, but the largest peak in the autocorrelation function, for one of the neighbouring frames, occurs at a delay which is a multiple of the true pitch value. Δ_p is a constant whose value has been determined as a result of testing with a variety of files of speech and background sounds . The value 3 is used in practice. Equation 6.39 strengthens the periodicity test further by requiring that the period should be found to change very little over three consecutive frames if the final frame is to be declared to be periodic.

6.5.2.3 Threshold Adaptation

The primary VAD decision threshold must be adapted so that an accurate decision can be made as to whether the input is speech or background sound. When stationary

background sounds are present at the VAD input, the adaptation algorithm attempts to adjust the threshold so that it exceeds the resulting $P_{if}[n]$ by a given margin. Additionally, if $P_{if}[n]$ changes due to changing background sounds, the adaptation process attempts to maintain this margin. The adaptation process is described in Figure 6.9 [WATS98]:



Primary VAD Threshold Adaptation Algorithm

Figure 6.9

Initially the energy of the current frame is compared with a fixed threshold and if less than this, the VAD threshold is set to a predetermined minimum value. This

corresponds to situations where the frame contains inaudible speech or very low-level background noise. However, if the frame energy is significant and the signal has been declared stationary and non-periodic for several consecutive frames, it is deemed that background noise is present. Under these conditions, the VAD threshold may be adapted and its model of the background sounds updated. The first stage in adapting the threshold is to reduce it by a factor of Δ_{dec}. If the resulting threshold is less than the tracking level, $P_{lf}[n]\times fac$, it is increased by $\Delta_{inc}=2\Delta_{dec}$. However, if this increase causes the threshold to become larger than the tracking level, it is set to the tracking level. Thus the threshold can increase or decrease to track changes in $P_{lf}[n]$ but can never exceed the target tracking level, $P_{lf}[n]\times fac$. Finally, a check ensures that the threshold does not exceed a value above which low amplitude speech could be incorrectly classified as background noise.

It is found that the rate of adaptation depends on the type of background noise [WATS98]. For example, the threshold adapts more quickly for car and babble noise than for multi-speaker noise. This is because the assumption that the background noise is stationary and non-periodic is satisfied less frequently during multi-speaker noise. Hence, it will take longer for the VAD to operate correctly in the presence of multi-speaker noise than in the presence of car noise. When used in the comfort noise injection system it is very important that the VAD should begin to operate correctly very quickly after start-up otherwise there will be too few sets of background parameters to permit realistic operation. It is likely that at the beginning of a call, there will be an initial period in which only background is present, before the start of near-end talker or echo. For this reason, when used in the comfort noise injection system for the first 0.5 seconds the VAD is forced to adapt its threshold and to indicate the presence of background. In this way, the VAD is quickly able to detect the presence of speech with good reliability and to capture sufficient background parameters to commence good quality operation.

6.6 Comfort Noise Performance

This section describes the tests that were used to examine the performance of the comfort noise system, which was implemented as a high-level language computer

simulation. The tests have two objectives – firstly to determine if both echo and near-end speech could be reliably detected by the VAD, and secondly to assess the quality of the comfort noise synthesised using the VAD linear prediction coefficients. To this end, the operation of the adaptive filter was simulated and prior knowledge of the test waveforms was used to ideally control the centre-clipper. This ensures that the clipper is only enabled when echo is present and that the clipping threshold is always larger than the amplitude of the background sounds.

In general, the ability of the VAD to distinguish speech (or echo) from background sound is diminished as the power of the speech decreases relative to that of the background. In the echo cancellation environment, this can happen in two ways: either the ERL can increase or the background power can increase. In both of these cases the speech/echo becomes less significant compared to the background sound. Ideally, the comfort noise system should function over a wide range of background sound levels and ERL values. In these tests ERL values of 6dB, 12dB and 20dB were used, whilst the background noise levels were set, relative to the near-end talker, at -20dB, -10dB and 0dB. Table 6.1 shows for each background to near-end talker power level, σ_b^2/σ_{s2}^2, the corresponding background to echo power ratios, σ_b^2/σ_e^2.

$10\log_{10}\left(\sigma_b^2/\sigma_{s2}^2\right)$	$10\log_{10}\left(\sigma_b^2/\sigma_e^2\right)$		
	ERL=6dB	ERL=12dB	ERL=20dB
-20dB	-24dB	-18dB	-10dB
-10dB	-14dB	-8dB	0dB
0dB	-4dB	2dB	10dB

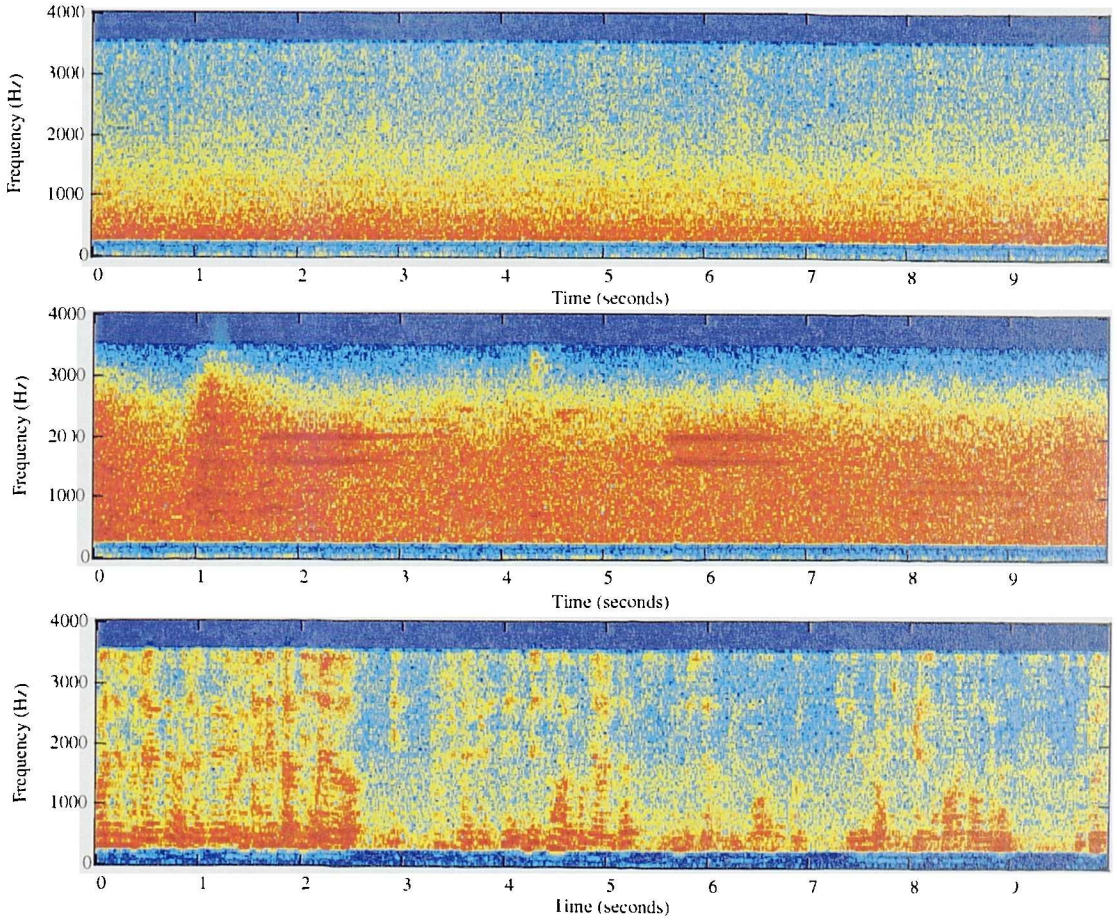
Table 6.1

The testing was carried out in two stages. First, the spectral characteristics of the processed residual echo were examined in order to assess the spectral shape and average power of the comfort noise in relation to the original background sounds. The spectrograms also give an indication of the time varying nature of the comfort

noise, when the background sounds are time varying. Secondly, a simple informal listening test with ten listeners was used to compare the DCNI system to the standard system that was described in chapter five.

6.6.1 Test Signals

The performance of the DCNI system has been evaluated using several different types of background sounds, which were provided by BT Labs. The noise types used in the tests are, car noise, babble noise and multi-speaker noise and, are believed to be representative of the type of sounds encountered within the network The spectrograms for ten seconds of each type are shown in Figure 6.10.



a) Car Noise, b) Babble Noise, c) Multi-speaker Noise

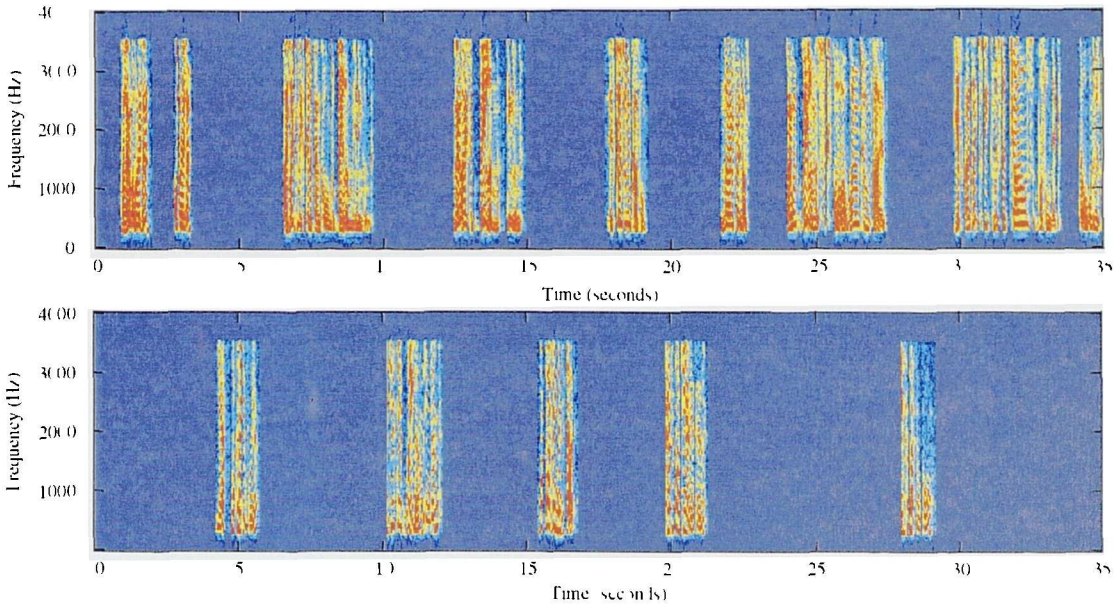
Figure 6.10

These three examples of background sounds have several important features. A common feature to all the background sounds is that they have more energy at low

frequencies than at high frequencies, and that they are bandlimited to between 330-3.5kHz, a typical transmission bandwidth of a PSTN.

The car noise used in the tests was recorded in a car that was travelling at a constant speed and consequently, its spectral characteristics are approximately constant. Many different types of car noise could have been used in the testing. For example, the car could be accelerating or decelerating, being driven with the windows down or the radio might be playing. The spectral characteristics of these types of car noise will obviously not be as constant as the characteristics of the car noise used here. In contrast to the car noise, the spectrum of the multi-speaker noise is highly variable. In multi-speaker noise, the power of the voiced (periodic) components is much larger than the power of the unvoiced (non-periodic) components. Conversely, the power of the non-periodic components in the car noise is much larger than the power of the periodic components. Thus the characteristics of the car and multi-speaker noises are 'opposite' in terms of spectral variability and the relative powers of periodic and non-periodic components. It was expected that because of these factors, the synthetic car noise would sound much more like the original than the synthetic multi-speaker noise. Babble noise, whose spectrum is shown in Figure 6.10b, has some of the characteristics of both the car and multispeaker noises. Its spectrum is more variable than that of car noise because it contains bursts of speech and other periodic sounds. For example, there are two one second bursts of ringing telephone that start at approximately 2.0 and 5.5 seconds.

In addition to the background noise, a recording of a conversation between two people that was supplied by BT Labs was used to simulate the near-end and far-end talkers. This consists of a male speaker at the far-end and a female speaker at the near-end. The spectrograms for 35 seconds of these signals are shown in Figure 6.11 below.



a) Male speaker (at far-end), b) Female speaker (at near-end)

Figure 6.11

It should be noted that the conversation does not contain any double talking Although the behaviour of the clipper during double-talking is important, only the single-talk characteristics of the comfort noise are considered here.

6.6.2 Spectral Performance

Figure 6.12, Figure 6.13 and Figure 6.14 show spectrograms for 35 seconds of the processed residual echo, that are obtained when car, multi-speaker and babble noise are used as the near-end background signal. In each case ERL=6dB and the background noise levels are set at 0dB, -10dB and -20dB.

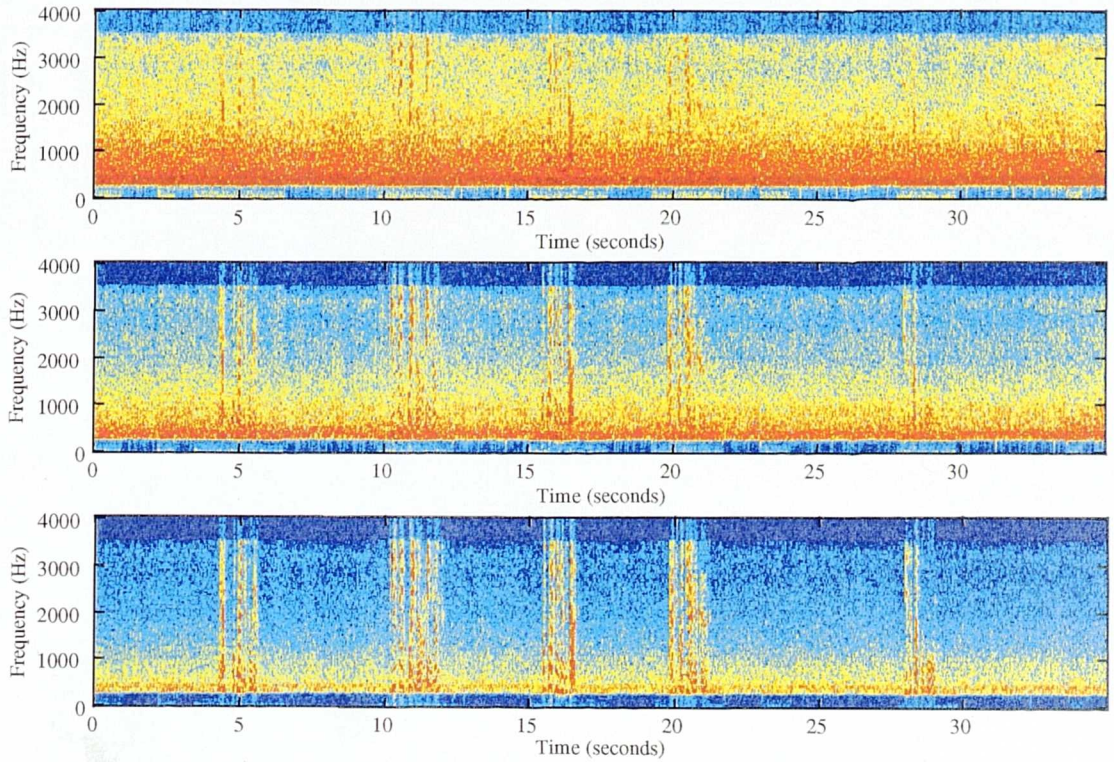


Figure 6.12

Car Noise at a) 0dB, b) -10dB and c) -20dB

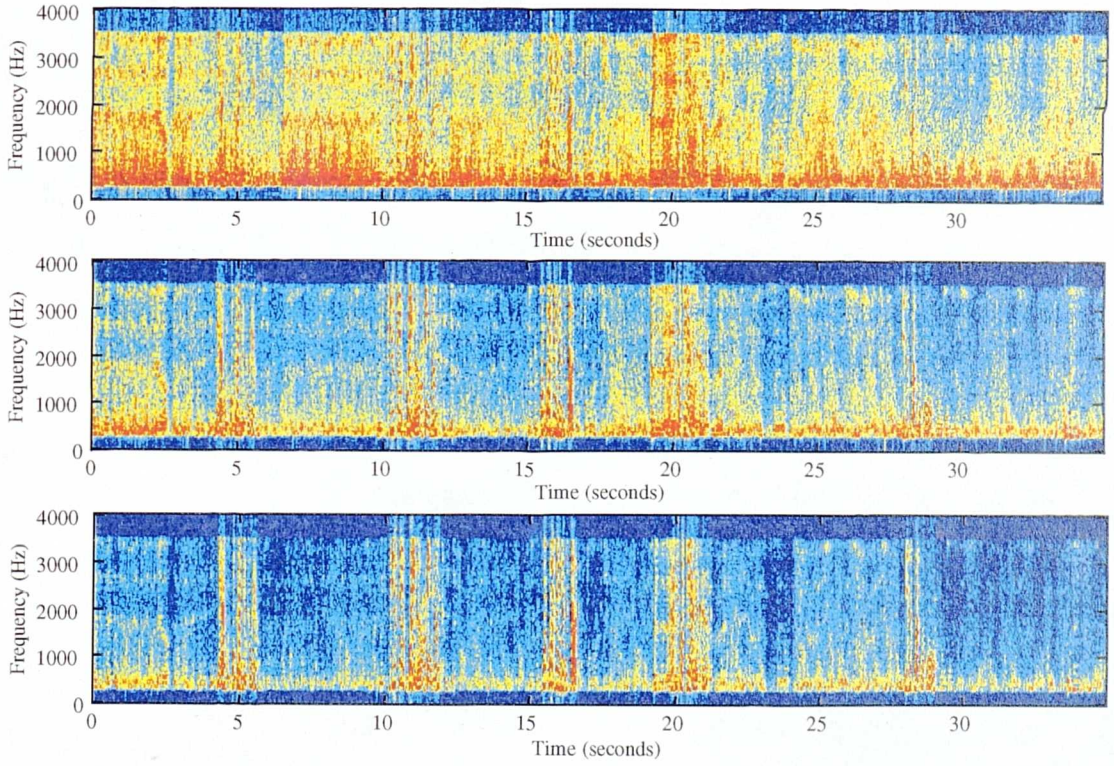


Figure 6.13

Multi-speaker Noise at a) 0dB, b) -10dB and c) -20dB

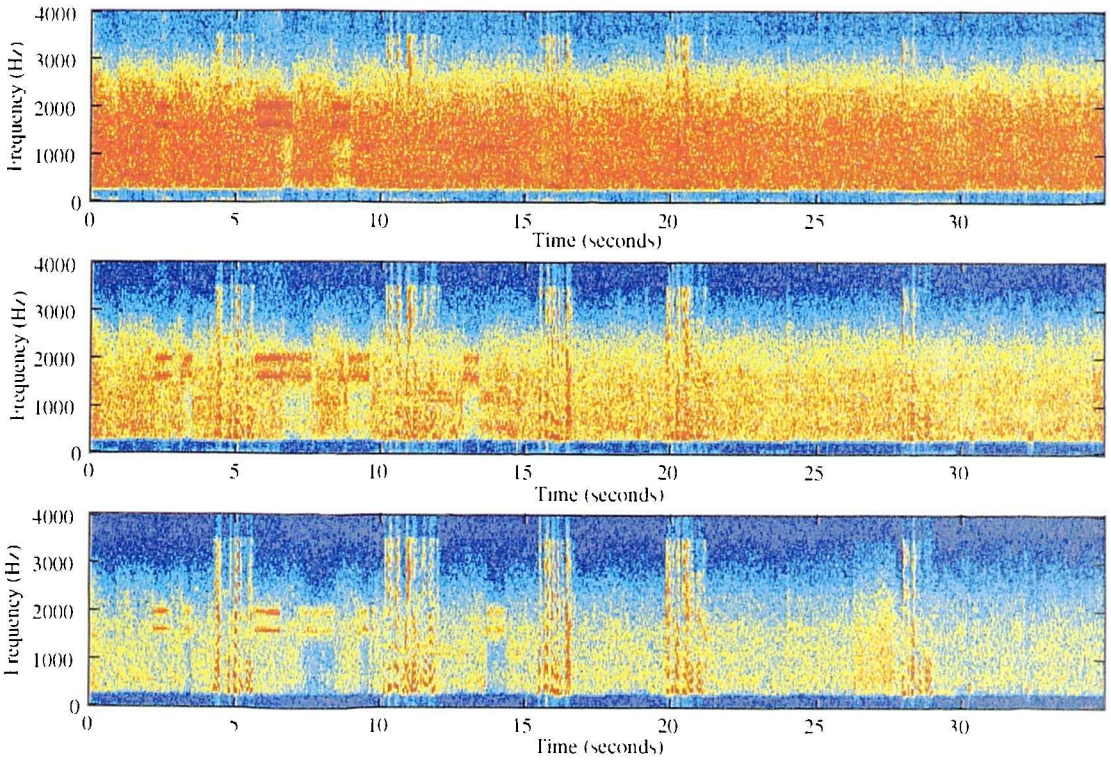


Figure 6.14

Babble Noise at a) 0dB, b) -10dB and c) -20dB

Figure 6.15, Figure 6.16 and Figure 6.17 are similar to the previous spectrograms, except that ERL=20dB.

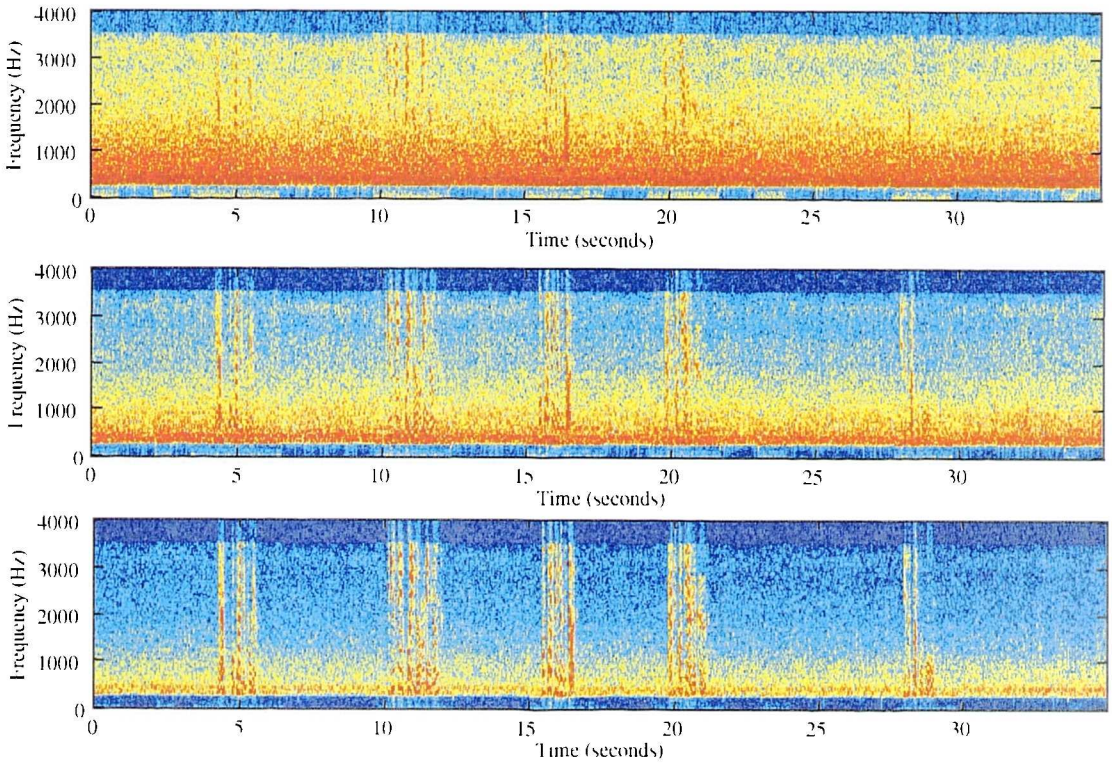


Figure 6.15

Car Noise at a) 0dB, b) -10dB and c) -20dB

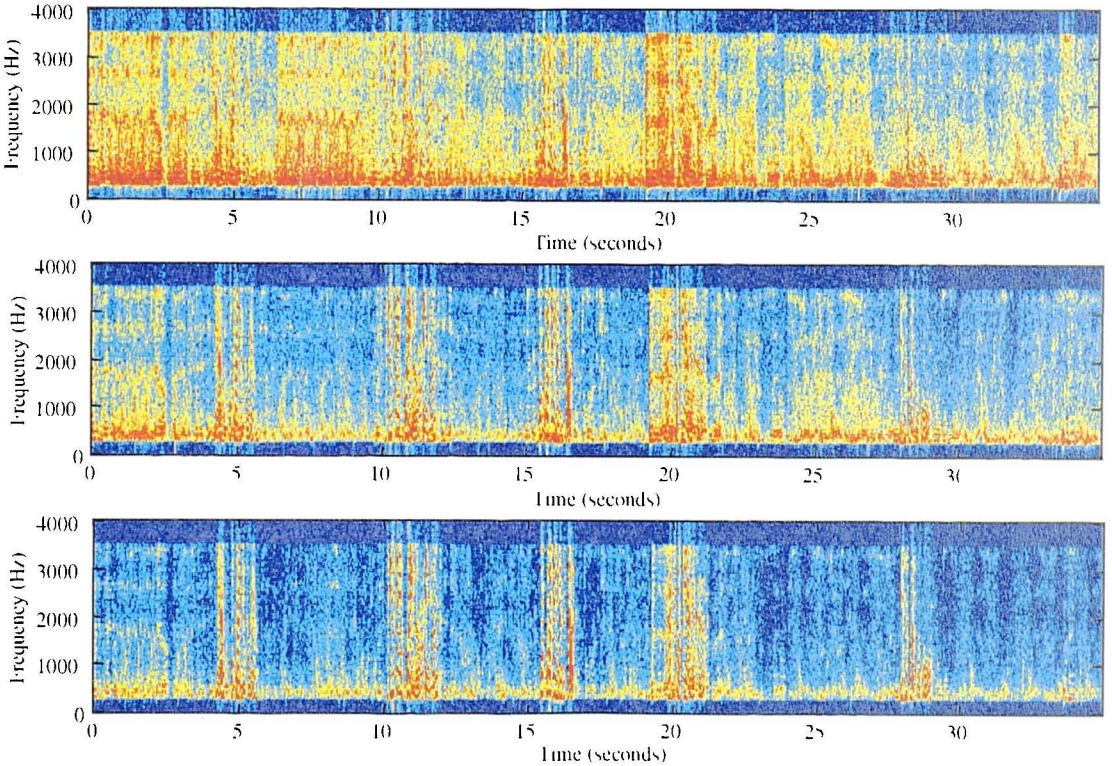


Figure 6.16

Multispeaker Noise at a) 0dB, b) -10dB and c) -20dB

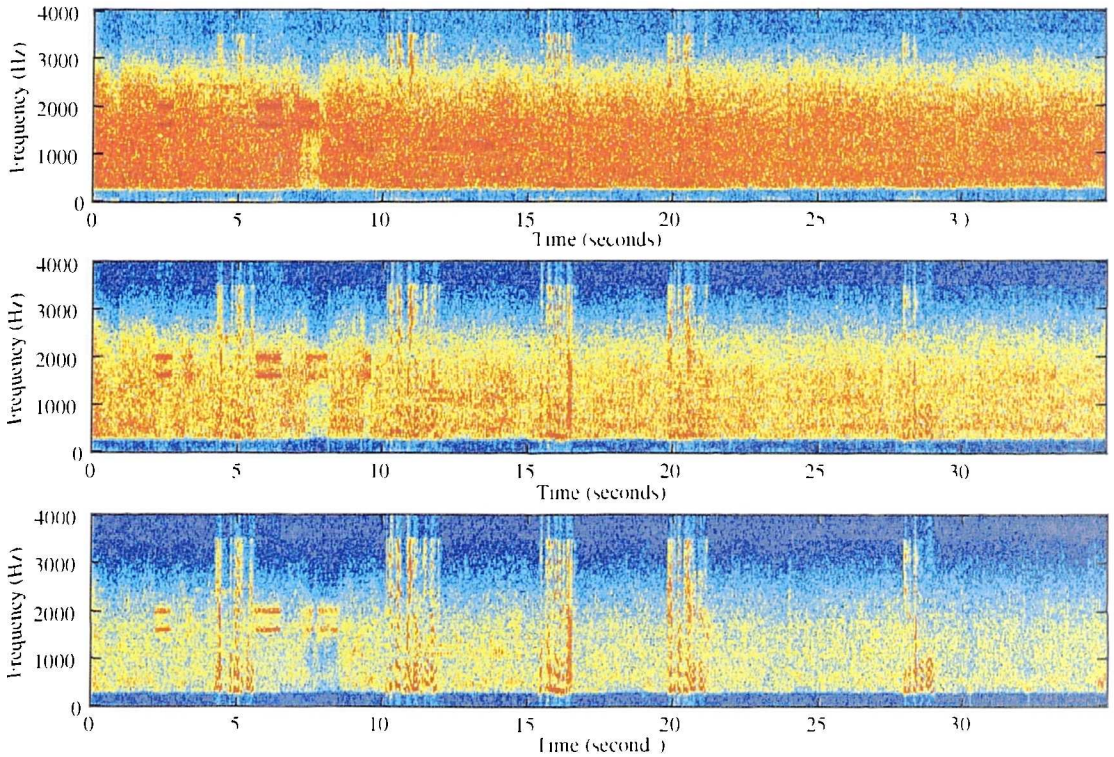


Figure 6.17

Babble Noise at a) 0dB, b) -10dB and c) -20dB

These spectrograms clearly show how the comfort noise system performs in terms of spectral and overall power matching, because any discontinuities are easy to see. The spectrograms suggest that, for these particular test conditions, the operation of the VAD is sufficiently reliable to detect the presence of echo or near-speech. When car noise is present at the near-end, the spectrum of the comfort noise is generally a good match to that of the actual background. However, at frequencies less than 330Hz, the actual background sounds have more energy than the comfort noise. This is because, as discussed in sections 6.3.4 and 6.3.5, the 10^{th} order all-pole model used here is unable to accurately estimate the spectrum in these regions.

For both ERL=6dB and ERL=20dB, the spectral shape of the car and multi-speaker noise appears to be a good match to the original background noise. For the multi-speaker noise, the 'playback' of the prediction coefficients yields comfort noise whose time variation does not appear to be unrealistic. The spectrograms obtained when car and multi-speaker noise are used suggest that it will not be obvious when the actual background sounds are being replaced with the synthetic versions, as there

are no obvious 'joins' when switching from actual to synthetic background and vice-versa.

For the car and babble noise, the spectral matching at start-up appears to be satisfactory, apart from in the stop-band regions. In the comfort noise generated when babble noise is present, a ringing telephone is re-inserted at times when not present in the original background. If sufficient coefficients are learned to replace those that represent the ringing telephone, it is possible than the ringing tones will not be injected. However, this cannot be relied upon to occur.

It is also noticeable that when multi-speaker noise is present, the comfort noise is 'repeated' for a considerable time after start-up. This is because the VAD takes several seconds to adjust its threshold to enable correct operation.

Apart from these general observations, it is difficult to draw any further conclusions by examination of these spectrograms. The following section describes an informal listening test that was used to assess the characteristics of the comfort noise.

6.6.3 Informal Listening Tests

To evaluate the audio performance of the comfort noise, several types of stereo file were constructed with different signals on the two channels. For each test condition the files are:

i)	Left:	Far-end talker	Right:	Unprocessed residual echo + background noise + near-end talker.
Ii)	Left:	Far-end talker	Right:	Residual echo processed using standard centre-clipper.
Iii)	Left:	Far-end talker	Right:	Residual echo processed using centre-clipper + white comfort noise.
Iv)	Left:	Far-end	Right:	Residual echo processed using

talker	centre-clipper + dynamic comfort noise.

The car, multi-speaker and babble noise was added at the levels of –30dB, -20dB and –10dB. Listening to both channels simultaneously gives an impression of what the comfort noise might sound like in real-time implementation. The files were played to ten different listeners, who were asked to decide which they preferred and explain why.

For all test conditions described here, the dynamic comfort noise was preferred over the other methods, and the following conclusions may be drawn. Noise modulation in the usual sense is inaudible, especially for car and multi-speaker noise, as indicated by the spectrograms. Compared to the white comfort noise injection, listeners preferred the spectrally shaped comfort noise because it sounded more realistic.

Looping of the comfort noise may sometimes be heard if the processed residual is listened to on its own, but it is much less obvious when listened to in conjunction with the far-end talker. These looping effects were much more obvious when multi-speaker noise was used, compared to the car and babble noise. For the multi-speaker noise at –10dB, 8 out of the 10 listeners noticed this effect and reported that it was undesirable.

Although the performance of the VAD was found to deteriorate as the background noise power becomes larger, or the ERL increases, the 'quality' of the comfort noise did not appear to deteriorate. This may be because although the comfort noise contains components that are due to the echo, the power of the background sounds is such that they dominate the overall spectrum.

The informal listening tests carried out here are limited in that they can only partially indicate what the comfort noise system would sound like in a real echo canceller. In order to justify the use of this technique, the comfort noise must be tested in a real-time environment. A real-time evaluation could be performed by instructing the two users to perform simple tasks which requires verbal communication. For example, the users could be asked to match up pairs of identical pictures by describing them to each other.

6.7 Conclusions

The operation of echo cancellers has a large effect on the perceived quality of telephone calls with a long round-trip delay. Due to the use of A/μ-Law companding and the presence of near-end background sounds, it is unlikely that an adaptive filtering structure will be able to completely cancel any echo and hence, the use of a centre-clipper, may be required to suppress the residual echo. However, when near-end background sounds are present, the switching action of the centre-clipper causes undesirable 'noise modulation' or 'noise pumping' to be heard at the far-end.

A comfort noise algorithm for echo cancellation has been developed which ensures that little or no residual echo, or noise-modulation is heard at the far end, assuming that the centre-clipper threshold is set correctly and that the system is enabled during periods of echo. The work carried out so far suggests that linear prediction may be used to model the actual background sounds, and that the resulting prediction coefficients may be used in a synthesis filter, excited by a random signal, to generate comfort noise that sounds similar to the original background. As the comfort noise is required to imitate the background sounds, a GSM voice activity detector (VAD) is used to determine when only near-end background sounds are present. During these periods, the linear prediction coefficients of the background are calculated and stored in a circular buffer.

Informal listening tests have suggest that the switching between real and synthetic background sounds is much less disturbing than the noise modulation generated by a

centre-clipper, and is more realistic than the injection of pseudo-random comfort noise.

Currently there are several limitations. A first limitation is that when the background sounds have significant energy at low frequencies, the bandlimiting of the excitation is not sufficient to compensate for the inaccurate modelling in the stop bands. As discussed, simply increasing the model order is not sufficient because the synthesis filter will attempt to model the periodic components of the background noise. The use of such coefficients results in unpleasant sounding comfort noise. An alternative spectral estimation technique that can more accurately model the background sounds in the stop-bands is required.

The simulations of the comfort noise system have been carried out using ideal control of the centre-clipper and its thresholds, which allows the VAD-comfort noise concept to be tested during periods of single-talk. A second limitation of the proposed comfort noise scheme is that this ideal control cannot be achieved in a real canceller and so, further tests must be carried out using practical control algorithms. Furthermore, the evaluation of comfort noise performance must be carried out in a real-time environment, as this will enable a more accurate assessment of how it would perform in a real network.

7. Tone Detection

7.1 Introduction

In the previous chapter, it was demonstrated that when babble noise is present at the near-end, the comfort noise system will 'capture' a ringing telephone sound that occurs intermittently in the test signal. Informal listening tests suggested that the resulting comfort noise sounds unrealistic and intrusive, because it contains repeated bursts of synthetic telephone at times when not present in the actual background. In the same way, it is undesirable to use any tonal signals, such as whistles or the DTMF dial tones that are used for in band signalling, for the generation of comfort noise. Although the VAD might detect and eliminate some of these sounds by incorrectly classifying them as speech, extra measures must be taken to ensure that the linear prediction coefficients derived during such sounds are not used in synthesising the comfort noise.

When tonal signals are present in the background sounds, there may be several peaks in the spectrum, which correspond to the frequencies that make up the tone. The exact number of peaks will not be known in advance of detection because they will vary from one sound to another. The detection of such tones may be accomplished by using any technique that is sensitive to these spectral peaks. One such method uses reflection coefficients, also known as partial correlation coefficients (PARCORs). This has the particular advantage that these coefficients are already available since

they are a by-product of Durbin's algorithm that is used to calculate the linear prediction coefficients in the VAD. This method [DAVI88] was originally proposed for the discrimination of speech and tonal modem training signals. When training signals are present, certain reflection coefficients will become approximately equal to unity, depending upon the number of peaks in the spectrum. The application of this method in the comfort noise system is slightly different from that previously described because the tones are likely to appear in the presence of other background sounds. This alters the behaviour of the reflection coefficients and has necessitated the investigation described here.

7.2 Partial Correlation Coefficients

Referring to the discussion of linear prediction in the previous chapter, the transfer function of the all-pole synthesis filter is given by

$$H(z) = \frac{G}{1 - \sum_{i=1}^{P} a_i z^{-i}} = \frac{G}{A(z)} \qquad (7.1)$$

where

$H(z)$ = all pole transfer function

$A(z)$ = $1 - \sum_{i=1}^{P} a_i z^{-i}$

G = gain parameter

a_i = linear prediction coefficients

P = the prediction order

The transfer function of the inverse filter A(z) may be written as:

$$A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \cdots\cdots - a_P z^{-P} \qquad (7.2)$$

$$= z^{-P}\left(z^{P} - a_1 z^{P-1} - a_2 z^{P-2} - \cdots\cdots - a_P\right)$$

$$= z^{-P}(z - z_1)(z - z_2)\cdots\cdots(z - z_P)(z - z_P) \qquad (7.3)$$

where

z_1 to z_P are the roots of A(z), which may be real or complex, and

a_1 to a_P are the linear prediction coefficients.

The right-hand side of equation (7.3), excluding the z^{-P} term, may be written as:

$$\left(z^P - a_1 z^{P-1} - a_2 z^{P-2} - \cdots\cdots - a_P\right) = (z - z_1)(z - z_2)\cdots\cdots(z - z_P) \qquad (7.4)$$

and therefore

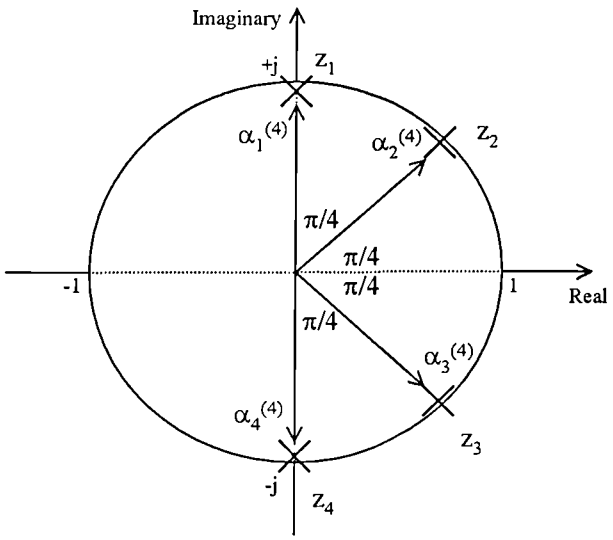
$$-a_P = (-1)^n z_1 z_2 \cdots\cdots z_P \qquad (7.5)$$

It is a feature of the Durbin's method that each new iteration of the algorithm generates the complete set of prediction coefficients of the next higher order predictor. It is normal practice within this algorithm, described by equations 6.15 to 6.19, to calculate a quantity k_i, which is known as a reflection coefficient, with one new value being found at each iteration. Inspection of 6.17 shows that k_i is equal to linear prediction coefficient a_i of the i^{th} iteration. Thus, if a 10^{th} order predictor is found for a particular waveform then, as an example, k_4 would represent a_4 for a 4^{th} order predictor, k_5 would represent a_5 for a 5^{th} order predictor, etc., for the same waveform.

Equation 7.5 demonstrates that the p^{th} prediction coefficient is the product of the roots of the p^{th} order inverse filter transfer function. Bearing in mind that for the p^{th} order predictor a_p is equal to reflection coefficient k_p equation (7.5) leads to the conclusion that for even orders, the reflection coefficients will always be negative when all of the poles occur in complex conjugates pairs, whilst for odd orders it may be either positive or negative. The reflection coefficients for even orders can only be positive if there is an odd number of real negative valued poles.

Now, suppose that we are attempting to model a spectrum that has two spectral peaks, for example two sinusoids at 1kHz and 2kHz, using a 4^{th} order system. After

solving the resulting normal equations using Durbin's method, the positions of the poles of $H(z)$ may be plotted and might be as in Figure 7.1:



The pole positions for a 4th order analysis

Figure 7.1

In this example the poles are given by:

$$z_1 = \alpha_1^{(4)} e^{j\pi/2} \tag{7.6}$$

$$z_2 = \alpha_2^{(4)} e^{j\pi/4} \tag{7.7}$$

$$z_3 = \alpha_3^{(4)} e^{-j\pi/4} = \alpha_2^{(4)} e^{j\pi/4} = z_2^* \tag{7.8}$$

$$z_4 = \alpha_4^{(4)} e^{-j\pi/2} = \alpha_1^{(4)} e^{j\pi/4} = z_1^* \tag{7.9}$$

where * denotes complex conjugation and the superscripts indicate the prediction order. Therefore from equation (7.5):

$$-a_4 = (-1)^4 (z_1 z_1^*)(z_2 z_2^*) \tag{7.10}$$

$$= (-1)^4 \alpha_1^{(4)} \alpha_2^{(4)} \tag{7.11}$$

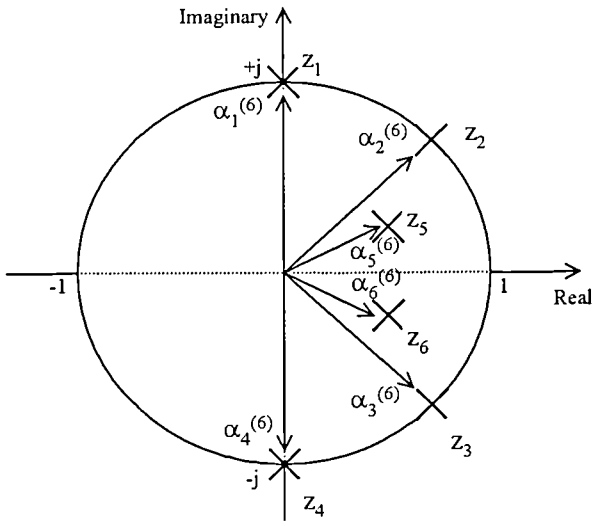
In this case, $\alpha_1^{(4)} \approx 1$ and $\alpha_2^{(4)} \approx 1$ because the poles will be very close to the unit circle, and therefore:

$$-a_4 \approx 1 \qquad (7.12)$$

For a P^{th} order all-pole model, $k_P = a_P$, therefore for the 4^{th} order case illustrated here:

$$k_4 \approx -1, \text{ or } |k_4| \approx 1 \qquad (7.13)$$

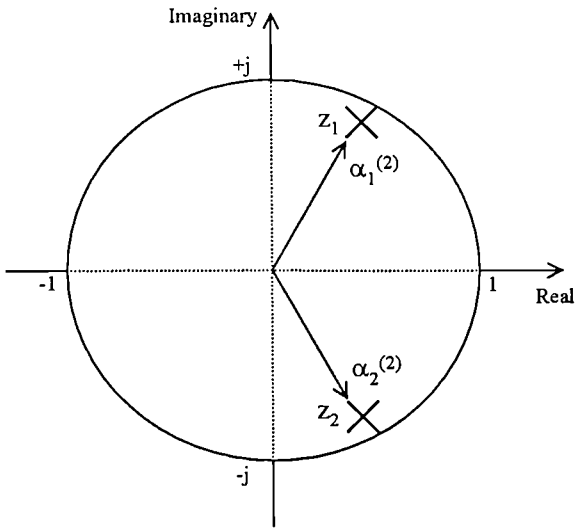
Now suppose that a 6^{th} order LP analysis is used to model the same system. Again, after solving the normal equations using Durbin's algorithm the poles are likely to be placed as shown in Figure 7.2.



The pole positions for a 6^{th} order analysis

Figure 7.2

Now, $|k_6| < 1$ even if the poles that are modelling the peaks are closer to the unit circle (i.e. $\alpha_1^{(6)} > \alpha_1^{(4)}$, $\alpha_2^{(6)} > \alpha_2^{(4)}$) because z_5 and z_6 are closer to the origin than the other poles. When modelling this particular spectrum, a similar argument applies for higher order partial correlation coefficients. Turning to lower order prediction, Figure 7.3 shows the pole positions when the spectrum is approximated using a 2^{nd} order model.



The pole positions for a 2nd order analysis

Figure 7.3

As discussed previously, the spectral estimation process minimises the integrated ratio of the actual spectrum to the estimated spectrum. This is best achieved here by placing the two poles near the unit circle at a frequency that lies between the two peaks in the original spectrum. In this example, the two peaks are separated by 1kHz and it is probable that $\alpha_1^{(2)} < \left(\alpha_1^{(4)}, \alpha_2^{(4)}\right)$ and hence $|k_2| < |k_4|$. Note that the exact values depend upon the frequency difference between the peaks.

In [DAVI88], the presence of modem training signals was determined by searching for values of $|k_n|$ that are approximately equal to unity. In addition, it was found that the number of peaks can be determined from knowledge of the largest reflection coefficient, i.e. if $|k_2|$, $|k_4|$ or $|k_6|$ is maximum then the spectrum has 1, 2 or 3 peaks respectively. When the tonal signals are embedded in background sounds, such as car or multi-speaker noise, the behaviour of the reflection coefficients is slightly different, but in general, they still change according to the principles described above. However, they can no longer be relied upon to indicate the number of peaks in the spectrum.

The following sections examine the general behaviour of the reflection coefficients firstly when the tones are embedded in white bandlimited noise, and secondly when present during other background sounds.

7.3 Tones in Bandlimited White Noise

Figure 7.4 plots the reflection coefficients that are obtained from two seconds of white noise that has been bandlimited to frequencies between 330 and 3.5kHz when 10^{th} order prediction is being used. Two continuous sinusoids at frequencies of 1kHz and 2kHz are present between 0 5 and 1.5 seconds. The SNR in this case is 20dB and corresponds to the case where, in the power spectrum, the peaks produced by the sinusoids are significantly larger than the randomly varying peaks of the white noise

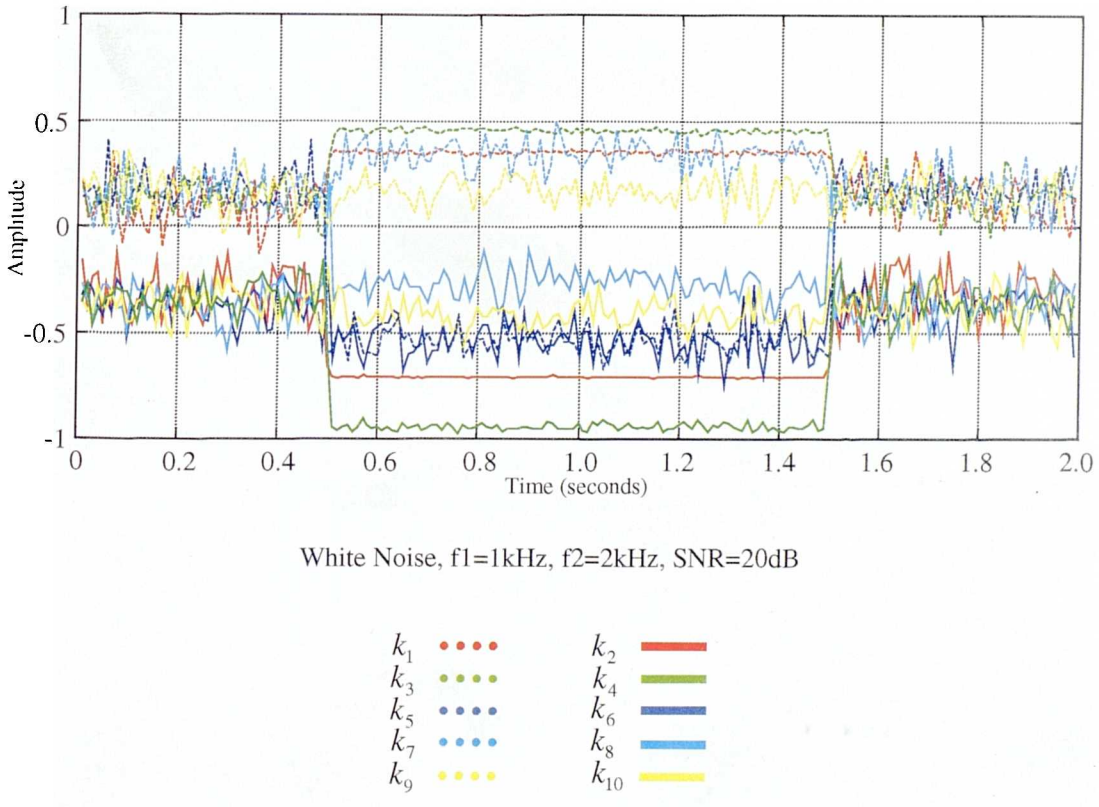
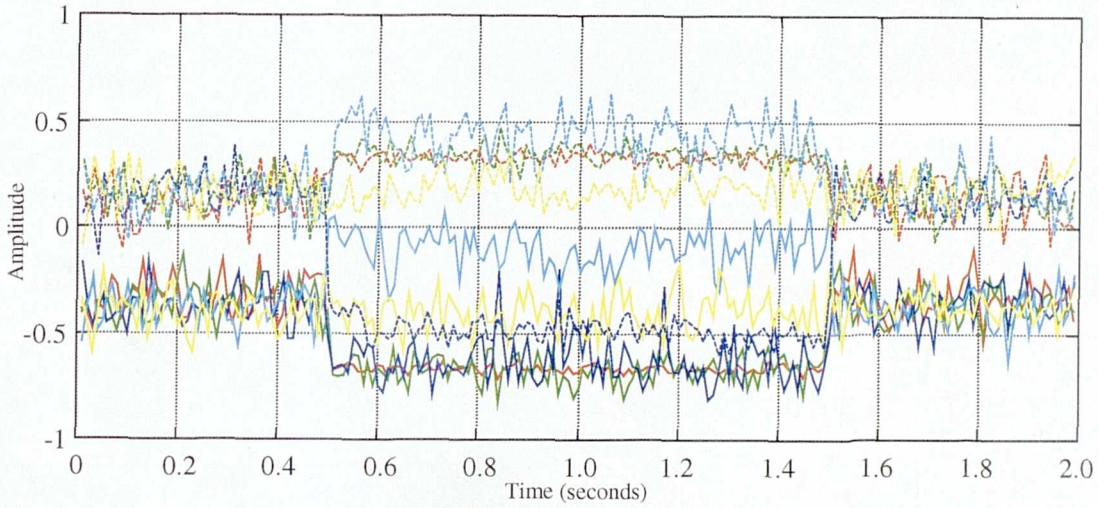


Figure 7.4

Figure 7.4 shows that for this SNR, the magnitudes of k_1 to k_7 generally increase when the sinusoids are present, with $k_4 \approx -0.9$ being the largest. This is in contrast to when the sinusoids are present alone, where k_4 is likely to be less than - 0.9, but greater than -1.0. When higher model orders are used, it is probable that the poles,

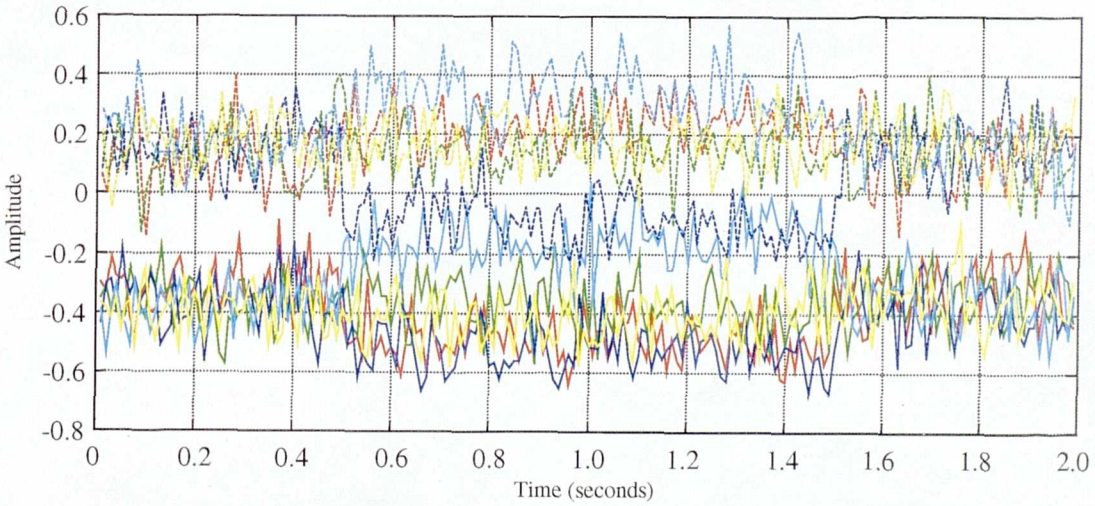
which represent the peaks in the spectrum, will be closer to the unit circle than those from the 4th order case. However, the other poles will be closer to the origin and therefore the magnitude of the higher order reflection coefficients will be smaller than k_4.

Figure 7.5 and Figure 7.6 show what happens when the SNR is reduced to 10dB and 0dB respectively.



White Noise, f1=1kHz, f2=2kHz, SNR=10dB : Refer to Figure 7.4 for colour key

Figure 7.5



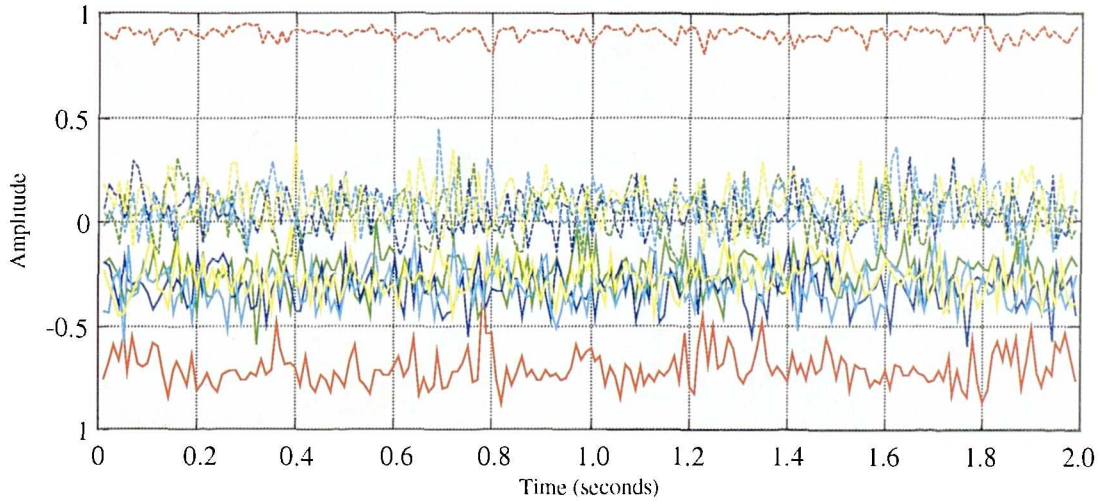
White Noise, f1=1kHz, f2=2kHz, SNR=0dB : Refer to Figure 7.4 for colour key

Figure 7.6

Decreasing the SNR reduces the influence of the peaks in the spectral estimation process, i.e. there is less of a difference between the peaks due to the tones compared with those that represent the 'white' background noise. Thus, the magnitudes of the reflection coefficients change by less when the tones are introduced, with k_1 becoming smaller in magnitude until it is no longer the largest. These diagrams show that unless the SNR is large, i.e. greater than approximately 20dB, it will be impossible to determine the maximum number of peaks in the spectrum by examination of the reflection coefficients. However, detection of the presence of tones can still be achieved by examining the magnitudes of the reflection coefficients. The thresholding process that is needed to perform the detection will be examined after discussing the general behaviour of the reflection coefficients when the tones are embedded in 'real' background sounds.

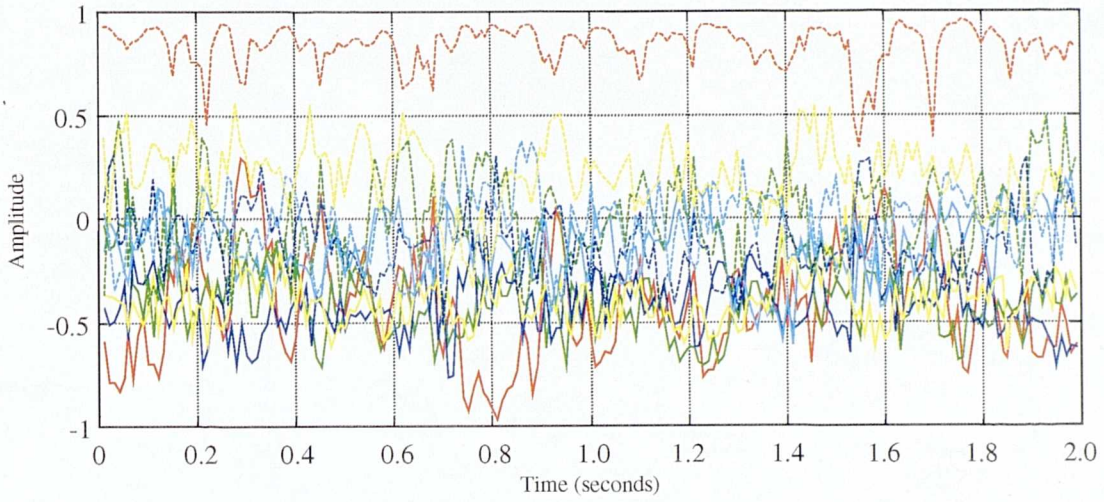
7.4 Tones in Environmental Noise

In practice, the near-end background sounds are unlikely to have a white spectrum. Figure 7.7, Figure 7.8 and Figure 7.9 show the reflection coefficients that are obtained by analysing two second segments of car, multi-speaker and babble noise respectively, that have been bandlimited to between 330Hz and 3.5kHz. For the particular segment of babble noise that is analysed, there is a ringing telephone present between 0.4 and 1.6 seconds.



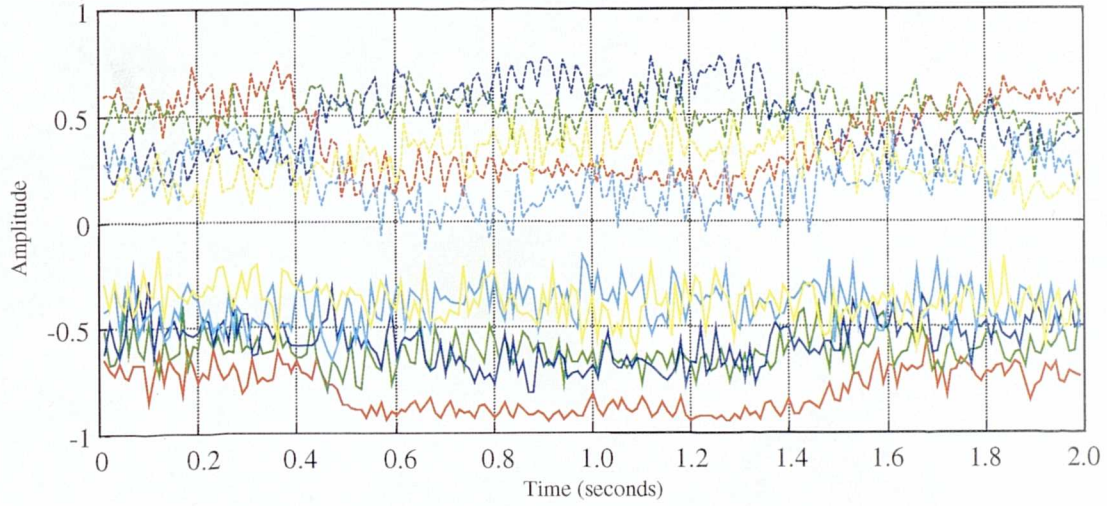
Car Noise : Refer to Figure 7.4 for colour key

Figure 7.7



Multi-speaker Noise : Refer to Figure 7.4 for colour key

Figure 7.8



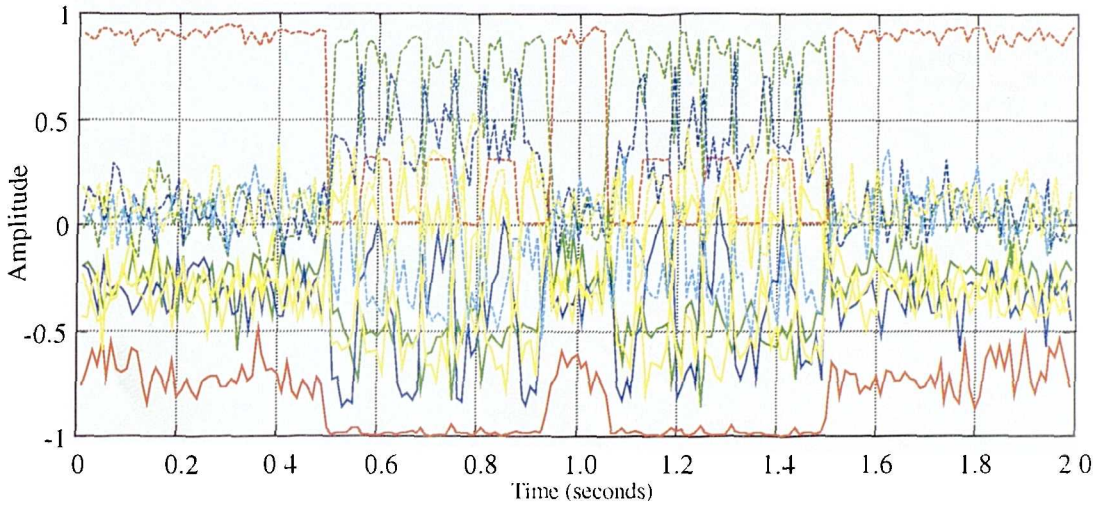
Babble Noise (containing a ringing tone) : Refer to Figure 7.4 for colour key

Figure 7.9

The partial correlation coefficients derived from car noise are similar to those obtained using white noise, except that the magnitudes of k_1 and k_2 are now much larger. The magnitude of k_1 is also large for the multi-speaker noise, with all the coefficients exhibiting a greater degree of variability compared with both the car and bandlimited white noise cases. This is unsurprising, since the spectral characteristics of speech vary more than those of both car and white noise. When the ringing tones are absent, the reflection coefficients, shown in Figure 7.9, that are derived from babble noise do not exhibit the same degree of variability as those for multi-speaker

noise. In addition, the reflection coefficients are generally found to be larger than for the other noise types.

Figure 7.10 shows the reflection coefficients that are obtained when a simulated ringing tone, consisting of two frequencies (1.6kHz and 2.0kHz) that alternate at 60Hz, is added to the car noise at an SNR of 20dB.



Simulated ringing tone in car noise with an SNR=20dB : Refer to Figure 7.4 for colour key

Figure 7.10

Although as expected, the magnitudes of the partial correlation coefficients generally increase during the presence of the ringing tone, comparing the magnitude of all the coefficients to a simple threshold is no longer sufficient to achieve detection This is because both k_1 and k_2 are large even when the ringing tones are absent. A similar behaviour is observed with both the multi-speaker and babble noise. However, the information provided by k_1 may be used intelligently to decide whether ringing tones are present.

7.4.1 The First Order Reflection Coefficient

The value k_1 is equal to a_1 when a first order prediction model is used, i.e. when the system only has one pole. This pole must be real, and will take a value between +1.0 and –1.0 because the autocorrelation method always results in the poles lying inside

the unit circle. The following table illustrates the general behaviour of k_1 when a single sinusoid, of frequency F, is added to white bandpass noise.

Frequency F (kHz)	k_1
1.0	0.7
1.8	0.15
2.0	0.0
3.0	-0.7

Table 7.1

It is seen that k_1 is positive when the frequency is less than 2 kHz, is equal to zero when the frequency equals 2 kHz and is negative when the frequency is greater than 2 kHz. Generally, k_1 acts as measure of the 'centre of gravity' of the spectrum – it is positive if low-frequency components dominate the spectrum and negative if high-frequency components dominate. For a white spectrum the mean of k_1 will be close to zero, but for the telephone bandlimited 'white' spectrum used here, k_1 (=0.15 on average) is positive because the spectrum is biased towards lower frequencies. This may be explained by considering how the spectral estimation process models the actual spectrum when using a first order model. If there is more low frequency energy than high frequency energy, the integrated ratio of the actual spectrum to the estimated spectrum is best minimised by placing the real valued pole somewhere in the RHS of the unit circle, i.e. $z_1 = +x$. Hence using equation (7.5) leads to $k_1 = +x$. Similarly, if there is more high frequency energy than low frequency energy, the pole will be given by $z_1 = -x$ and hence $k_1 = -x$. Inspection of Durbin's algorithm shows that, in fact, the first order partial correlation coefficient is given by:

$$k_1 = \frac{r[1]}{r[0]} \qquad (7.14)$$

were $r[i]$ are the autocorrelation coefficients for the current frame.

In Figure 7.10 the high positive value of k_1 shows that in car noise, most of the energy is located at low frequencies. However, at the onset of the ringing tone k_1 alternates between approximately 0 and 0.3, which corresponds to times when the frequency of the ringing tone is 2kHz and 1.6kHz respectively. The presence of these tones results in a shift in the 'centre of gravity' of the spectrum and this is reflected in the change of k_1. Now, if k_1 is large and positive one might expect $|k_2|$ to also be large because the spectral estimation process will attempt to model the low frequency energy using a single spectral peak. However, this will only be true when the background noise has significantly more energy at low frequencies, as in car noise. A similar situation also exists with the multi-speaker and babble noise, although k_1 and k_2 are now not significantly larger than the other reflection coefficients.

7.5 The Detection Test

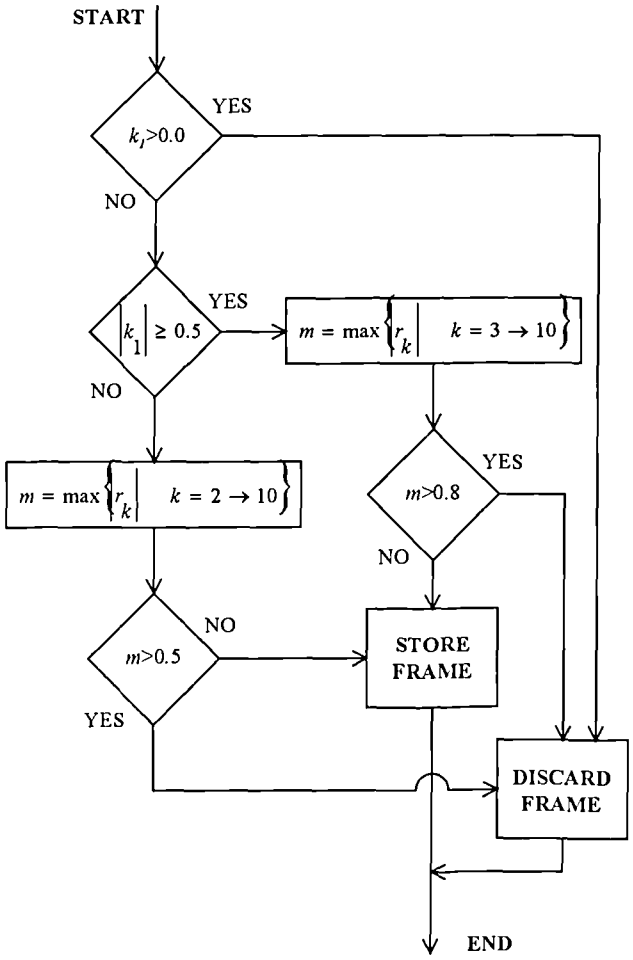
The detection of tones present in white noise, when the SNR is greater than 0dB, is straightforward. For the examples given previously, detection may be achieved by comparing each reflection coefficient to a threshold of 0.5. However, this method on its own is unsatisfactory since $|k_1|$ and $|k_2|$ are likely to be large when background sounds are present.

It is assumed here that the environmental sounds will always have more energy at low frequencies, although this may not be true for all type of background sounds. By examining the reflection coefficients generated for car, babble and multi-speaker noise [JONE96], it was found that $|k_1|$ was generally greater than 0.5 for these background sounds. When the simulated ringing tones were added, with signal-to-noise ratios greater than 0dB, it was observed that $|k_1|$ decreased to less than 0.5. Hence, $|k_1|$ may be used to indicate the presence of background sounds.

If $|k_1|>0.5$, this suggests that only background sounds are present, and therefore that the current LP coefficients should be added. However, by inspection of the partial correlation coefficients for different signal-to-noise ratios and background sounds, it was found that there were short durations when $|k_1|>0.5$ during the presence of ringing tones. During these times some of the other reflection coefficients could be large, i.e. greater than 0.75.

When $|k_1|<0.5$, it is likely that either ringing tones are present, or that the background sounds are white. These two conditions can be differentiated by thresholding $|k_2|$ to $|k_{10}|$ using a value of 0.5.

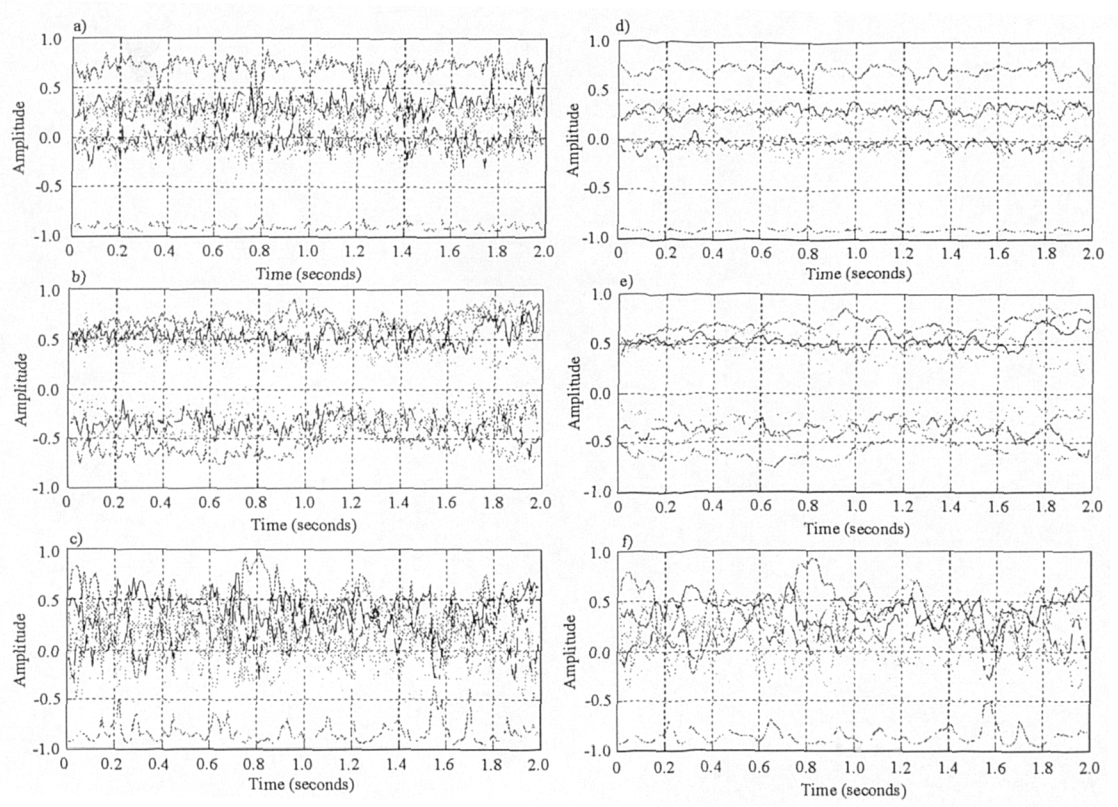
The following flowchart shows the detection process.



Tone Detection Procedure

Figure 7.11

The accuracy of the above scheme is likely to be affected by the rapid variation of the reflection coefficients. This will compromise the accurate detection of ringing tones whose power is similar to that of the background sounds. However, in the DCNI system, the linear prediction and reflection coefficients are obtained from the VAD, where the are calculated from the average autocorrelation coefficients of the last four frames. Therefore, the linear prediction coefficients describe an average or smoothed spectrum and the resulting reflection coefficients do not exhibit the same degree of variability. Figure 7.12 shows how the reflection coefficients calculated by averaging the autocorrelation coefficients over 4 frames (1 frame = 160 samples) compare with those calculated from single frames. Note that the sign of the coefficients is inverted in this figure.



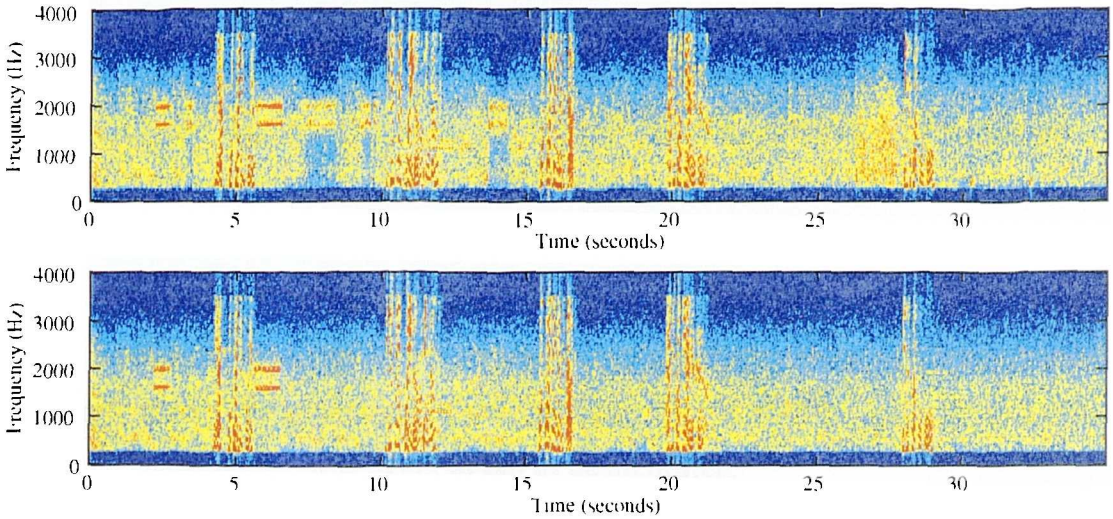
a) to c) unsmoothed reflection coefficients, d) to f) smoothed reflection coefficients

Figure 7.12

This shows that the variation of the reflection coefficients is indeed reduced. It is believed that the use of these 'smoothed' coefficients will give more accurate

detection of the tonal sounds, although no specific testing has yet been carried out to verify this.

Figure 7.13 shows two spectrograms for comfort noise generated when babble noise containing the ringing tone is present.



a) without tone detection, b) with tone detection

Figure 7.13

Figure 7.13a shows that between 6.5 and 10 seconds, the comfort noise contains unwanted ringing. Figure 7.13b shows that the unwanted ringing is removed by the use of the simple procedure described here.

7.6 Conclusions

The behaviour of the reflection coefficients that are obtained when typical background sounds are present has been studied. Although it is not possible to detect the number of peaks in the unwanted tonal signal, it is possible to detect their presence, using the assumption that these background sounds have more energy at low frequencies than at high frequencies.

The thresholds that are used in the detection process were determined empirically, by inspecting the reflection coefficients that were obtained for three different types of background sounds when a simulated ringing tone was added, at varying signal-to-noise ratios. The resulting thresholds appear to give satisfactory detection of the simulated tones and a ringing telephone that is embedded in babble noise. The detection system was tested by embedding simulated ringing tones in real background sounds, and listening to the comfort noise that was generated. In all cases, the comfort noise did not contain any audible synthetic ringing. As the loudness of the ringing relative to the background was diminished the number of occasions when the ringing was not detected increased, however this was only found to happen when the loudness of the ringing was so low that its effect in the comfort noise was inaudible. It should be noted that the performance of the detection system has not been numerically classified in any way, and this should definitely be a subject for further work. The generation of such statistics would enable the conditions under which the algorithm is most likely to fail to be identified.

It is not the fact that the comfort noise contains tones that is the problem, but rather, that they are present repeatedly until new background characteristics are captured. As a further enhancement, if sufficiently accurate detection of the tonal sounds can be achieved, these sounds could be stored in a short term buffer and used only once to generate comfort noise, before being discarded.

8. Conclusions and Further Work

8.1 Conclusions

This thesis has been concerned with some of the problems that are caused by the generation of echoes in Public Switched Telephone Networks. Specifically, only echo generated at the hybrids, which interface the four-wire transmission system to the two-wire local-loop circuits, has been considered. In the UK, such echoes are only a problem on international telephone calls that are routed via satellite, submarine cable or a long landline. In order to maintain the usability and quality of these connections some form of echo control is necessary. Today, the echo cancellation technique is used exclusively on all connections where the echo is known to be a problem.

Echo cancellers use adaptive filtering to generate a replica of the echo, which, when subtracted from the actual echo, will lead to complete cancellation if the two are identical. However, cancellers operate in an environment in which the echo is corrupted by near-end background sounds, and by non-uniform quantisation noise introduced by the companding which occurs in the near-end of the network. Network echo cancellers that employ linear adaptive filtering techniques are unable to remove this quantisation noise even if the optimum filter response is used. If the quantisation noise is returned to the far-end subscriber, it is likely to be perceived as distorted echo, and hence will compromise the quality of the connection.

Many network echo cancellers use a Non-Linear Processor, often a centre-clipper, to remove the unwanted residual echo. It would be advantageous if the use of a centre-clipper could be avoided, because accurate control of the clipper is required to prevent mutilation of the near-end speech. An alternative cancellation method, that avoids the need for a centre-clipper, has been investigated in this thesis. In principle, the residual quantisation noise could be removed by introducing the same non-linearity into the predicted echo as is introduced into the actual echo by the companding process. Now, provided the adaptive filter weights, and hence predicted echo, are sufficiently accurate, the residual echo will be removed completely when subtraction occurs. However, when the standard NLMS algorithm is used to adapt the filter weights, it is likely that the misadjustment error will be such that large unwanted spikes will be frequently introduced into the canceller residue. The modified NLMS algorithm presented in this thesis attempts to reduce the steady-state error by using an estimate of the echo, before it is corrupted by the quantisation noise, in the adaptation equation. In the absence of near-end background sounds, it has been found that the modified technique can be very effective and can reduce the true error by over 40dB for speech waveforms. However, even with this level of improvement, some quantisation level mismatches between the predicted and actual echo still occur, and hence companding of the filter output will still result in some unwanted spikes in the resulting waveform. Since these spikes are generally isolated it would not be difficult to identify and remove them. Unfortunately, the performance of the modified algorithm is severely degraded by the presence of any near-end background sounds or noise. A method of countering these degradations has not been found. Under these conditions, the performance will be no better than that of the standard NLMS algorithm.

The remainder of the thesis has been concerned with the suppression of residual echo and in particular, with the question of how the near-end background sounds are affected by the process that is used to suppress the residual echo. Several forms of centre-clipper have been examined, including single-threshold, twin-threshold, single-band and multi-band devices. A recently proposed 'echo shaping' technique has also been investigated.

All of the methods considered here were found to suffer from a problem known as noise modulation. This is the variation of loudness of the background sounds, caused by the residual echo control process, as heard by the far-end user. Of the clippers tested, the noise modulation caused by the single threshold clipper is the most severe because the clipper output is set to zero during periods when residual echo is to be suppressed. The twin threshold clipper attempts to reduce the effects of noise modulation by transmitting small amplitude samples without modification. A lower threshold setting equal to the standard deviation of the background sounds appears to offer a good compromise between the degree of echo suppression achieved and the amount of noise modulation introduced. Ideally, the lower threshold should vary to track the standard deviation of the background sounds. However, if the background sounds are loud prior to the onset of echo and quiet when the echo starts, it is likely that uncancelled echo will be transmitted because the lower threshold will now be too large. If it is decided that complete echo suppression is to be ensured then a single threshold clipper must be used. These conclusions also apply to single and twin threshold clippers when they are used in a multi-band configuration. The main advantage of the multi-band configuration is that the distortion introduced during double-talk, although not eliminated, can be greatly reduced in comparison to the equivalent single-band system.

In the echo shaping technique, the residual echo is filtered using a low order FIR filter. Ideally, the frequency response of this filter should be adjusted so that the resulting signal has the same spectral shape as the actual background sounds. It was found that this technique offers good suppression of the uncancelled echo, provided that the echo canceller is providing more than 10dB of echo attenuation. However, the use of the proposed method for determining the required filter transmission characteristic does not, in fact, result in the processed signal having the same spectral shape as the background sounds. The analysis suggests that the proposed characteristics are likely to introduce too much attenuation, and this is confirmed by simulations of the echo shaping system. The filtered signal is found to have 'holes' in the spectrum at frequencies where the residual echo has more power than the background sounds. Therefore, noise modulation is also likely to be a problem with this technique. A matter of concern in the use of network echo cancellers is that there

will be an increase in the future in the use of signal processing such as ADPCM in the end-path. Such usage will lead to the appearance of a higher level of uncancellable echo products and hence a reduction of the ERLE. The value of the echo shaping method would be expected to diminish under these circumstances.

Many echo cancellers attempt to mask the operation of their residual echo control devices by injecting so called comfort noise into the output of their NLP's when they are operating. A dynamic comfort noise injection (DCNI) system has been developed, to overcome the deficiencies of conventional comfort noise injection schemes, in which the comfort noise is spectrally shaped to match that of the background. Additionally, the comfort noise is given a temporal variation that, is similar to that of the original background. A GSM voice activity detector has been adapted for use in this system and this is used to detect periods when only background sound from the near-end of the network is present. During these periods, the background sounds are characterised using an all-pole model and the resulting linear prediction coefficients are stored for use in future comfort noise generation. The linear prediction coefficients that represent the spectral envelope of the background sounds are calculated as part of the VAD operation and are therefore available at no extra processing cost. The comfort noise system has been implemented as a high-level language computer simulation, and its performance has been tested for a range of different conditions. The tests suggested that listeners prefer DCNI rather than injection of constant power, white, comfort noise, or the use of a single-band single-threshold centre-clipper without any comfort noise injection.

It was found that the use of certain types of background sounds, in the DCNI system, that have strong intermittent periodic components, e.g a ringing telephone, may result in unrealistic sounding comfort noise. This is because the resulting comfort noise is likely to contain a synthetic ringing telephone at times when not present in the true background sounds. It has been found that by inspecting the reflection coefficients, calculated as a by-product of the linear prediction analysis in the VAD, the presence of such background sounds may be detected, and then excluded from comfort noise generation. The system was found to work reliably for both the babble

noise containing a real telephone, and for other types of background sounds that contain an artificial ringing telephone sound.

It may be concluded that the introduction of DCNI into the operation of conventional echo cancellers would lead to a significant improvement of the perceived quality of performance when calls are made from environments with high levels of background sound. If there is an increase of the level of uncancellable echo, as might occur if ADPCM or some other coding scheme is used in the near-end circuit, and it is therefore necessary to increase the clipping threshold then this would be expected to lead to no increase of complexity of the DCNI system and no deterioration of its quality of performance. However, as noted previously, the increase of clipping threshold would lead to an increase in the mutilation of the near-end speech.

8.2 Further Work

In view of the results presented in chapter 4, the introduction of non-linearities into the predicted echo is likely to be unsuccessful at completely eliminating the residual quantisation noise, even if the steady-state error is very small. Thus, all further work should assume that the near-end background sounds and noise is likely to affect the performance of the echo canceller. Although the algorithm was found unsuccessful when background sounds are present, there may be some connections where the background is quiet enough for the method to offer some value. It would then be more attractive if its implementation could be simplified. The implementation investigated here was complicated by the assumption of a Gaussian distribution of true error. It would be interesting to know if this added complexity is necessary for obtaining a reduction of misadjustment. The algorithm could be tested using simpler probability distributions, for example a triangular distribution might be used. Even simpler, would be to use as the error, the same measure as is used in the calculation of the standard distribution of true error i.e. the difference between the predicted echo and the nearer threshold of the level that contains the quantised echo value.

Currently, the comfort noise system has been implemented and tested using a high-level language computer simulation. In order to evaluate its characteristics more thoroughly, tests must be carried out using a real-time system. Although the simulations suggest that the DCNI technique is potentially of value, the overall performance of a real system depends upon other factors such as the setting of the clipper threshold and the operation of the control algorithms. These factors affect the dynamics of a conversation, in a way that cannot be replicated by 'off-line' simulations. In the real-time system it is proposed that a multi-band centre-clipper should be used in order to reduce the near-end speech distortion. Initially, a single threshold clipper whose thresholds can be set using a fast-attack, slow-decay version of the reference envelope, should be used in each band with DCNI being added after the multi-band stages.

The comfort noise is synthesised at present by using a white excitation which is then spectrally shaped, and thus it does not contain any of the periodic components that might be present in the original background sounds. Ideally, some of these components should be present to make the comfort noise sound more like the original background. As discussed in chapter 6, using an impulse-like excitation to introduce the periodicity is unlikely to be able to achieve this. One approach could be to characterise the periodic content of the background in relation to the spectral shape and variation of the background. The periodic components could then be synthesised using the modelled characteristics, but using random rather than recorded parameters.

A simpler alternative to the DCNI injection scheme proposed here, is to use a multi-band comfort noise scheme. In this scheme, white noise could be injected into each of the sub-bands directly after the clippers. In order to obtain temporal variation of the comfort noise spectrum, the power in each band could be measured and 'replayed' when comfort noise is required. An alternative approach could be to measure the variation, i.e. its distribution, and generate comfort noise based on this distribution. This type of system has the advantage that sets of coefficients representing the background spectrum need not be stored, although it is likely that the resulting comfort noise will be inferior to that produced by the DCNI system.

As noted in chapter 6, one improvement that can be made to the DCNI system is to use a different spectral estimation technique so that the stop band regions can be modelled more accurately. Currently, the simulations use band-pass filtered excitation that requires knowledge of the system bandwidth, in order to reduce the power of the comfort noise in the stop bands, and even then this does not always result in correct matching in these regions. However, the disadvantage of such a technique is that the spectral parameters are no longer obtained 'for free' and therefore the computational load is increased. It is however, possible to combine the proposed multi-band comfort noise scheme, with the DCNI system described in this thesis. In this system, comfort noise in the passband region of the spectrum would be generated by the DCNI system, whilst a two-band comfort noise system would generate comfort noise in the stop bands.

It is to be expected that the increased use of computers in homes and offices will lead to an increase in the number of subscribers who are connected digitally to their local exchanges using, effectively, a four wire full duplex channel. This might be done, for example, by means of an ISDN or other, higher bit rate, connection. This, in turn will lead to the disappearance of the need for four-wire to two-wire conversion within the network using a hybrid and, with it, will disappear the source of network echo. The only remaining echo would then be acoustically generated and, presumably, it would be the responsibility of the terminal equipment rather than of the network to remove such an echo. However, it will only be when a very high proportion of subscribers have made the change to four-wire digital connections that the network operators will be able to decrease the installed echo cancellation capacity. This is likely to be at some considerable distance in the future and in the mean time there is, therefore, very good reason to continue to attempt to improve the quality of performance of network echo cancellers.

9. References

[BART91] New Considerations For Echo Control in the Evolving Worldwide Telecommunications Network, T.L. Barto, Proceedings Telecom91, pp321-325.

[BATE91] Network Structure, J. Bateman, Telecommunication Engineering:A Structured Information Programme, The Institution of British Telecommunication Engineers, 7/1991.

[BELL82] Digital Telephony, J.C. Bellamy, J. Wiley & Sons, New York, 1982.

[BOYD88] A Speech Coder for the Skyphone Service, I. Boyd and C.B. Southcott, BT Technology Journal, vol. 6, no. 2, pp50-59, April 1988.

[BRAD63] Echo Suppressor Design in Telephone Communications. P. T. Brady and G. K. Helder. The Bell System Technical Journal, vol. 42, November 1963, pp2893-2917.

[CHOI97] Pitch Synchronous Waveform Interpolation for Very Low Bit-Rate Speech Coding, H.B. Choi, Ph.D. Thesis, Liverpool University 1997.

[CURT81] Use of a Digital Echo Canceller in the AT&T DOMSAT Intertoll Network, T. H. Curtis, S. J. D'Ambra, R. H. Tegethoff and L. E. Ashkenazi, Proceedings 5th International Conference on Digital Satellite Communications, March 1981.

[DAVI88] PhD Thesis, A. D. Davis, Liverpool University 1988

[DURB60] The fitting of time-series models, J. Durbin, Rev. Inst. Int. Statistics, vol. 28, no. 3, pp233-243, 1960

[DUTT78] A Twelve-Channel Digital Echo Canceller, D. L. Duttweiler, IEEE Transactions on Communications, vol. COM-26, no. 5, May 1978.

[EMLI63] The Effects of Time Delay and Echoes on Telephone Conversations. J.W. Emling and D. Mitchell. The Bell System Technical Journal, vol. 42, November 1963, pp2869-2891.

[FANG83] Voice Channel Echo Cancellation, G. S. Fang, IEEE Communications Magazine, December 1983.

[FREE89] The Voice Activity Detector for the Pan-European Digital Cellular Mobile Telephone Service, D.K. Freeman, G. Cosier, C.B. Southcott and I. Boyd, Proceedings IEEE ISCASP89, pp369-372, 1989.

[FUJI96] Double Talk Detection Method with Detecting Echo Path Fluctuation, K. Fujii and J. Ohga, Electronics and Communications in Japan, part 3, vol. 78, 1996.

[FURU85] High Performance Custom VLSI Echo Canceller, N. Furuya, Y. Fukushi, Y. Itoh, J. Tanabe and T. Araseki, Proceedings IEEE International Conference on Communications 1985.

[GUST97] Combined Acoustic Echo Control and Noise Reduction for Mobile Communications, S. Gustaffson and R. Martin, Proceedings Eurospeech97.

[HARRI86] A Variable Step (VS) Adaptive Filter Algorithm, R. W. Harris, D.G.Chabries and F. A. Bishop, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-34, no. 2, April 1986.

[HAYA83] Echo Canceller with Effective Double Talk Control, T. Hayashi, S. Unagami, M. Koshikawa, K. Murano, Proceedings Globecom 1983.

[HAYK96] Adaptive Filter Theory (third edition), S.Haykin, Prentice-Hall 1996.

[HILL73] Transmission Systems, M.T Hills and B.G. Evans, George Allen & Unwin Ltd.

[HUGH92] Adaptive filters – a review of techniques, P. Hughes, S. F. A. Ip, J.Cook, BT Technology Journal, vol.10, no. 1, January 1992.

[ITUT88] ITU-T Recommendation G.711 Pulse Code Modulation (PCM) of Voice Frequencies.

[ITUT90] ITU-T Recommendation G.726 40, 32, 24 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)

[ITUT91] ITU-T Recommendation G.736 Digital Circuit Multiplication Equipment using 32kbit/s ADPCM and DSI.

[ITUT93] ITU-T Recommendation G.164 Echo Suppressors.

[ITUT94] ITU-T Recommendation G.165 Echo Cancellers.

[ITUT96a] ITU-T Recommendation G.131 Control of Talker Echo.

[ITUT96b] ITU-T Recommendation G.114 One Way Transmission Time

[ITUT96c] ITU-T Recommendation G.113 Transmission Impairments

[ITUT97] ITU-T Recommendation G.168 Digital Network Echo Cancellers

[JONE94] Derivation of SQNR for μ-Law and A-Law companding, Dept. Electrical Engineering & Electronics, The University of Liverpool, 1994.

[JONE96] Reflection Coefficient Behaviour for Environmental Sounds: Simulation Results, D.J. Jones, The University of Liverpool, 1996.

[JONE97] A Network Speech Echo Canceller with Comfort Noise, D. J. Jones, S.D. Watson, K. G. Evans, B. M. G. Cheetham and R. A. Reeves, Proceedings Eurospeech 1997

[JONE98] The Design & Test Results of Real-time echo canceller, D. J. Jones, The University of Liverpool, 1998.

[KWON92] A Variable Step Size LMS Algorithm, R. H. Kwong and E.W.Johnson, IEEE Transactions on Signal Processing, vol. 40, no. 7, July 1992.

[LEWI92] Adaptive filtering – applications in telephony, A. Lewis, BT Technology Journal, vol. 10, no. 1, January 1992.

[LEWI94] Aeronautical facsimile – over the oceans by satellite, A.V. Lewis, C.D. Gostling, K.G. Evans and W.T.K. Wong, BT Technology Journal, vol. 12, January 1994.

[MAKH76] Linear Prediction: A Tutorial Review, J. Makhoul, Proceedings of the IEEE, April 1975.

[MART95] Coupled Adaptive Filters for Acoustic Echo Control and Noise Reduction, R. Martin and J. Altenhöner, Proceedings ICASSP95, pp3045-3046.

[MART96] The Echo Shaping Approach to Acoustic Echo Control, R. Martin and S. Gustafsson, Speech Communication, vol. 20, 1996, pp181-190.

[MCCO76] Principles and Applications of Adaptive Filters: A Tutorial Review, J.M. McCool and B. Widrow, IEE Conference Publication No.144, The Impact of New Technology on Signal Processing, Aviemore, 1976.

[MESS84] Echo Cancellation in Speech and Data Transmission, D. G. Messerschmitt, IEEE Journal on Selected Areas In Communications, vol. SAC-2, no. 2, March 1984.

[MINA85] A Double Talk Detection Method for an Echo Canceller, S. Minami and T. Kawasaki, Proceedings IEEE Conference on Communications 1985.

[MITC71] A full-duplex echo suppressor using centre clipping. O. M. M. Mitchell and D.A. Berkley. The Bell System Technical Journal vol. 40, May-June 1971, pp1619-1630.

[MOFF87] Echo and Delay Problems in Some Digital Communication Systems. R.H. Moffett, IEEE Communications Magazine, vol. 25, no. 8, August 1987.

[MULG88] Adaptive Filters and Equalisers, B. Mulgrew and C.F.N. Cowan, Kluwer Academic Publishers, 1988.

[MUND88] Noise reduction using frequency-domain non-linear processing for the enhancement of speech, E. Munday, BT Technology Journal, vol. 6, no.2, April 1988.

[OCHI77] Echo Canceller with Two Echo Path Models, K. Ochiai, T. Araseki and T. Ogihara, IEEE Transactions on Communications, vol. COM-25, no.6, June 1977, pp589-595

[PRES94] Numerical Recipes in C : The Art of Scientific Computing (2^{nd} Edition), W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Cambridge University Press, 1994.

[RABI78] Digital Processing of Speech, L. R. Rabiner, R. W. Schafer, Englewood Cliffs, New Jersey, Prentice-Hall, 1978.

[RICH69] Echo Suppressors for telephone connections having long propagation times. D. L. Richards and J.Hutter. Proceedings of the IEE, vol. 116, no. 6, June 1969, pp955-964.

[RIES63] Subjective Evaluation of Delay and Echo Suppressors in Telephone Communications. R. R. Riesz and E. T. Klemmer. The Bell System Technical Journal, vol. 42, November 1963, pp2919-2943.

[SOND80] Silencing Echoes on the Telephone Network. M. M. Sondhi and D.A. Berkley. Proceedings of the IEEE, vol. 68, no. 8, August 1980, pp948-963.

[TURB97] Comparison of Three Post-Filtering Algorithms for Residual Acoustic Echo Reduction, V. Turbin, A. Gilloire and P. Scalart, Proceedings ICASSP97, pp307-310.

[WATS97] A Voice Activity Detector for the ITU-T 8kbit/s Speech Coding Standard G.729, S.D. Watson, B.M.G. Cheetham, P.A. Barrett, W.T.K. Wong and A.V. Lewis, Proceedings Eurospeech 97, vol. 3, pp 1571-1574, 1997.

[WATS98] Low and Variable Bit-Rate Speech Coding for Asynchronous Transfer Mode Networks, S.D. Watson, Ph.D. Thesis, Liverpool University 1998.

[WEST96] Speech Technology for Telecommunications, F.A. Westall, R.D. Johnston and A.V. Lewis, BT Technology Journal, vol. 14, no. 1, pp9-27, January 1996.

[WIDR76] Stationary and Nonstationary Learning Characteristics of the LMS Adaptive Filter, B. Widrow, J. M. McCool, M. G. Larimore, C. R. Johnson, Proceedings of the IEEE, vol. 64, no. 8, August 1976.

[YE91] A New Double Talk Detection Algorithm Based on the Orthogonality Theorem, H. Ye and B. Wu, IEEE Transactions on Communications, vol. 39, no. 11, November 1991.