Enhancement of Acoustically Reverberant Speech Using Cepstral Methods

by

Duncan Charles Bees

A thesis submitted to the Faculty of Graduate Studies and
Research in partial fulfillment of the requirements
for the degree of Master of Engineering

Department of Electrical Engineering
McGill University
Montreal, Canada
July, 1990

Abstract

Acoustical reverberation has been shown to degrade the intelligibility and naturalness of speech. In this thesis, we discuss the application of cepstral methods to the enhancement of acoustically reverberant speech.

We first study previously described cepstral techniques for removal of simple echoes from signals. Our results show that these techniques are not directly applicable to the enhancement of speech of indefinite extent. We next recast these techniques specifically for speech. We propose new segmentation and windowing strategies, in combination with cepstral averaging, to accurately identify the acoustical impulse response. We then consider inverse filtering based on an estimated acoustical impulse response, and find that finite impulse response filters designed according to the least mean squared error criterion provide satisfactory performance. Finally, we synthesize and test an algorithm for enhancement of reverberant speech. Although significant difficulties remain, we feel that our methods offer a substantial contribution to the solution of the reverberant speech enhancement problem.

Sommaire

Il a été démontré que la réverbération acoustique dégrade l'intelligibilité et l'aspect naturel de la parole. Dans cette thèse, nous traitons de l'amélioration, par l'application de méthodes "cepstrales", de la parole ayant subi une réverbération acoustique.

En premier lieu, nous étudions certaines techniques "cepstrales" connues, dans l'élimination de simples échos d'un signal arbitraire. Nos résultats montrent que ces techniques ne sont pas directement utilisables dans l'amélioration de portions de paroles de durée indéterminée. Par la suite, nous adaptons ces méthodes au cas spécifique de la parole. Nous proposons de nouvelles méthodes de segmentation et de pondération, combinant une moyenne "cepstrale", pour identifier de faqn précise la réponse impulsionnelle acoustique. Nous considérons alors une méthode de filtrage inverse basée sur une estimation de la réponse impulsionnelle acoustique, et constatons que les filtres à réponse impulsionnelle finie conqs par la méthode des moindres carrés donnent une performance satisfaisante. Finalement, nous concevons et testons un algorithme pour l'amélioration de la parole ayant subi une réverbération acoustique. Bien que quelques problèmes demeurent, nous pensons que nos méthodes offrent une contribution substantielle à l'étude de l'amélioration de la parole ayant subi une réverbération acoustique.

Acknowledgements

I would like to thank my supervisor Professor Maier Blostein for his guidance and spirited discussion of the technical issues of this work, and for financial support during my studies. I also thank Professor Peter Kabal for the initial suggestion of my thesis topic and for his helpful conversations along the way. Daniel Boudreau translated the text of the abstract into French, and I thank him for this. This research was conducted mainly at the facilities of the Institut national de la recherche scientifique of the Université du Québec, and I am grateful to the administration of INRS-Télécommunications for making this possible.

I would like to thank my fellow students Daniel Boudreau, Vasu Iyengar, Ravi Ramachadran, and above all Guylain Roy, for their companionship and help.

Most of all, I would like to thank my wife Fiona for her patience, understanding, and love under trying circumstances.

Table of Contents

List of Figures

Chapter 1 Introduction

The communication of information via speech from one person to another involves many steps which are collectively known as the *speech train* [1]. These steps, in a "natural" context, include production of acoustic pressure patterns at the talker's mouth, transmission of the acoustic pressure to the listener's ears, and the perception of this acoustic pressure pattern by the listener as meaningful speech. In this context, interferences such as competing noise sources may occur at the transmission step which may hinder the listener's ability to perceive speech. Another source of interference may be *reverberation*, in which delayed copies of the speech acoustic waveform, called echoes, are received at the ears in addition to direct speech. In normal situations when a speaker and listener are in reasonably close proximity such reverberation is not a significant source of interference.

Through technology, the speech chain is modified to allow communication by speech at distances greater than a natural context would allow. With such modification, new sources of interference may be added to the transmission stage. For example, in the conversion of acoustic pressure to electric waveforms or to digits in telephony, electrical noise is unavoidable, and further electrical noise may be added during transmission of the electrical signal. Furthermore, the frequency bandwidth of

speech is usually diminished during such conversion or transmission. Finally, speech is normally transmitted over one channel only and is therefore presented to the listener *diotically*; in other words, the natural advantage of binaural hearing is lost. The effect of these interferences may not, under typical conditions, be severe enough to impede speech perception. However, they may make the listener more vulnerable to other interferences, such as reverberation, which under natural circumstances would not cause perceptual problems.

In this thesis, we study the problem of dereverberation of speech. The dereverberation of speech may be useful under a variety of situations, including the enhancement of speech recorded monaurally in reverberant enclosures such as rooms, and the enhancement of speech transmitted over telephone lines which has become reverberant either before transduction by microphone, or during transmission over the telephone network. We propose and study a method of dereverberation which is suitable where one has access only to a single channel of speech. The method employed is a class of digital signal processing called *homomorphic* signal processing [2]. This method, which requires that the speech acoustic pressure signal be converted to a digital sequence and subsequently re-converted to an acoustical signal after processing, may be applied to previously recorded speech or it may be applied in real time through the application of digital signal processing technology.

1.1 Effects of Reverberation on Speech Perception

Speech in enclosures such as rooms is subject to reverberation. Direct sound from the speaker to the listener is followed by reflections from walls and other surfaces which may arrive at delays from a few milliseconds to a few seconds. The perceptual effects

of reverberation are usually classified according to the delay time. Echoes which occur with delays up to a few tens of milliseconds modify the short-time spectrum by introducing periodically spaced nulls into the speech spectrum. This effect, known as spectral colouration, is particularly apparent in small rooms with highly reflective walls, because the echoes produced have high amplitudes and small delay times [3]. Beyond a delay of perhaps 50ms, echoes may be perceived as distinct copies of the direct path speech [3], and cause temporal rather than spectral distortion.

Many studies have demonstrated that reverberation degrades the intelligibility of speech under certain circumstances. In a "live" situation a normal listener is able to use his binaural hearing and possibly other screening methods to compensate for reverberation [4,5,6]. In a normal room, intelligibility between talkers and listeners of normal hearing is not affected severely by reverberation (at least in the absence of noise) [5]. However, whenever a single microphone is used in a room to record speech, the binaural hearing advantage is lost. In this case, reverberation reduces intelligibility [4,5,7].

Most studies have found that long time echoes affect intelligibility more severely than shorter time echoes. For a binaural environment, studies have shown that the earliest echoes actually improve intelligibility because they effectively increase the energy of the speech signal [8,7]. As long as the delay is less than the integration time of the ear, some of the reverberant energy may be useful. As the delay increases, the likelihood of interference between a phoneme and a succeeding phoneme increases and less of the reverberant energy is useful. This boundary occurs normally between 30 and 80 ms [5]. A practical cut-off of 50 ms between useful and disturbing energy [9] was suggested.

For monaural hearing the cut-off between useful and disturbing energy is lower [4], and the effect upon intelligibility is worse [5]. The masking level difference (difference between the noise level required to mask a signal) was found to be about 3dB between monaural and binaural hearing [6]. The intelligibility of consonants following a vowel seem to be affected most [10], which supports the theory that "overlap masking" of one phoneme by the following is the most important effect on intelligibility. A recent study suggests, however, that the echo from a consonant may also affect the intelligibility of that consonant [11], implying a degree of "self masking". This would suggest that relatively short delay echoes may be an important source of disturbance for monaural speech perception. For monaural hearing, an unpleasant "hollow" sound is also associated with short time echoes [3] producing colouration.

Perceptual degradations associated with reverberation are related to the geometry and acoustic characteristics of the room. For rooms with larger values of reverberation time T (see Section 1.2 for definition) the perception of reverberation may be worse. Larger rooms and those with hard reflective walls have larger T values. For typical small to medium rooms with T of 0.3 to 0.6 sec, a reduction in T was found to improve monaural intelligibility [5]. However, the time pattern of the echo may play a more important role than the actual value of T. Small rooms with more compact echo patterns [10] may affect intelligibility more severely.

For a monaural recording situation, the reverberant effect is also related to the source-microphone distance. The loss of intelligibility increases with this separation [12]. As the separation passes the "critical distance", the direct and reverberant energies become equal and intelligibility decreases markedly. The psychophysical preference for reverberant speech was found to decrease monotonically with source-

microphone separation [13]. The source-microphone distance is also related to the minimum phase criterion; for a minimum phase echo impulse response, the source-microphone distance must be kept small [14].

Finally, reverberation effects are more severe for hearing impaired listeners, children, and the elderly than for normal listeners. For hearing impaired listeners who had bilaterally equal hearing, reverberation affected their speech perception under binaural hearing conditions where normal hearing listeners were not affected [5].

1.2 Characterisation and Simulation of Room Reverberation

The acoustic echo property of a room has typically been characterised by the reverberation time T, which is the time for the acoustical reverberant energy to decay 60 dB relative to the direct path energy. This time may or may not be specified with respect to frequency. Typical values of T range from 1/3 second for small offices [3] up to 2.5 seconds for auditoria [15].

Because not all rooms exhibit the smooth exponential decay implied by the reverberation time measure, and because the echo pattern may change at different delay times, a more useful measure for studying the nature and effects of reverberation is the time domain acoustical impulse response. The impulse response is defined as the acoustic pressure pattern induced at a particular point in a room in response to a pressure impulse of unity magnitude at another point in the room. We thus view the room's acoustical properties between these two points as a linear, time-invariant filter which produces a known acoustical output for every input. As a practical matter, the impulse response between two points in a room is often calculated from the measured response to a continuous white noise source.

The spectral content of the impulse response may also provide useful information relating to perception of reverberation. Curtis [16] found that the subjective evaluation of the degree of colouration in speech is correlated with the second moment of the frequency spectrum of the impulse response. Curtis thus suggests that this measure be used as an indicator of the acoustic quality of rooms for hands free telephony. We employ a modified form of this measure in later sections of this thesis.

The measurement of the impulse response of a room is a complex undertaking. Often, it is simpler to simulate the impulse response on a computer. Also, computer simulation allows one to estimate an impulse response of a room which has not yet been built. Simulation requires a knowledge of the geometry and reflective characteristics of the surfaces in a room. Two methods described in the literature for the numerical simulation of room impulse responses are the ray tracing method and the image method. The ray tracing method [9] is useful for many room geometries. One considers a large number of randomly selected "rays" of sound emanating from a source and follows them through a series of reflections. This method requires a definition of the size of the receiver, so that the number of rays impinging upon it may be calculated.

The image method is especially useful for rectangular enclosures [9]. A source is considered as having a number of images caused by reflections from the six walls. This model is well justified for rigid walls, but it is an approximation for non-rigid walls [17]. The image method has been chosen to simulate acoustic impulse responses in this thesis for three reasons: the algorithm has been presented in a computer program [17] which has gained acceptance in the literature; the algorithm is well suited for rectangular enclosures such as small offices; and, it is useful for the calculation of

point to point impulse responses (no assumptions must be made as to the size of the receiver).

1.3 Previous Methods of Speech Dereverberation

Previous methods of speech dereverberation have had largely disappointing results. Typically, previous methods have focused on removing effects of either short or long time echoes. These methods can be classified into one-microphone methods and two or multi-microphone methods.

For the class of one-microphone methods of dereverberation, long-time echoes appear to have been the most successfully suppressed. The earliest such method suppressed long-time echoes by centre-clipping speech in frequency bands. This method exploits the fact that the "reverberant tails" of room impulse responses result in noise-like, low amplitude, approximately additive disturbances to speech. The aim of centre-clipping is to remove these disturbances while preserving the speech. Good results were claimed for this method [18,19].

Several studies have been published in which reverberant speech recorded with one microphone was processed using an envelope convolution model for speech [20,21]. In these studies it was assumed that the spectral amplitude envelope of the acoustical impulse response of a room was known in each of several frequency bands. The enhancement procedure was to calculate the amplitude envelope of the reverberant speech, and then apply inverse filters designed to remove the impulse response amplitude envelopes. The enhanced speech was then reconstructed using the original phase. The results were found somewhat effective for long time echoes but a loss in fine detail was noted. In a similar study [22], this method was found effective except

that zeroes outside the unit circle were not properly removed and a tone-like noise remained.

A signal processing scheme effective for minimum phase acoustical impulse responses was described in [14]. In this study, the enhancement procedure was to design a causal, linear, time-invariant filter from the (known) impulse response. When the impulse response was minimum phase, such a filter was effective in removing the "room effect". For a non-minimum phase impulse response, a filter, designed to remove the minimum phase component of equivalent spectral magnitude, was also effective in removing the reverberant quality from the speech, but left in its place a chime-like distortion. The chime distortion was attributed to peaks in the group delay of the remaining allpass function.

Signal processing schemes using more than one microphone have claimed somewhat better results. Flanagan [23] attempted to remove early echoes using a two-microphone technique. In each of many frequency bands, the spectral amplitudes of the two signals were compared, and that with the higher amplitude was selected to be the contribution for the reconstructed speech. This method exploits the periodic nature of the spectral distortion of speech caused by a simple echo. Because the two microphones, spaced at different locations, have echoes of different delays, nulls appear at regular but different intervals in the spectra of the microphone outputs. Thus in any particular frequency "bin", there is a good chance that one or both of the spectra does not contain a null. This technique, which may be extended to more than two microphones, was only effective for very simple and short delay patterns.

A two microphone technique designed for both short and long time reverberation exploited the uncorrelated nature of the tail of the impulse response of a room at two

different locations [24]. Processing was again in frequency bands. For each band the correlation between the two microphone signals was computed, and used as a gain factor for that band. The goal was to suppress spectral bands with low correlations and presumably long time echoes. To remove short time echoes the phase difference between the two signals in each band was eliminated and the amplitude was averaged. This procedure appeared effective, for rooms with T from 0.1 to 2.0 seconds, in reducing perception of reverberation. But a separate study found no increase in intelligibility [25] for this technique, and also found that vowel to consonant masking was not reduced. The latter finding may indicate that this method is ineffective for shorter delay echoes.

Finally, another study exploiting the uncorrelatedness between the reverberant tails at two locations used LMS (least mean-squared-error) adaptive filters to provide the correlated component of the two microphone signals [26]. Processing was again performed in frequency bands. Although no claim was made to remove short time delays, performance comparable to [25] was cited.

1.4 Outline of Thesis Research

None of the signal processing techniques outlined above can be used to enhance, for both short and long delays, reverberant speech recorded with a single microphone without knowledge of the characteristics of reverberation. In this research, the object was to develop such a method, which could potentially have wide use, in for example hands-free telephony. In view of the restrictions that only a single channel of speech be available and that the reverberation characteristics not be known *apriori*, we selected the homomorphic technique of complex cepstral filtering as a candidate. This

technique was developed in the 1960's by Oppenheim as a method of deconvolution and was applied by Schafer to the removal of simple echoes from speech [27].

1.4.1 Goals

The goal of the enhancement of reverberant speech can be considered twofold: first, we wish to improve the intelligibility of speech; and second, we wish to process the speech in a way that makes it more pleasing to listen to. In the definition of "pleasantness" we may consider factors such as hollowness, distortion, and fatiguing effect. As discussed in Section 1.1, the effects of reverberation upon intelligibility and pleasantness are not clearly understood. In general, we know that discrete, audible copies of speech degrade intelligibility. In typical rooms, this sort of reverberation is not usually present, but typical room reverberation can have adverse effects upon speech intelligibility especially when a monaural recording is made. Less is known about the pleasantness of speech under reverberation, but it is known that under many conditions reverberation is considered annoying by listeners. Although in some circumstances a small degree of short delay reverberation may be beneficial to binaural perception, this is much less clear for monaural perception. With respect to this point, a processing system will allow for arbitrary amplification and so we will not consider amplifying effects of reverberation to be useful. In view of the above factors, the goal of the processing system will be to eliminate reverberation over as broad a range of delays as possible.

1.4.2 Processing Technique

The complex cepstrum is described in [2]. It is a two-sided (non-causal), infinite

sequence related to the time domain sequence by a non-linear, homomorphic transformation. For the discrete time signal $x(n)$, the characteristic system [2] by which the complex cepstrum is calculated is shown in Figure 1.1.

Fig. 1.1 Block diagram for calculation of complex cepstrum

The complex cepstrum has several properties which make the technique a candidate for deconvolution of echoes from speech. First, signals which are combined convolutionally in the time domain have complex cepstra which are combined additively. As a result, deconvolution is reduced to subtraction in the cepstrum. Second, the complex cepstrum is a measure of the "frequency" of variation in the log spectrum, and so signals which vary slowly in the log spectrum may be separated from quickly varying signals by windowing the complex cepstrum. Speech is usually considered to be primarily slowly varying in the log spectrum and has a complex cepstrum concentrated about the cepstral origin. Echoes which are delayed from the direct path speech can be represented by impulse responses which have cepstra concentrated farther from the cepstral origin.

Complex cepstral filtering techniques suitable for the separation of speech and simple echoes are described in [2]. They involve computation of the complex cepstrum, removal of components in the complex cepstral domain, and reconversion to the time domain. We found these techniques, however, to be unsuitable for the dereverberation of speech subject to acoustic reverberation. The major problems involve

the segmentation and windowing procedures in the time domain, and the imperfect separation between the speech and reverberation complex cepstrum.

In the research described in this thesis, new approaches to solving the above problems are presented. These techniques allow a much more accurate estimate of the original speech than was possible with previous techniques. The results of this research also show that significant difficulties remain, and that complete removal of reverberation is not yet possible. However, we believe that the methods which we have developed offer significant promise in many dereverberation applications.

Chapter 2

Echo Removal with Complex Cepstrum Methods

The removal of simple patterns of discrete echoes from speech using complex cepstral techniques was developed by Schafer [27]. This work was based on the development by Oppenheim of homomorphic systems. Techniques based on Schafer's work are attractive candidates for the dereverberation of speech because they require as input only the reverberant speech, and because they require little knowledge about the reverberation to remove it.

In this chapter, the computation and properties of the complex cepstrum will be reviewed. The suitability of Schafer's methods to the general dereverberation problem will be evaluated.

2.1 Complex Cepstrum: Introduction

The complex cepstrum is a sequence in the *quefrency* domain which is the result of a homomorphic transformation on a sampled time signal. A homomorphic transformation is one that alters a rule of combination between signals [2]. In the case of the complex cepstrum, the transforming operation is the complex logarithm

in combination with a forward and an inverse z-transform. The convolutional rule of combination in the time domain is transformed into the additional rule of combination in the quefrency domain. Since the quefrency domain and the time domain are both related to the z domain via an inverse z-transform, quefrency has the same units as time. The *characteristic system* relating the complex cepstrum to the time domain sequence is shown in Figure [2.1].

Fig. 2.1 Characteristic system relating complex cepstrum and time domain

Because convolution in the time domain is "converted" into addition in the complex cepstral domain, deconvolution of time domain signals may be achieved by passing them through the characteristic system, simple subtraction in the cepstral domain, and passing the result through an inverse characteristic system. This would be no more effective than division in the frequency domain, however, were it not for the fact that time domain signals have complex cepstra concentrated in locations dictated by their *log spectral* properties. Thus, signals may be separated in the cepstrum if their log spectra have properties different enough that their cepstra are concentrated in different locations. In this case, deconvolution of one component from another may be achieved by simply multiplying one cepstral region by zero and another by unity. This procedure, known as cepstral windowing, is the basis of Schafer's echo removal techniques.

The term cepstrum was coined by Bogert, Healey, and Tukey [28] to mean the

power spectrum of the log of the power spectrum of a signal. This definition of the cepstrum is also useful for the detection of echoes but is of more limited use because the phase information of the signal is not used. The word cepstrum will generally be used in this thesis to refer to complex cepstrum.

2.2 Principle of Cepstral Filtering of Speech

The location of cepstral components is dictated by the log spectral properties of the time domain signals. Specifically, the log spectrum of a signal is considered as the sum of components which are slowly varying and those which are quickly varying. As a consequence of the definition of the characteristic system, slowly varying log spectral have cepstral components concentrated about the quefrency axis origin, and quickly varying components have cepstra concentrated farther from this origin. Just as the value of the Discrete Fourier Transform $X(k_0)$ expresses the content of a time domain signal which varies with period $\frac{2\pi}{k_0}$, the value of the cepstrum at quefrency n_0 expresses the content of the log spectrum which varies with frequency $\frac{2\pi}{n_0}$.

The utility of the complex cepstrum for speech enhancement or analysis is seen by considering the nature of the log spectrum of speech. Voiced speech is usually considered to be a convolution of an impulse train $p(n)$ and an impulse response $v(n)$ representing the combined effects of the glottal wave shape, the vocal tract impulse response, and the radiation impulse response due to the lips [29]. Ignoring time window effects, the z-transform is

$$S(z) = P(z)V(z) \tag{2.1}$$

Because the time extent of $v(n)$ is relatively small, $V(e^{j\omega})$ varies slowly with ω. On the other hand $P(e^{j\omega})$ varies periodically with ω, with spacing between harmonics given by 2π times the reciprocal of the pitch period. After the logarithm is applied the two components are additive and a linear filter acting on the log spectrum can be designed to separate them. If the filter is low pass the logarithm of $V(z)$ may be retained and that of $P(z)$ removed. Such techniques have been successfully applied to speech analysis [30] and pitch determination [31].

The separation of echo from speech is an extension of the above principle. The effect of an echo on speech can be described by a convolution of speech with an impulse response $h(n)$. For a single echo of time delay n_0 and amplitude α the spectrum of the impulse response is

$$H(e^{j\omega}) = 1 + \alpha e^{-j\omega n_0} \tag{2.2}$$

The spectrum and the log spectrum are periodic with period $\frac{2\pi}{n_0}$. If n_0 is greater than the pitch period, the log spectrum of the echo varies faster than that of speech and a high pass filter operating on the log spectrum can separate them. The filtering operation may be carried out by convolution with a kernel in the log frequency domain, or it may be carried out by multiplication in the complex cepstral domain. It is computationally more efficient to multiply in the cepstrum than to convolve directly [27].

2.3 Computation of Complex Cepstrum

The complex cepstrum $\hat{x}(n)$ is a two sided, real sequence which has, in general, non-zero values for all positive and negative values of quefrency n. It is related to

the sampled time domain signal $x(n)$ by the series of transformations known as the characteristic system which are shown in Figure [2.1]. We compute an approximation to $\hat{x}(n)$ by employing the Discrete Fourier Transform to evaluate samples of the complex logarithm of the z-transform around the unit circle, as shown in Figure [2.2]. In this way, because of the infinite extent of $\hat{x}(n)$, aliasing is introduced in the computation. Fortunately the cepstrum has an envelope with decays rapidly in time, and aliasing may be controlled by zero padding $x(n)$ and using longer DFT's.

Fig. 2.2 Calculation of approximation to complex cepstrum

The definition of the complex cepstrum requires that $\hat{x}(n)$ be representable by the convergent power series $\widehat{X}(z)$. Consequences of this requirement are that $\widehat{X}(z)$ must be uniquely defined, and must be continuous. We must be careful in the calculation of $\widehat{X}(k)$ in order to satisfy these consequences and provide unique samples of a continuous $\widehat{X}(z)$. This care is required because of the ambiguity inherent in the imaginary component of the resultant of the complex logarithm. The complex logarithm gives

$$\widehat{X}(k) \;=\; \log|X(k)| + j\arg X(k) + j(2\pi i),\; i \text{ any integer} \tag{2.3}$$

The definition adopted to resolve the ambiguity and satisfy the requirement for continuity is that $\mathrm{Im}[\widehat{X}(k)]$ be defined to be the *unwrapped phase* of $X(k)$, where the unwrapped phase is the integral of the derivative of the phase of $X(k)$. The calculation of the unwrapped phase is performed by the adaptive integration algorithm of Tribolet [32], which adds multiples of 2π to the principal phase value to satisfy this

definition. It should be noted that the adaptive integration of the unwrapped phase is a computationally intensive task.

There are two parameters of the signal $x(n)$ which are not represented by $\hat{x}(n)$: sign, and linear phase. These two parameters are determined during the calculation of the complex logarithm and "removed" from the signal. Thereafter, they must be retained along with $\hat{x}(n)$ in order to reconstruct the signal.

2.3.1 Computation of Complex Cepstrum Without Phase Unwrapping

There are several methods which allow computation of the complex cepstrum without explicitly calculating the unwrapped phase. In one method, instead of calculating $\hat{x}(n)$ directly, $n\hat{x}(n)$ is calculated first [33]. This calculation is simpler because only the derivative of $\widehat{X}(z)$ is needed, for which there is no ambiguity. However, we found that this method leads to unacceptable aliasing distortion, because the envelope of $n\hat{x}(n)$ decays much more slowly than does $\hat{x}(n)$.

For functions $x(n)$ which are minimum phase, the complex cepstrum is readily calculated without phase unwrapping. A minimum phase function has an entirely causal complex cepstrum (as shown in Section 2.4.2). For causal functions, the even part,

$$\hat{x}_e(n) \;=\; \left[\frac{\hat{x}(-n)+\hat{x}(n)}{2}\right] \tag{2.4}$$

completely specifies $\hat{x}(n)$. It is shown in [2] that

$$\mathrm{DFT}[\hat{x}_e(n)] \;=\; \mathrm{Re}\,(\mathrm{DFT}[\hat{x}(n)]) \tag{2.5}$$

The complex cepstrum of a minimum phase function can thus be computed from twice the inverse Discrete Fourier Transform of the magnitude of the logarithm of $X(k)$. This fact is exploited in later sections of this thesis.

2.4 Properties Relevant to Cepstral Dereverberation

Some basic properties of the complex cepstrum and of exponential weighting of convolved sequences are described here. A full list of such properties is contained in [2].

For illustration let us deal with finite length signals such as $h(n)$ which has complex cepstrum $\hat{h}(n)$. The signal $h(n)$ is assumed mixed phase, with zeroes inside and outside the unit circle in the z-plane, and can be constructed from its minimum and maximum phase components:

$$h(n) = h_{min}(n) * h_{max}(n) \tag{2.6}$$

These properties are:

2.4.1 Cepstral Additivity

Let $x(n) = s(n) * h(n)$. Then

$$\begin{aligned} X(z) &= S(z)H(z) \\ \widehat{X}(z) &= log[S(z)H(z)] \\ &= log[S(z)] + log[H(z)] \\ &= \widehat{S}(z) + \widehat{H}(z) \end{aligned}$$

So,

$$\hat{x}(n) = \hat{s}(n) + \hat{h}(n) \tag{2.7}$$

2.4.2 Minimum/Maximum Phase Separability

It it shown here that minimum phase components have cepstra which occupy the right side or causal region of the quefrency axis, and that maximum phase components

occupy the anticausal region.

$$h_{min}(n) \iff \hat{h}(n)u(n+1)$$

$$h_{max}(n) \iff \hat{h}(n)u(-n-1)$$

where

$$u(n) = \begin{cases} 1 & n \geq 0 \\ 0 & n < 0 \end{cases}$$

so that

$$\hat{h}_{min}(n) + \hat{h}_{max}(n) + \hat{h}(0) = \hat{h}(n) \tag{2.8}$$

This can be seen in the following way. The z-transform $\widehat{H}(z)$ of the complex cepstrum has no zeroes and has poles which occur at the poles and zeroes of $H(z)$. $H_{min}(z)$ has poles and zeroes only inside the unit circle and so $\widehat{H}_{min}(z)$ has poles inside the unit circle only. When the inverse z-transform is evaluated on the unit circle it thus produces a causal sequence $\hat{h}_{min}(n)$. Because $H_{max}(z)$ has poles and zeroes only outside the unit circle, $\hat{h}(n)$ is an anticausal sequence. The term at the origin $\hat{h}(0)$ reflects the energy of the signal (the average value of the log magnitude spectrum). We usually adopt the convention that $\hat{h}(0)$ is associated with the causal portion of $\hat{h}(n)$.

2.4.3 Complex Cepstrum of Single Zero Function

The complex cepstrum of a finite length sequence is the sum of the complex cepstra of each of its zeroes. These zeroes can be separated into minimum and maximum phase depending on their location with respect to the z-plane unit circle. The cepstrum of a single zero sequence is calculated [2] using the Taylor series

$$\begin{aligned} log[1+f(z)] &= 1 + f(z) - \frac{1}{2}f^2(z) + \frac{1}{3}f^3(z) - \ldots \\ |f(z)| &< 1 \end{aligned} \tag{2.9}$$

The two cases are:

Zero Inside Unit Circle (Minimum Phase)

$$
\begin{aligned}
h(n) &= \delta(n) + a\delta(n-1) && |a| < 1 \\
H(z) &= 1 + az^{-1} \\
\widehat{H}(z) &= log[1 + az^{-1}] \\
&= az^{-1} - \frac{a^2}{2}z^{-2} + \frac{a^3}{3}z^{-3} - \ldots \\
\widehat{h}(n) &= a\delta(n-1) - \frac{a^2}{2}\delta(n-2) + \frac{a^3}{3}\delta(n-3) - \ldots
\end{aligned}
\tag{2.10}
$$

The minimum phase zero thus has a complex cepstrum consisting of an infinite series of pulses in the causal side of the cepstrum.

Zero Outside Unit Circle (Maximum Phase)

$$
\begin{aligned}
h(n) &= \delta(n) + a\delta(n-1) && |a| > 1 \\
H(z) &= 1 + az^{-1} \\
&= az^{-1}(1 + \frac{1}{a}z) \\
\widehat{H}(z) &= log[az^{-1}(1 + \frac{1}{a}z)] \\
&= log[a] + log[z^{-1}] + log[1 + \frac{1}{a}z] \\
&= log[a] + log[z^{-1}] + \frac{1}{a}z - \frac{1}{2a^2}z^2 + \frac{1}{3a^3}z^3 - \ldots \\
\widehat{h}(n) &= log[a]\delta(n) + \frac{1}{a}\delta(n+1) - \frac{1}{2a^2}\delta(n+2) + \frac{1}{3a^3}\delta(n+3) - \ldots
\end{aligned}
\tag{2.11}
$$

The maximum phase zero thus has an entirely anticausal cepstrum. Since the term $log[z^{-1}] = log[e^{-j\omega}]$ is an additive linear phase term $-j\omega$ when evaluated on the unit circle, it is excluded from the computation of the complex cepstrum. This linear phase term determines the

time position of the original function. In order to invert the cepstrum the linear phase component must be known.

A useful consequence of the above identities is that the linear phase component of an impulse response with no pure delay directly identifies the number of z-plane zeroes outside of the unit circle. This information is a byproduct of the computation of the complex cepstrum with the phase unwrapping method.

2.4.4 Complex Cepstral Envelope

It is shown in [2] that for functions with rational z-transforms the complex cepstrum envelope $|\widehat{h}(n)|$ decays at least as fast as $|\frac{1}{n}|$. Thus when viewing complex cepstra when graphed against quefrency, it is normal to scale the cepstrum by a linear function of n. In this thesis, most plots are scaled by $|n|$. This scaling is also applied before cepstral peak-picking or peak-elimination described in later sections.

2.4.5 Exponential Weighting of Convolved Sequences

The exponentially decreasing, truncated window plays a major role in cepstral analysis, partly because of the properties described here. First, under exponential weighting, convolution of sequences is preserved. For $x(n) = s(n) * h(n)$, each of duration N samples,

$$\begin{aligned} \gamma^n x(n) &= \gamma^n \sum_{m=0}^{N-1} s(m)h(n-m) \\ &= \sum_{m=0}^{N-1} \gamma^m s(m)\gamma^{n-m}h(n-m) \\ &= \gamma^n s(n) * \gamma^n h(n) \end{aligned} \tag{2.12}$$

Thus the cepstrum of the weighted signal $\gamma^n x(n)$ is the sum of the cepstra of the two weighted signals $\gamma^n s(n)$ and $\gamma^n h(n)$.

Second, exponential weighting with $|\gamma| < 1$ moves z-plane zeroes inward radially because the z-transform of $\gamma^n h(n)$ is $H(\frac{z}{\gamma})$. Minimum phase zeroes remain inside the unit circle after weighting, while maximum phase zeroes may or may not be moved inside the unit circle. The effect on the cepstrum is

$$\gamma^n h_{min}(n) \iff \gamma^n \hat{h}_{min}(n)$$

$$\gamma^n h(n) = \gamma^n h_{min}(n) * \gamma^n h_{max}(n) \iff \gamma^n \hat{h}_{min}(n) + \hat{g}(n)$$

where $\hat{g}(n)$ is the complex cepstrum of $\gamma^n h_{max}(n)$. If the exponential weighting is severe enough that all of the zeroes of $\hat{h}_{max}(n)$ are moved inside the unit circle, then $\hat{g}(n)$ will be entirely causal. In summary, exponential weighting tends to reduce in amplitude the cepstrum of minimum phase components and tends to "flip" to the causal region the cepstrum of maximum phase components.

Exponential weighting is often used in an effort to transform the analysed signal from mixed phase to minimum phase. A direct benefit of this is that the cepstrum may be computed from the spectral magnitude only.

2.5 Detection and Removal of Echo using Complex Cepstrum

Schafer [27] and others [34,35] have developed techniques to exploit the properties of the complex cepstrum to remove patterns of echoes from speech. These techniques can be grouped into two classes: those which seek to identify the echoes through cepstral peak picking, so that they may be removed in the cepstrum; and those which remove all cepstral terms in certain cepstral ranges, thereby assuming that the echo

cepstrum is restricted to one range and the speech cepstrum in another. In this section, we describe these techniques.

Both methods model the cepstrum of (voiced) speech as the sum of components due to the combined impulse response $v(n)$ and the pitch excitation $p(n)$, as described in Section 2.2. The cepstrum $\hat{v}(n)$ is assumed to be restricted to the "low-pass" region around $n = 0$, and $\hat{p}(n)$ is assumed to be concentrated in large peaks at the pitch period T, and rapidly decaying peaks at multiples thereof. Echoed speech is modelled as the convolution of clear speech with an impulse response $h(n)$:

$$x(n) = p(n) * v(n) * h(n) \tag{2.13}$$

Since all components are combined convolutionally, their cepstra are combined additively. The key assumption made is that the cepstrum of $h(n)$ is confined to regions outside the region containing most speech cepstrum, that is, outside $|n| < T$. For the peak-picking methods, it is further assumed that $\hat{h}(n)$ is composed of a series of peaks which are discernable from the background cepstrum.

We consider a sequence of speech which has a single echo of amplitude α and delay n_0. For $\alpha < 1$ we have from (2.2) and (2.9) the log spectrum of the echo

$$\widehat{H}(e^{j\omega}) = 1 + \alpha e^{-j\omega n_0} - \frac{\alpha^2}{2} e^{-j2\omega n_0} + \frac{\alpha^3}{3} e^{-j3\omega n_0} - \ldots \tag{2.14}$$

Because the echo amplitude is less than unity all of the impulse response zeroes lie within the unit circle and the impulse response is minimum phase. The complex cepstrum

$$\hat{h}(n) = \alpha\delta(n - n_0) - \frac{\alpha^2}{2}\delta(n - 2n_0) + \frac{\alpha^3}{3}\delta(n - 3n_0) - \ldots \tag{2.15}$$

is an infinite, causal series of decaying equispaced pulses of alternating sign. In particular, there are no non-zero components in the region $n < n_0$. Figure 2.3 shows

the (un-scaled) complex cepstrum calculated using a short sequence of echoed speech sampled at 8000 samples per second, with an echo of amplitude $\alpha = 0.9$ and delay $n_0 = 200$ samples. In this calculation, the entire record of the speech plus its echo were used to calculate the complex cepstrum.

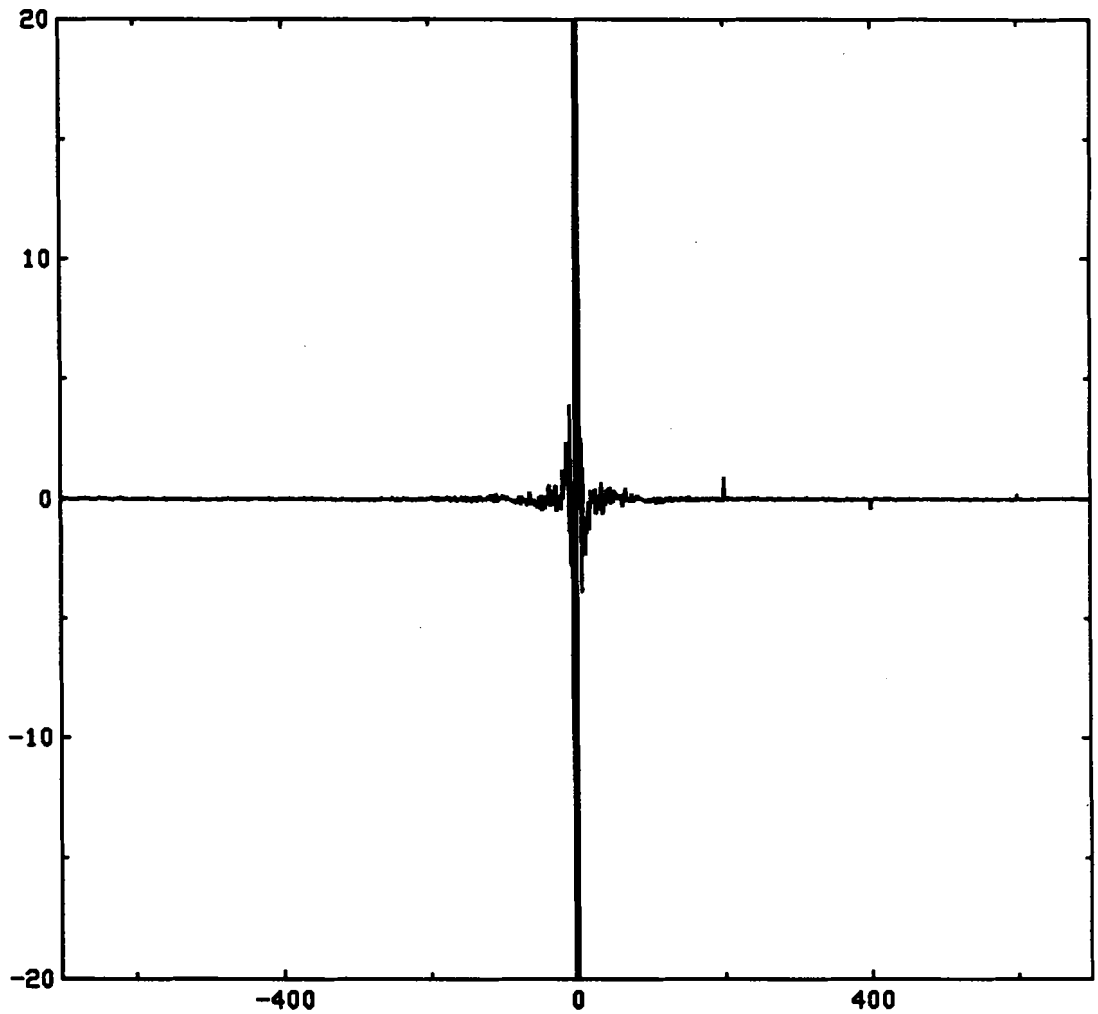

Fig. 2.3 Complex cepstrum of echoed speech

The plot shows the quick fall-off of the amplitude of the cepstrum. The first three cepstral echo peaks are clearly identifiable at 200, 400, and 600 quefrency samples. The speech cepstrum is seen to be concentrated about $n = 0$. Cepstral peaks due to

the pitch train $p(n)$ can be seen at ± 34 samples, approximately. Thus the pitch of the (female) speaker during this interval was about $8000/34 = 235$ Hz. The first pitch "harmonics" at ± 68 samples can also be seen. Beyond this region, the speech cepstrum dies out rapidly. In this example, it is a simple matter to "detect" and remove the cepstral peaks due to the echo by either setting the cepstrum at these points equal to zero, or performing interpolation with the neighbouring points. The remaining cepstrum, when transformed to the time domain, would yield the clear speech.

For multiple echoes, the cepstrum of the impulse response is a much more complicated series of pulses. If $h(n)$ is minimum phase, the first cepstral pulse occurs at the delay time of the first echo [2]. For example $h(n) = \delta(n) + \alpha_1\delta(n-n_1) + \alpha_2\delta(n-n_2)$, which is guaranteed to be minimum phase [27] for $|\alpha_1 + \alpha_2| < 1$, has cepstral pulses at quefrencies n_1, n_2, $2n_1$, n_1+n_2, $2n_2$,.... These pulses may not be easily identifiable from the background cepstrum if they are of low amplitude or if they are numerous. In particular, for impulse responses representing the acoustic response of a room, an impulse response will have many terms and the peak-picking of individual cepstral pulses may be difficult. In this case, the low-passing "window" method could be used. With this approach, all cepstral samples outside of, say, $|n| < 200$ would be set to zero. This "low-passing" approach would result in some distortion due to elimination of remaining speech components in the high-pass region. However, if the cepstrum of the echo is confined to the high-pass area, and the cut-off value does not truncate the speech cepstrum too severely, it may still provide a good estimate of the clear speech.

Mixed phase impulse responses pose further complications. In these cases, the cepstrum of the impulse response has both causal and non-causal components. These

may be removed using either of the above methods, but any linear phase component due to maximum phase zeroes of the impulse response will not be identifiable. Thus, the resulting estimated speech may be displaced in time by an arbitrary positive or negative delay. Furthermore, there may be echo components inside the pitch period region, and these are difficult to separate from the speech cepstrum. For these reasons, exponential weighting is often applied [27,36] to attempt to convert all echo components to minimum phase.

2.5.1 Cepstral Filtering of Continuous Speech

In the previous section cepstral echo removal techniques are explained which are effective in cases of finite records of speech. However, for echo removal from continuous reverberated speech $x(n) = s(n) * h(n)$, the cepstral filtering must be applied on a segment by segment basis. In this case, special care must be taken to compensate for the errors introduced by truncating each segment $x_i(n)$. The following development is due to Schafer [27].

We define the sequence $x_i(n)$ to be a segment of length N of the signal $x(n)$,

$$x_i(n) = \begin{cases} x(n) & iN \leq n < (i+1)N \\ 0 & n \text{ otherwise} \end{cases} \tag{2.16}$$

This sequence is composed of a convolution of a segment $s_i(n)$ of the original, clear speech with the impulse response $h(n)$, plus an error term

$$x_i(n) = s_i(n) * h(n) + e_i(n) \tag{2.17}$$

The error term $e_i(n)$ is due to the truncation of the echo of $s_i(n)$ which occurs at the end of the segment, and to the intrusion of the echo of $s_{i-1}(n)$ at the beginning

of the segment. These errors are the amount by which $x_i(n)$ deviates from a pure convolution.

In Schafer's method, the complex cepstrum of $x_i(n)$ is calculated, and either peak elimination or low-pass windowing is applied in order to remove the cepstrum of the echo. If the cepstral terms removed are approximately those due to the echo impulse response $h(n)$, then the cepstral filtering procedure is equivalent to applying a linear inverse filter to remove $h(n)$. Thus we can say that the result after transformation, $\tilde{s}_i(n)$, approximates

$$\tilde{s}_i(n) \approx s_i(n) + e_i(n) * h^{-1}(n) \tag{2.18}$$

The last term on the right side of equation (2.18) represents an error term which would be present in this segment were we to use $\tilde{s}_i(n)$ as the estimate for $s_i(n)$. The error can be removed by taking advantage of the fact that the truncation error at the end of segment $x_{i-1}(n)$ is the negative of the intrusion error at the beginning of segment $x_i(n)$. Therefore, a correction sequence generated as a byproduct of the generation of $\tilde{s}_{i-1}(n)$ added to $\tilde{s}_i(n)$ should cancel this error term. Thus continuous filtering may be carried out under the assumption that $h(n)$ is relatively constant from segment to segment and is minimum phase. If $h(n)$ is mixed phase, $h^{-1}(n)$ is two sided and correction sequences from both $\tilde{s}_{i-1}$ and $\tilde{s}_{i+1}$ must be used to produce the estimate for $s_i(n)$.

This filtering procedure assumes that the cepstrum removed in the filtering process corresponds closely to the actual echo cepstrum. It does not model any distortions in the calculated cepstrum caused by finite block length processing but assumes that such distortions are small if the block length is sufficiently large. In the results pre-

sented in [27] it was indicated that this assumption was valid for certain types of echo functions. We evaluate the effectiveness of these techniques in the next section.

2.5.2 Evaluation of Cepstral Filtering

In order to illustrate the cepstral filtering methods described above and to evaluate their potential effectiveness for the general dereverberation problem, we consider a sentence of speech digitized at 8000 samples per second to which a simple echo of impulse response $h(n) = \delta(n) + 0.5\delta(n - 250)$ has been digitally added. The sentence, as spoken by a female, is "Cats and dogs each hate the other", and consists of 17664 samples. A segment of this speech is shown in Figure 2.4(1). In Figure 2.4(2) is shown the additive error of the echoed speech compared with the original speech. It should be noted that the echo used, which gives the speech a slightly unpleasant quality, is simpler than the impulse responses representing reverberation which would be encountered in, for example, typical rooms. However, the delay of 250 samples, or 31 ms, is a value which would be expected to contribute to spectral "colouration" and for which reverberation components would be expected in a typical room.

Three cepstral filtering experiments, using different windowing and cepstral filtering techniques, were conducted with the same echoed speech as input. The first and second experiments used peak-elimination which could be used in general if the impulse response had well defined cepstral peaks. The third experiment uses cepstral low-passing which could be used in general for impulse responses having broadly dispersed cepstra confined to the cepstral high-pass region. In each experiment, the complex cepstrum was calculated for speech segments of length 2048 samples, and the cepstrum was calculated with DFT's of length 4096 points using the phase un-

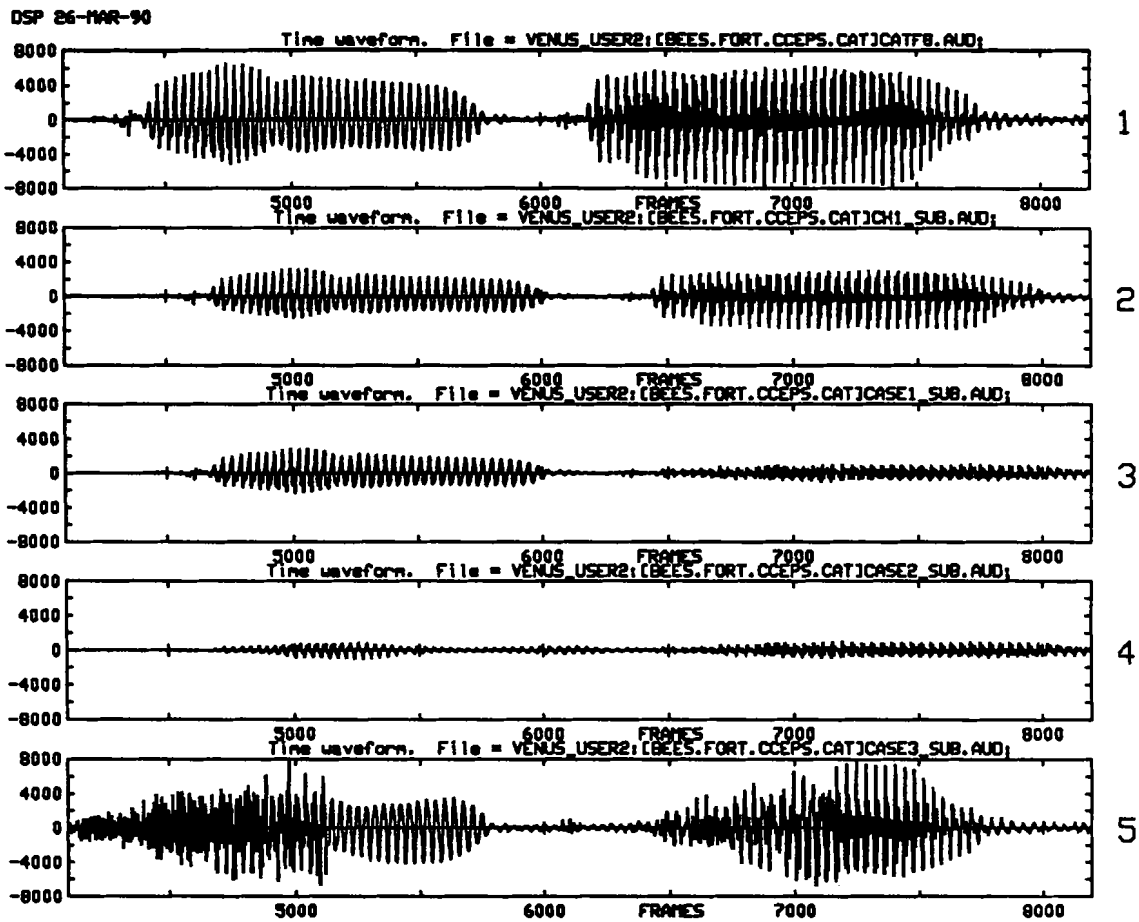

Fig. 2.4 Speech waveforms:
(1) Original speech
(2) Additive error for echoed speech
(3) Additive error for cepstrally filtered speech, experiment 1
(4) Additive error for cepstrally filtered speech, experiment 2
(5) Additive error for cepstrally filtered speech, experiment 3

wrapping programs contained in [37]. These procedures are summarized in Figure 2.5.

In the first experiment, no window function was multiplied with the speech data. The cepstrum calculated for each of the 9 segments, with the average over all 9 shown at the bottom, is shown in Figure 2.6. In this figure, the cepstrum has been linearly

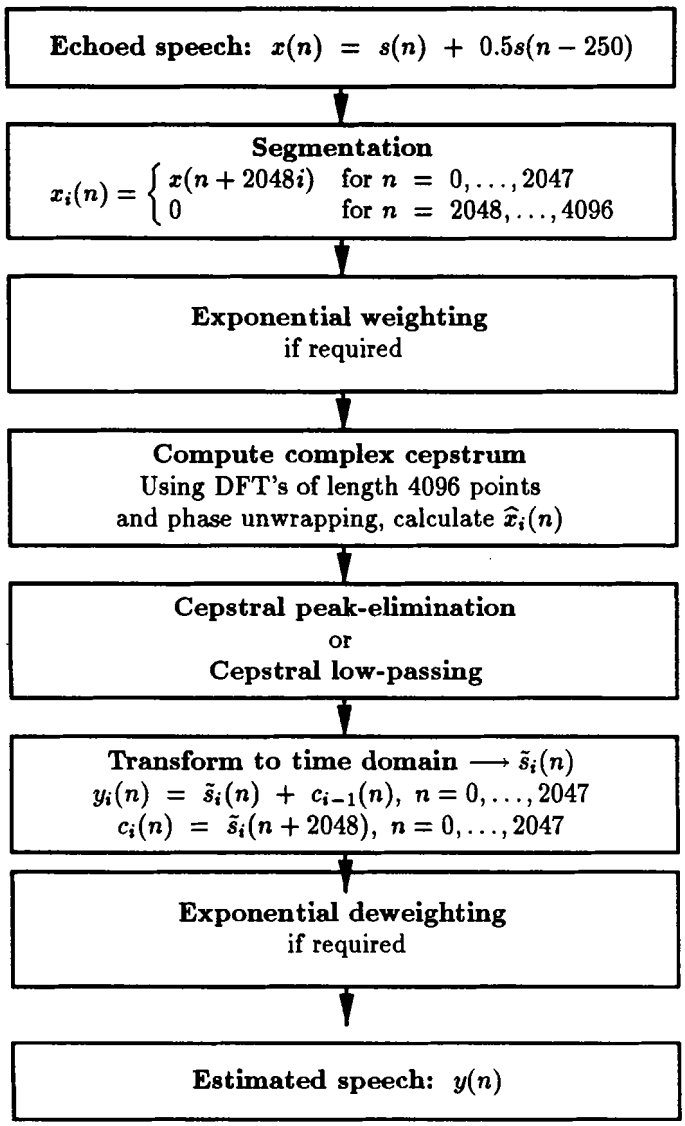

Fig. 2.5 Cepstral filtering procedures for evaluation

rescaled as described in Section 2.4.4. As can be seen from Figure 2.6, a cepstral peak is visible in most of the cepstral segments at $n = 250$. However, at $n = 500$, a peak is discernible for only a few segments. At higher multiples of 250, which are not shown in the plot, no peaks could be discerned from the background "noise". A surprising feature of this plot is that cepstral peaks are visible for many segments at $n = -250$, which is not expected for the minimum phase echo. For this experiment,

the causal cepstral peaks which were discernible from the background "noise" were set to zero. The correction described in Section 2.5.1 was used with 1024 points of correction sequence generated for each segment. The resulting output speech showed little reduction in reverberation and the sound was not improved. The time domain error in comparison with the original speech is shown in Figure 2.4(3).

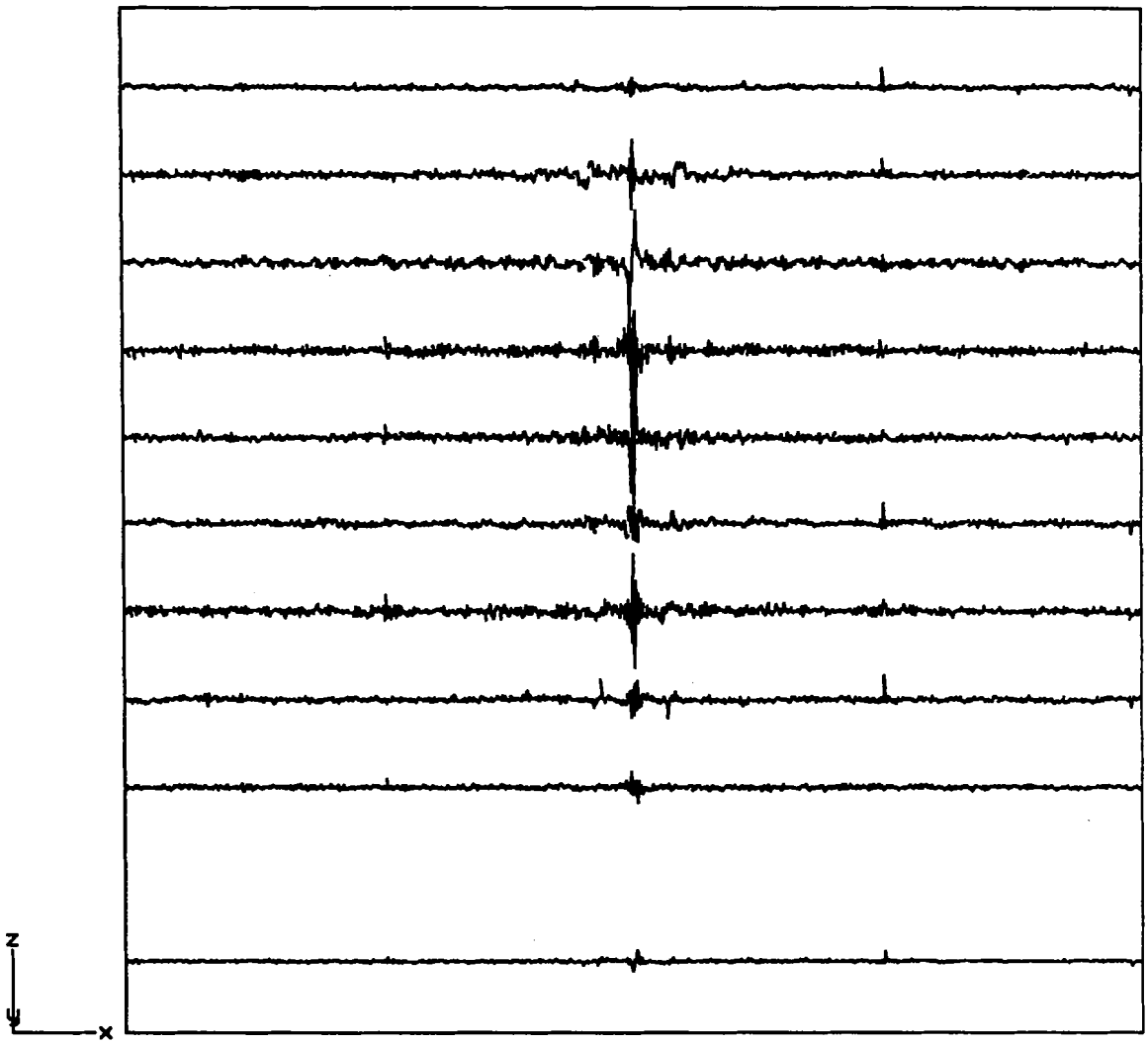

Fig. 2.6 Complex cepstrum for $n = -512, \ldots 511$, calculated with rectangular windows

In the second and third experiments, an exponential weighting was applied to the

echoed speech using a weighting function of 0.999^n (this was the procedure followed for the echo removal experiments in [27]). The resulting cepstrum is shown in Figure 2.7. In comparison to Figure 2.6 most of the cepstral peaks at $n = 250$ are slightly larger. Again, however, there are few visible peaks at $n = 500$, and none at larger multiples. The peaks seem to be restricted to the causal region in this plot, as would be expected from theory. In the second experiment the cepstral peaks were set to zero; in the third experiment, all cepstral components at quefrencies greater than $n = 200$ were zeroed. For both cases, after cepstral filtering and reconversion to the time domain the effect of exponential weighting was undone by multiplication with 0.999^{-n}. Again, correction sequences of 1024 points were used. The results for the second experiment were similar to but better than the first experiment, in that more of the echo was removed and with less distortion. The third experiment produced highly distorted speech which was much worse than the original echoed speech in quality. The time domain errors in comparison with the original speech are shown in Figure 2.4(4) and 2.4(5) for the second and third experiments.

In order to verify that the filtering mechanism was sound, an experiment was conducted in which the theoretical cepstrum of the echo was subtracted from the calculated cepstrum. For both rectangular window and exponential weighting cases, the resulting filtered speech removed the echo almost completely and with little distortion. Thus we conclude that the failure of the cepstral filtering method is due to a failure to calculate the echo cepstrum correctly. As shown in Table 2.1, for both rectangular and exponential cases the average values of the cepstral peaks attributable to the echo were lower than the values which should have been obtained. Furthermore, as seen in the plots, the peaks were highly variable from segment to segment.

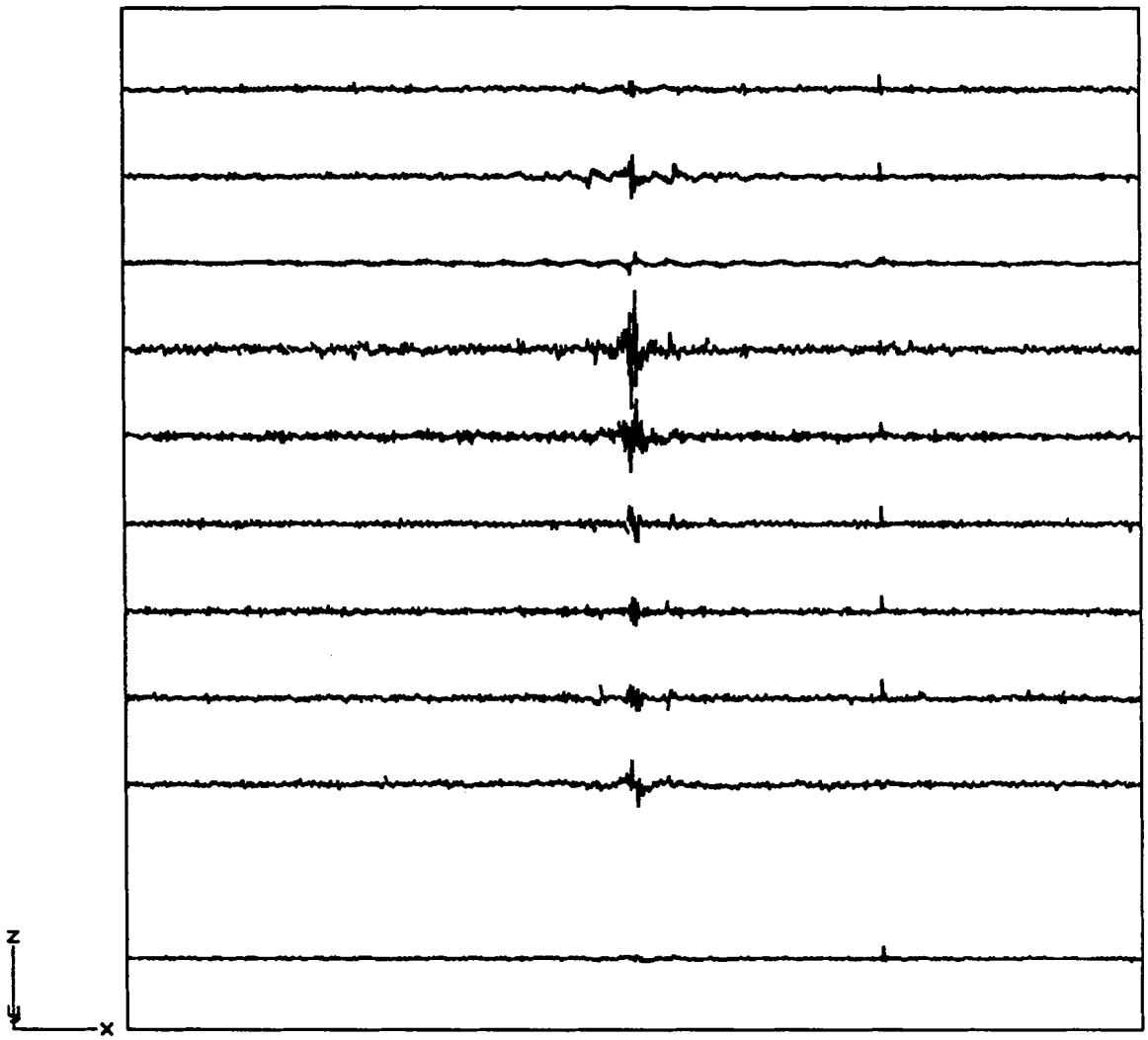

Fig. 2.7 Complex cepstrum for $n = -512, \ldots 511$, calculated with exponential weighting

By removing these cepstral peaks and replacing them with zero the cepstral filtering procedures were subtracting insufficient and variable values from these cepstra to perform effective filtering. In the next chapter we examine the causes of this failure

and propose alternate filtering structures which can circumvent such problems.

QUEFRENCY	RECTANGULAR WINDOW		EXPONENTIAL WEIGHTING	
	THEORY	AVERAGE	THEORY	AVERAGE
250	0.500	0.210	0.389	0.250
500	-0.125	-0.045	-0.076	-0.040

Table 2.1 Cepstral peaks from theory and experimental averages

Chapter 3

Development of Dereverberation Technique

We saw in Chapter 2 that the cepstral filtering techniques proposed and developed in [27] failed to perform well for the dereverberation of continuous, echoed speech. The reverberation impulse response used was a simple minimum phase echo of large amplitude for which the peak-elimination procedure should have been expected to work well. However, because the complex cepstrum of the echo was not well calculated, the elimination of the detected peaks did not correspond to perfect inverse filtering. Furthermore, only the first peak of the echo was generally detectable. The low-passing technique with a cut-off quefrency of 200 samples (31 ms) was demonstrated to cause unacceptable distortion . It is unrealistic to raise this cut-off quefrency substantially because the impulse responses encountered in rooms can be expected to have echo components at such values. If the impulse response is mixed phase, then impulse response cepstral components at all values of quefrency can be expected. Therefore, we must find new ways to exploit the cepstral method.

In this chapter, we first investigate the causes for the inaccurate computation of the impulse response cepstrum. Based on these findings, we then suggest procedures to allow the estimation of the impulse response complex cepstrum more accurately.

Finally, we introduce a new filtering structure which exploits these estimation procedures.

3.1 Investigation of Window Effects

Of the filtering procedures investigated in Chapter 2, the best results were obtained when the cepstrum was calculated from segments to which exponential weighting was applied. This raises the question as to which window function would in general allow us to calculate the cepstrum most accurately.

3.1.1 Commonly Used Windows for Cepstral Analysis

There are several types of window functions commonly used in cepstral work. In applications for speech analysis in which the deconvolution of the pitch train from the vocal tract response is desired, the Hamming window has typically been used [31, 30,36]. For echo removal, the exponential window has generally been used, both in speech [27,36], and in seismic signal processing in which an impulse response of the earth is to be deconvolved from a seismic "wavelet" [38,39]. The exponential window use is often motivated by the attempt to convert echo series to minimum phase, as in the seismic case where in general the entire seismic "wavelet" and its reflections are available for processing. The use of the exponential window both for reduction of cepstral aliasing and for reduction of truncation effects is mentioned in [36].

In terms of the time windowing properties which would normally be associated with the processing of continuous signals in finite length blocks, little is written in the cepstral literature. However, it has been noted [40] that for short-time cepstral

analysis, the success of the cepstral calculation is highly dependant upon the shape and the time registration of the window function.

3.1.2 Motivations for Window Use for Continuous Signals

Let us review the reasons, apart from those directly concerning the movement of z-plane zeroes, for which we may wish to employ time domain window functions. One of the goals of a window function, in general, is to minimize edge effects without destroying the information to be extracted. By edge effects, we mean the effect upon the frequency or cepstral domain representation caused by abruptly truncating a continuous signal.

Windowing is most often applied in digital signal processing with a view towards the frequency domain properties of the windowed signal. Let us recall that multiplication in the time domain corresponds to convolution in the frequency domain. For example, the multiplication of a continuous signal by a rectangular window corresponds to a convolution in frequency by a sinc function which induces ripple in the spectrum. Multiplication by a smoothly tapering window such as a Hamming window, on the other hand, corresponds to convolution with a more smoothly varying function in frequency, which has the effect of reducing spectral ripple at the cost of spectral "blurring". It is shown in [41] that the effect of short time Hamming windowing on speech is to non-linearly interpolate the log spectrum of the vocal tract impulse response between pitch harmonics; this leads to cepstral aliasing of the vocal tract cepstrum.

Similar results which describe the frequency domain effects in the context of dereverberation of impulse responses from speech are not available, seemingly because

of the mathematical complexity introduced by the complex logarithm operation. As an example, consider the finite length signal $x(n) = s(n) * h(n)$. The spectrum is

$$X(\omega) = S(\omega)H(\omega) \tag{3.1}$$

The complex logarithm operation produces

$$\widehat{X}(\omega) = \widehat{S}(\omega) + \widehat{H}(\omega) \tag{3.2}$$

Now, however, consider the windowed signal $y(n) = x(n)w(n)$ for which the spectrum is

$$\begin{aligned} Y(\omega) &= [S(\omega)H(\omega)] * W(\omega) \\ &= \int_{-\pi}^{\pi} [S(\lambda)H(\lambda)]W(\omega - \lambda)d\lambda \end{aligned} \tag{3.3}$$

For the Hamming window $w(n) = 0.5[0.54 - 0.46\cos(\frac{2\pi n_0}{N-1})]$,

$$Y(\omega) = 0.5\left[.54X(\omega) - .23X(\omega - \frac{2\pi}{N-1}) - .23X(\omega + \frac{2\pi}{N-1})\right] \tag{3.4}$$

It is extremely difficult to find an expression for the complex logarithm of equation (3.4) which will allow us to predict the effect upon the cepstrum.

In view of the difficulty of analysis of the window effects in the frequency domain, let us reconsider the windowing problem in the time domain. From Section 2.5.1, each segment of reverberant speech may be written as

$$x_i(n) = s_i(n) * h(n) + e_i(n) \tag{3.5}$$

We may describe the error term as the sum

$$e_i(n) = v_i(n) - w_i(n) \tag{3.6}$$

where $v_i(n)$ is the "extra" echo which intrudes from the previous segment and $w_i(n)$ is the "missing" tail of the echo of the speech of the current segment. Intuitively,

the goal of windowing would be to reduce the importance of these error components by smoothly tapering the segment boundary, while at the same time not introducing distortion into the calculated cepstrum. Rectangular windows obviously provide no tapering at segment boundaries. Functions such as Hamming windows are tailor-made for reduction of truncation error, but their effects upon the convolutional combination of signals (of extent on the same order as the window) are not known. Exponential windows provide smooth taper at the segment finish only, but because they do not destroy the convolutional combination between signals, they affect the cepstrum in a known way.

3.1.3 Experimental Investigation of Window Effects on Cepstrum

In this section we compare the effects of the rectangular, Hamming, and exponential windows upon the computation of the complex cepstrum from continuous speech. We wish to determine which if any of these windows will allow us to accurately calculate the cepstrum so that the filtering process can proceed successfully.

In order that the experiments not reflect any peculiarities of the complex cepstrum of speech, experiments were carried out using white noise as the "speech" signal. Simple minimum and maximum phase echoes were added to this white noise, and the complex cepstrum was calculated for all three types of window functions, and for segment sizes of various lengths. The reverberation functions used were

$$h_{min}(n) = h(n) + 0.5h(n-200)$$

and

$$h_{max}(n) = h(n) + 2.0h(n-200)$$

Note that $h_{min}(n)$ and $h_{max}(n)$ are minimum and maximum phase sequences having the same spectral magnitude.

In Chapter 2, it was noted that for the rectangular and exponential windows the main causal cepstral peak obtained for the echo was less in size than predicted by theory, and in the rectangular window case, "spurious" cepstral peaks were found at the mirror-image anticausal location. For the present experiments, the computed cepstral values for two quefrencies $n = \pm 200$ are are tabulated in Table 3.1.

WINDOW TYPE	MIN/MAX	C CEPS AT	NX 512	NX 1024	NX 2048	NX 4096
RECTANGULAR	MIN	−200	0.123	0.183	0.156	0.211
		+200	0.120	0.174	0.256	0.240
	MAX	−200	0.126	0.191	0.230	0.091
		+200	0.121	0.167	0.182	0.359
HAMMING	MIN	−200	0.078	0.174	0.185	0.231
		+200	0.077	0.175	0.263	0.267
	MAX	−200	0.075	0.188	0.218	0.244
		+200	0.081	0.175	0.226	0.251
EXPONENTIAL $\alpha = 0.996$	MIN	−200	0.006	−0.002	0.008	0.002
		+200	0.152	0.185	0.185	0.173
	MAX	−200	0.021	0.002	0.007	−0.002
		+200	0.147	0.178	0.175	0.177

Table 3.1 Cepstral peaks calculated for different windows
The "speech" is white noise and the impulse responses are
$h_{min}(n) = \delta(n) + 0.5\delta(n-200)$ and
$h_{max}(n) = \delta(n) + 2.0\delta(n-200)$
Each reported value is an average. The number of segments averaged is equal to 20,10,5, and 2 for NX lengths 512, 1024, 2048, and 4096, respectively. For rectangular windows with NX 4096 the cepstral noise was extremely large.

The results are consistent with the findings of Chapter 2, indicating that they are not peculiar to speech. For the rectangular window, all values of window length show that peaks appear in both causal and anticausal quefrency locations where only one peak is expected. This is true both for the minimum phase echo, which from

theory has a peak of value 0.5 at $n = 200$, and for the maximum phase echo, which from theory has a peak of value 0.5 at $n = -200$. Furthermore, in no case was the peak value at the expected location close to its theoretical value. The peak values did show some increase with window length, but it was also found that the cepstral noise encountered increased as it appeared that phase unwrapping became increasingly error prone at the longer window lengths.

For the Hamming window, similar results to the rectangular window results were obtained. In this case, the causal and anticausal quefrency peaks were also "mixed up", and the peak values were too low. Thus it appears that in spite of the tapering effect of the Hamming window, it is not suitable for the calculation of the impulse response cepstrum.

The exponential window results are somewhat different. For the minimum phase echo, application of the exponential weighting factor of 0.996^n should produce a cepstral pulse of amplitude 0.224 at $n = 200$. The results show that a pulse of amplitude somewhat less than to this value was achieved for most window lengths. For the maximum phase echo, application of the exponential window converts the echo to minimum phase, and should produce a cepstral pulse of amplitude $0.996^{200} \times 2.00 = 0.897$ at $n = 200$. In this case, a peak is detected at $n = 200$, but it is much too low in amplitude. However, for neither echo is there a "spurious" peak at $n = -200$.

3.1.3.1 Interpretation of Results

In the light of the above results, it appears that both rectangular and Hamming windows do not allow the accurate computation of the complex cepstrum of signals of indefinite duration. It is interesting to note that the sum of the causal and the

anticausal peaks for the rectangular and Hamming windows seems to approach the expected value of the peak. This may indicate that the *even* part of the cepstrum $\widehat{h}(n)$ is being computed correctly, and hence that the spectral *magnitude* is computed correctly, while the *phase* is computed incorrectly. This interpretation would appear to support the speculation of Tribolet et. al. in [40], that

> "... windowing may smooth the logmagnitude and phase curves near the high Q poles and zeros. Although this effect is localized for the logmagnitude characteristic, the phase curve is globally affected and it can drastically change shape."

To this we add the speculation that the effect of the segmentation errors $v_i(n) - w_i(n)$ in the rectangular window case may similarly affect the unwrapped phase more severely than the spectral magnitude.

The exponential window case is again different. It seems that the maximum phase echo, which is converted to minimum phase by exponential windowing, produces a cepstral peak of the same amplitude as the minimum phase echo after exponential windowing. This surprising effect is thus similar to that described above for the rectangular and Hamming windows. The amplitude of both minimum and maximum phase echo peaks is less than would be expected from theory for the minimum phase case. Hence we find that the exponential window is also unsuitable for the direct calculation of cepstra from continuous signals.

In the next section, we further investigate the computation of the complex cepstrum, but concentrate directly upon the effects of segmentation error for each window.

3.2 Investigation of Segmentation Error Effects

The findings in the previous section showed that neither the rectangular, Hamming, or exponential windows were directly suitable for the computation of cepstra from continuous signals. In this section we conduct experiments to determine to what extent these problems are caused by segmentation error and to what extent they are due to inherent limitations of the window functions. Because the Hamming window provides smooth taper at both segment ends, we speculate that the poor computation seen for this window is not due to segmentation error but is an inherent property of this window. The exponential window provides taper at the segment finish; we therefore speculate, in view of the convolution preserving property of this window, that the poor computation seen for it is due to the segmentation error $v_i(n)$ at the segment start. Finally, the poor computation for the rectangular window seems clearly due to segmentation error at both segment start and finish.

To investigate these assumptions, a short segment of white noise of length 1024 samples was used as a "speech" signal, $s(n)$. To this speech signal, a minimum phase and a maximum phase echo were added. The reverberation impulse response is the convolution of the minimum phase and maximum phase impulse response

$$\begin{aligned} h(n) &= h_{min}(n) * h_{max}(n) \\ &= [\delta(n) + 0.5\delta(n-50)] * [\delta(n) + 2.0\delta(n-100)] \\ &= \delta(n) + 0.5\delta(n-50) + 2.0\delta(n-100) + \delta(n-150) \end{aligned} \tag{3.7}$$

The echoed signal $x(n) = s(n) * h(n)$ is of length 1174 points and contains no segmentation error. From $x(n)$, three additional sequences were constructed: $x_v(n)$, which is the sum of $x(n)$ and an error sequence $v(n)$ added at the start of $x(n)$ in order to simulate the intrusion of an echo from the previous segment; $x_w(n)$, which is

the first 1024 points of $x(n)$ and which simulates the segment-end error which occurs because of segment truncation; and $x_{v-w}(n)$, which combines the two errors. In this experiment we compared the cepstra calculated for $x(n)$, $x_v(n)$, $x_w(n)$, and $x_{v-w}(n)$ using each of the three window types.

The complex cepstrum of $h(n)$ is the sum of a causal sequence of pulses spaced at 50 quefrency samples and an anticausal sequence of pulses spaced at 100 quefrency samples. For each of the results, plots are given showing the cepstrum as calculated above with the cepstrum of the original sequence $s(n)$ subtracted. Thus we can ignore the additive cepstrum $\hat{s}(n)$ as it is irrelevant to the desired information.

The results calculated using a rectangular window are shown in Figure 3.1. They show that the cepstrum calculated using the sequence with no segmentation error, $x(n)$, is exactly of the correct form. However, the cepstra calculated from sequences with segmentation at either and both ends are extremely distorted.

The Hamming window results are shown in Figure 3.2. In this case, all cepstra are distorted. Even the cepstrum calculated from the sequence with no segmentation error appears noisy and does not resemble $\hat{h}(n)$.

For the exponential window results, we show the cepstrum calculated using three degrees of exponential weighting. It should be noted that the exponential weighting changes the peak amplitudes, but since the exponential weighting factor γ is known, no information is lost. The first results, for which $\gamma = 0.998$, are shown in Figure 3.3. Without segmentation error the cepstrum is calculated perfectly. With segmentation error $v(n)$ at the segment start, the cepstrum is noisy and distorted. With segment truncation error $w(n)$, the cepstrum is much less distorted, but still remains somewhat noisy.

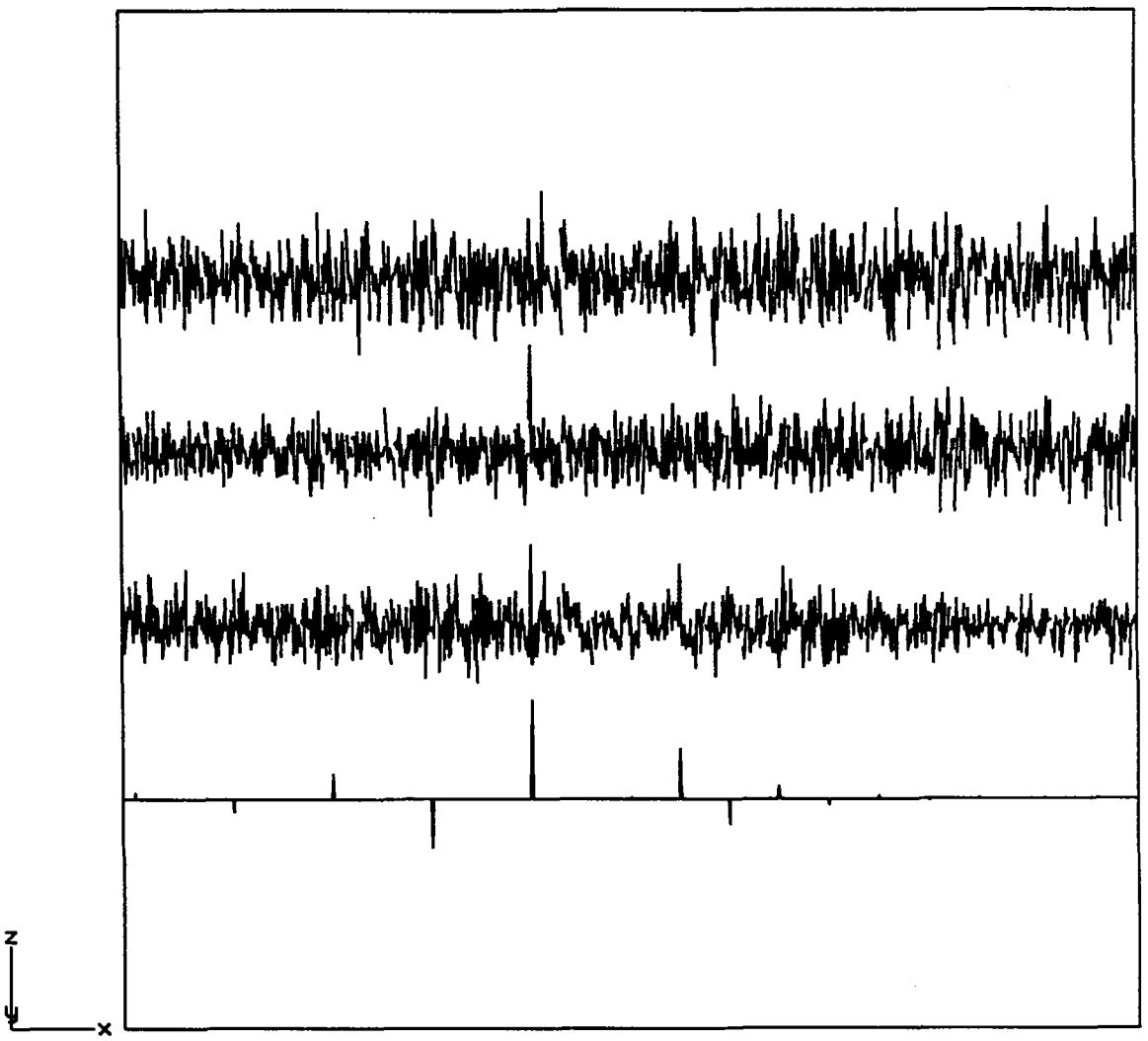

Fig. 3.1 Cepstrum calculated with rectangular window
From top:
$\widehat{x}_{v-w}(n) - \widehat{s}(n)$
$\widehat{x}_{w}(n) - \widehat{s}(n)$
$\widehat{x}_{v}(n) - \widehat{s}(n)$
$\widehat{x}(n) - \widehat{s}(n)$

For $\gamma = 0.995$, shown in Figure 3.4, the results are striking. Again, with segment-start error, the result is distorted, but with segment truncation error, the cepstrum is calculated virtually perfectly.

Finally, for $\gamma = 0.992$, the segment-start error also produces a distorted and noisy

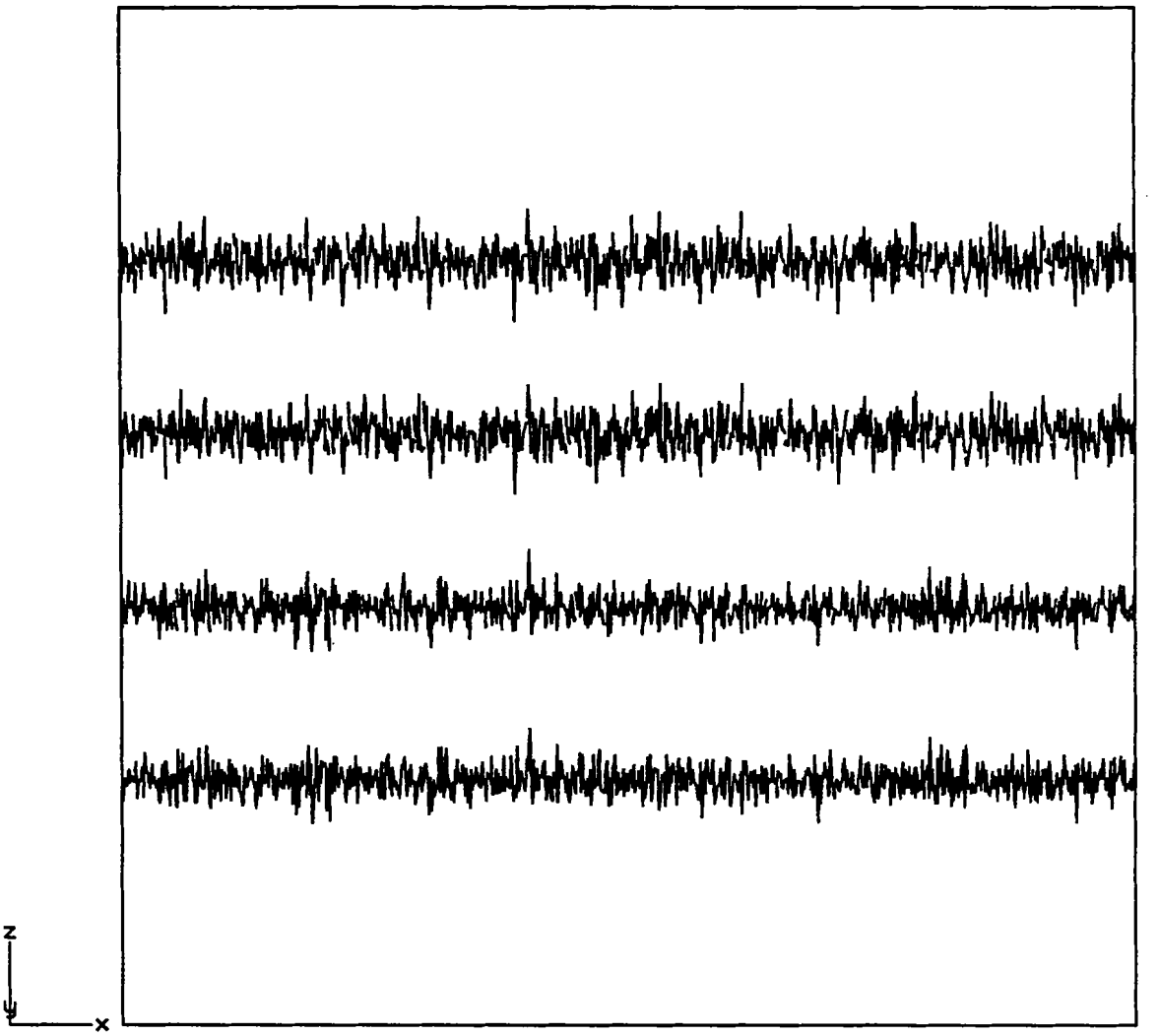

Fig. 3.2 Cepstrum calculated with Hamming window
From top:
$\hat{x}_{v-w}(n) - \hat{s}(n)$
$\hat{x}_w(n) - \hat{s}(n)$
$\hat{x}_v(n) - \hat{s}(n)$
$\hat{x}(n) - \hat{s}(n)$

cepstrum while the segment finish error has no effect whatever on the calculation. For this value of γ, the maximum phase echo has been converted to minimum phase and the echo cepstrum is entirely causal.

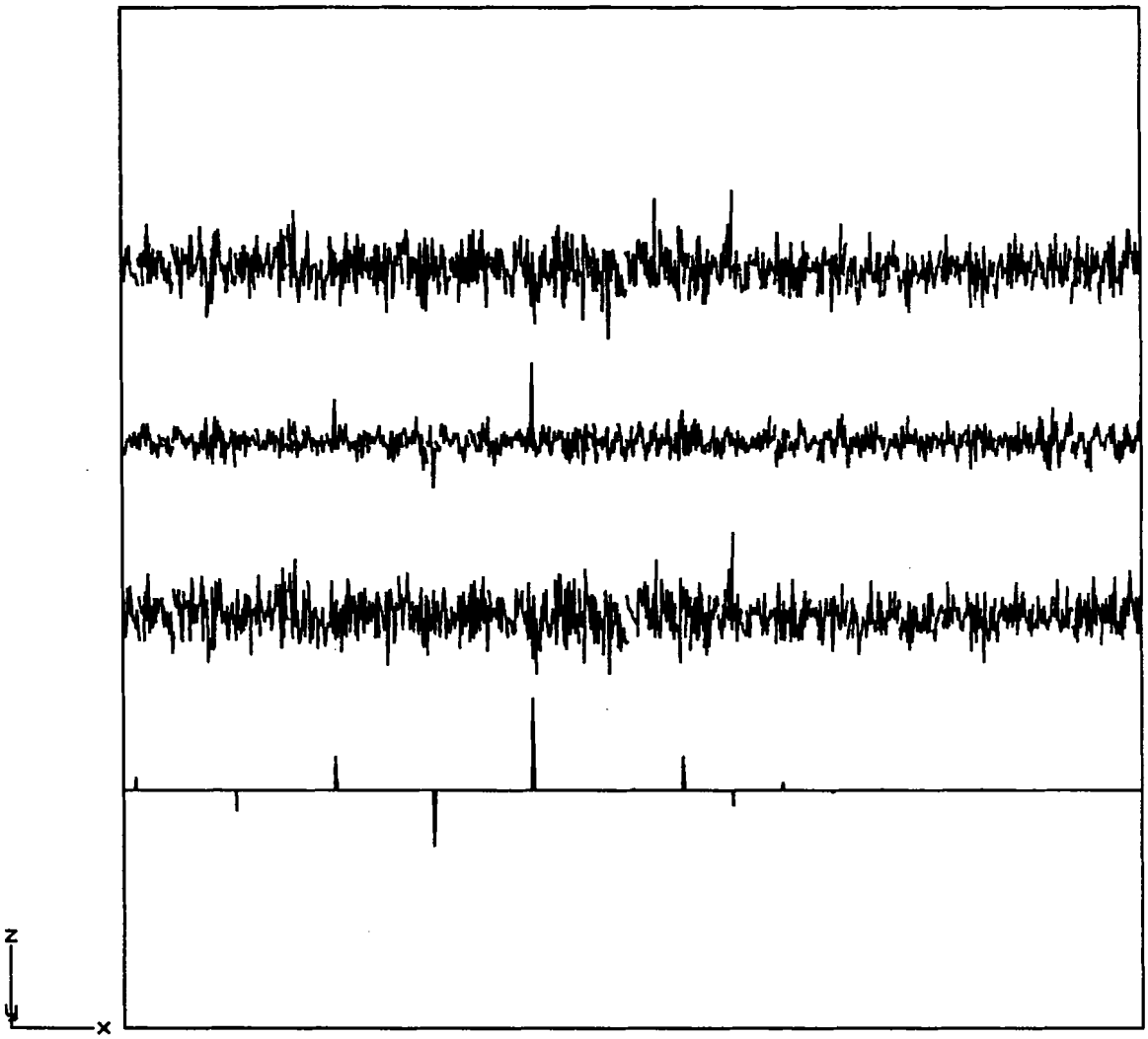

Fig. 3.3 Cepstrum calculated with exponential window,
$\gamma = 0.998$,
From top:
$\widehat{x}_{v-w}(n) - \widehat{s}(n)$
$\widehat{x}_{w}(n) - \widehat{s}(n)$
$\widehat{x}_{v}(n) - \widehat{s}(n)$
$\widehat{x}(n) - \widehat{s}(n)$

3.2.1 Interpretation of Results

The results of these experiments confirm the assumptions that the distortions noted in the calculations of cepstra from finite length segments of continuous signals

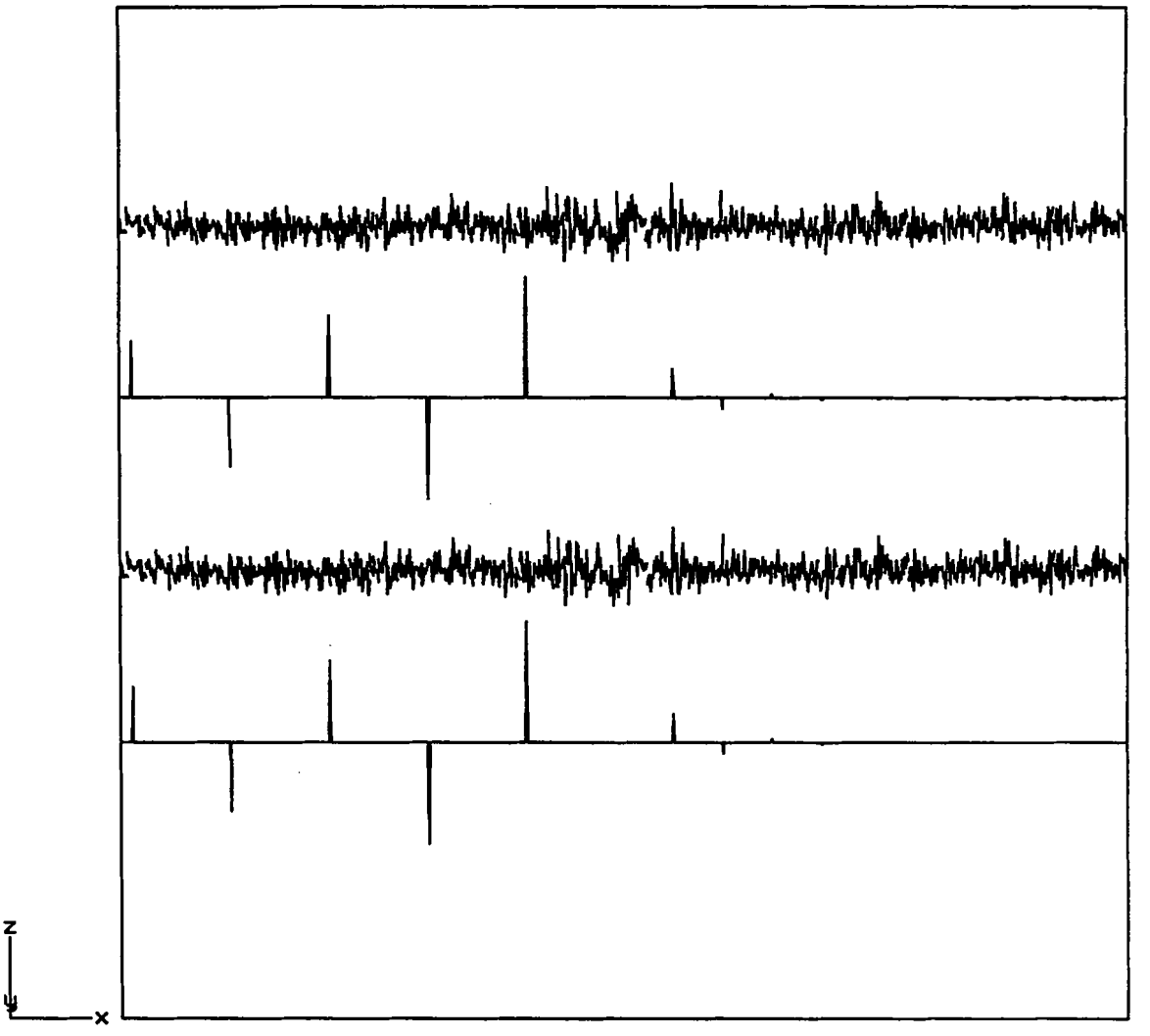

Fig. 3.4 Cepstrum calculated with exponential window,
$\gamma = 0.995$,
From top:
$\hat{x}_{v-w}(n) - \hat{s}(n)$
$\hat{x}_w(n) - \hat{s}(n)$
$\hat{x}_v(n) - \hat{s}(n)$
$\hat{x}(n) - \hat{s}(n)$

may be caused both by segmentation error and by window function. The Hamming window was shown to cause serious distortions even without segmentation error. The rectangular window, without segmentation error, caused no distortions but with segmentation error at either end resulted in serious distortions. For the exponential

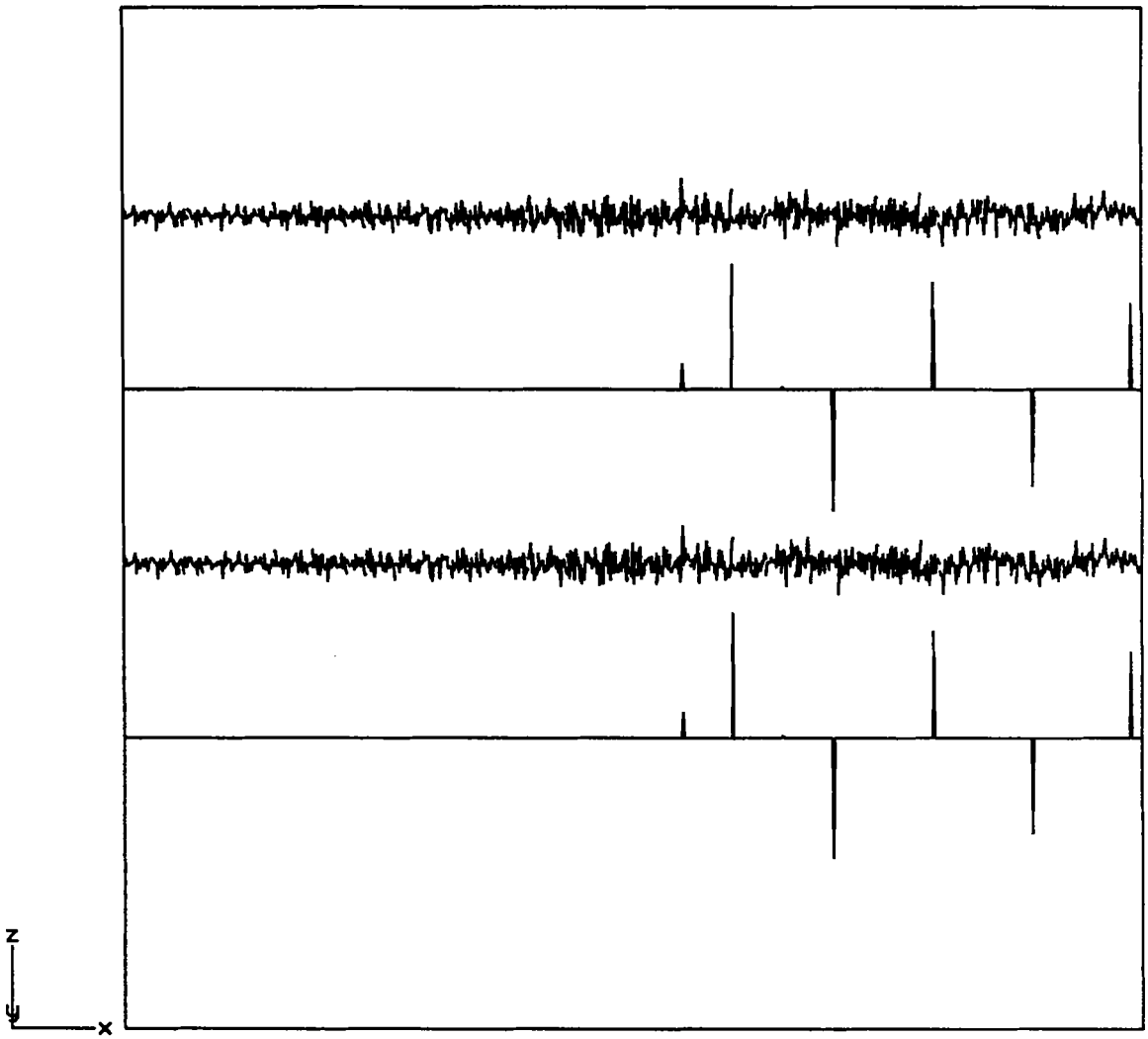

Fig. 3.5 Cepstrum calculated with exponential window,
$\gamma = 0.992$,
From top:
$\hat{x}_{v-w}(n) - \hat{s}(n)$
$\hat{x}_w(n) - \hat{s}(n)$
$\hat{x}_v(n) - \hat{s}(n)$
$\hat{x}(n) - \hat{s}(n)$

window, we found that segment-start error caused serious distortions while segment-end error caused little or no cepstral distortions for certain values of γ. Specifically, heavy exponential weighting (small values of γ) was most effective in the reduction of distortion due to segment-end error.

While it is stated in [27] that heavy exponential weighting may lead to roundoff errors, in the present experiments computation was performed using floating point arithmetic and this problem did not occur. With floating point arithmetic, heavier weighting factors are most successful in eliminating segment-end error because, we speculate, the value of the window at the segment end is lower and the effect of truncating the segment becomes negligible. This suggests that values of γ not effective using a window length of 1024 points, such as $\gamma = 0.998$, may be effective if longer windows are used. For example, with $\gamma = 0.998$, increasing the window length to 4096 points brings the final value of the window to 0.000275 from 0.129. It was confirmed by experiment that with this window length, $\gamma = 0.998$ effectively eliminated cepstral error for the segment-end error case.

3.3 Averaging in the Complex Cepstrum

In many scenarios in which we may wish to dereverberate speech, it is reasonable to assume that the impulse response of reverberation may change relatively slowly. For example, if reverberation occurs as a result of a fixed speaker playing speech into a room or other enclosure, the impulse response characterising the reverberation may be constant. Or, a person seated while talking in room during, for example, a teleconference, may move his head only slowly. A similar assumption was successfully made in the restoration of acoustic recordings [31]. In order to exploit such an assumption, we may wish to include averaging in the processing technique used to separate $s(n)$ and $h(n)$.

In the Figures 2.6 and 2.7, the calculated cepstra of a number of segments of echoed speech are displayed together with their averages. The noise like variations in

the cepstrum appear to be lower in the averages than in the individual cepstra, and the result is that the peaks due to the echoes are more clearly defined. In particular, we notice that prominent peaks due to voicing which are visible in most individual segments are not visible in the averaged plots. This effect is to be expected where the speech pitch is not perfectly constant.

3.4 Proposed Dereverberation Algorithm

We have demonstrated that the dereverberation of continuous speech by direct application of the cepstral filtering ideas in [27] is not feasible because the cepstrum can not be accurately computed due to segmentation errors. A further problem with this method is that for every output segment of speech, the cepstrum must be computed through phase unwrapping. Since the adaptive integration algorithm of [32] occasionally fails to produce an estimated phase curve, a practical implementation would be difficult.

However, we have seen that with the use of the exponential window, cepstra can be accurately computed when no segment-start error is present. Furthermore, the exponential window allows the conversion of mixed phase impulse responses to minimum phase and hence the calculation of the cepstrum of the impulse response without phase unwrapping. Finally, we have seen that averaging in the complex cepstrum can exploit the variability of the speech cepstrum to emphasize the impulse response cepstrum. Therefore, we propose the dereverberation system shown in Figure 3.6.

In the proposed system, we seek to use the cepstrum to *identify* the reverberation impulse response, and do not propose to filter in the cepstrum. The removal of the impulse response would be be performed using a linear inverse filtering step separate

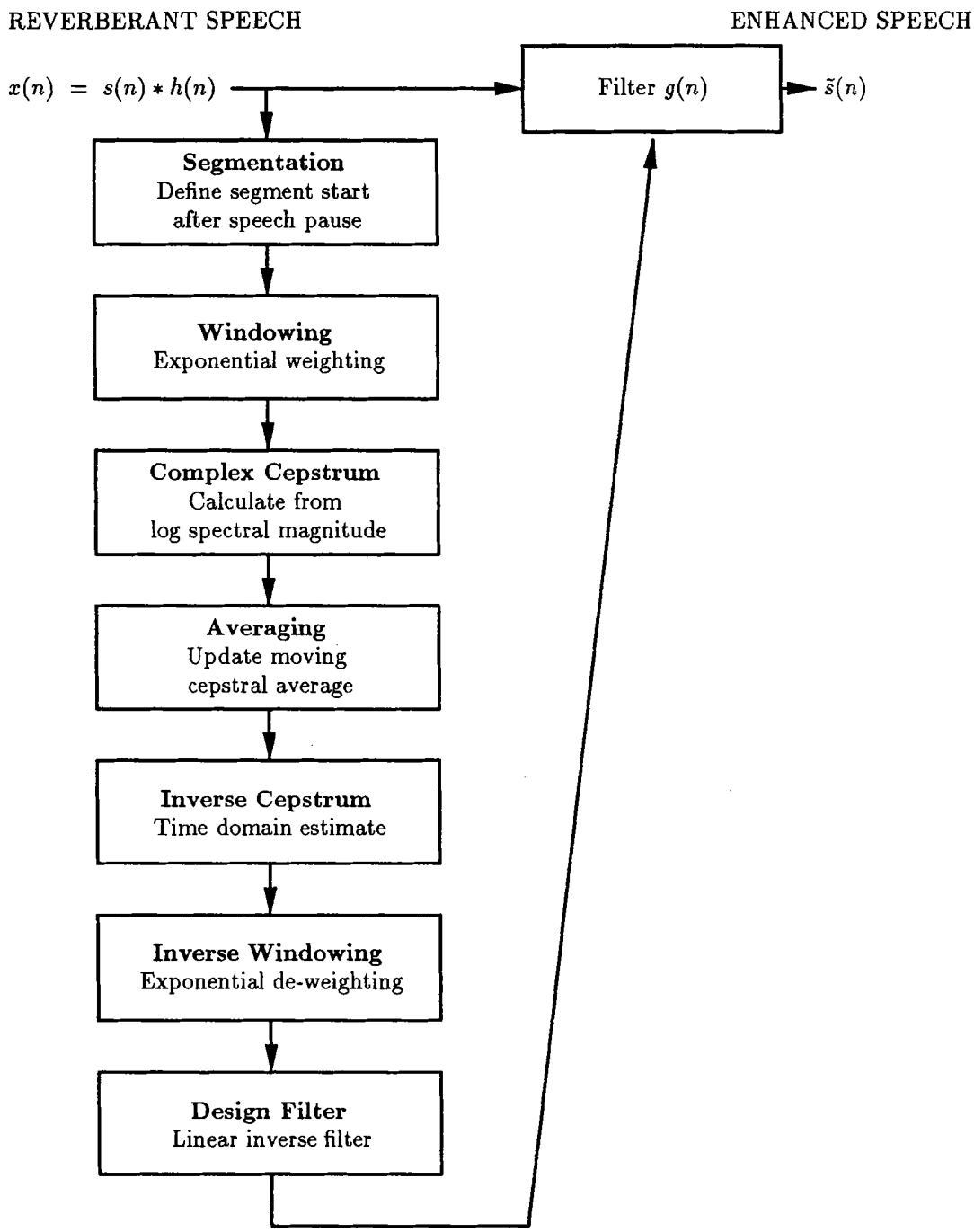

Fig. 3.6 Proposed dereverberation system

from the cepstral identification step. For the calculation of the cepstrum, we employ

an exponentially weighted window function. The key to the success of the cepstral identification would be to align the windowed segment in time so as to reduce the segment-start error. Thus a cepstral analysis segment would be defined to begin only after a period of relative speech silence. In this way, the intrusion error which normally occurs at the start of a segment which is caused by the intrusion of the echo from the previous segment would be reduced in amplitude.

Because the filtering and cepstral identification steps are decoupled in the proposed system, it is potentially suitable for real-time operation. We can view the cepstral identification procedures as operating in the background, continually updating the estimated impulse response. The filter would be updated from the latest available estimate. The only delay involved between reverberant and enhanced speech would be delay related to the operation of the linear filter. This delay could be of two types: delay required for filter causality (for mixed phase impulse responses - see Chapter 4); and delay required for filter implementation. We discuss this delay in Chapter 4.

The quality of the impulse response estimate and of the filter would depend upon the number of speech segments which are used in the computation of each estimate. This factor would be limited by the the speed at which the reverberation impulse response changed. The many parameters in this system, which include exponential weighting factor, number of segments to be averaged, size of DFT operations, and linear filter type and size, will be the subject of investigation in the remainder of this thesis.

Many algorithms exist for the detection of speech silent periods (discussed in [1]). They employ measures such as short-time energy and zero crossing rates to distinguish

between speech and non-speech activity. For the purposes of evaluating the proposed system, we will examine the reverberant speech waveforms "by eye" and determine segment starts after apparent speech pauses.

The goal of this research, as explained in Chapter 1, was to produce a dereverberation system capable of reducing or eliminating reverberation in all delay ranges. We now comment on this goal with respect to the proposed system. First, the maximum echo delay which the system could handle would be restricted by the length of the DFT operations involved in the cepstral calculations and the sampling rate of the system. For sampling rates of 8000 samples per second, a DFT length of 8192 samples would in theory allow the identification of 4096 cepstral components which corresponds to a (minimum phase) impulse response of 4096 points, or one half second. This "ballpark" seems adequate in the context of small room dereverberation. Secondly, we must reasonably assume that the accuracy of the segmentation process would increase for speech with long pause lengths. Pause lengths of between 0.1 and 0.2 seconds are common in conversational speech [42], and the proposed system would be feasible at least for reverberation within these times. Finally, we must ask for how small an echo delay can this system be feasible. The speech cepstrum is concentrated within the first pitch period, and is characterised by a large peak at this period. Normally, the cepstrum can not identify echo components at this delay time. However, it is conjectured that by averaging the cepstrum over a number of segments, the pitch peak and the speech cepstrum in general will become "smeared" to the extent that large peaks due to echo components can be identified within this range. This view is supported by the findings of relatively large variability of pitch within sentences of durations on the order of one to two seconds by speech recognition researchers [43].

This system and the above conjectures will be tested experimentally in Chapter 5. In the next chapter, we discuss the techniques by which we may design linear inverse filters to remove the estimated reverberation impulse response.

Chapter 4 Linear Inverse Filter Design

A major component of the proposed dereverberation system described in Chapter 3 is the design of the linear inverse filter. This filter is to be designed from an estimate of the acoustical impulse response, $h(n)$. If we represent the impulse response of the filter by $g(n)$, the filtering step replaces the acoustical impulse response by a new "impulse response", $f(n)$, which is the convolution of $h(n)$ and $g(n)$. This is shown in Figure 4.1. If the filter design is successful, $f(n)$ will be close to a unit impulse at $n = 0$ or at some delay time, and the perceptual effects of reverberation will be diminished or eliminated.

This chapter is concerned with the design of the filter based on a finite length estimate of the acoustical impulse response. We must keep in mind several limitations. First, for stability, we restrict the filter choice to be among finite impulse response methods. Because perfect inverse filtering of a finite length impulse response (with z-plane zeroes only) would require an infinite impulse response filter (with poles), the restriction to FIR design implies some error. We can only attempt to minimize this error by appropriate selection of filter coefficients. Second, although in this chapter we do not consider it, any error in the acoustical impulse response estimate resulting from the cepstral identification process is bound to affect the filter performance.

Finally, the modelling of the acoustical impulse response itself as finite length is an approximation.

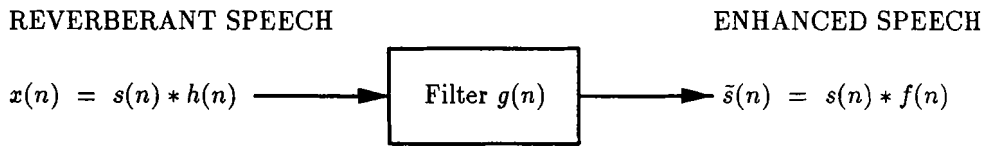

Fig. 4.1 Dereverberation using linear filter

We discuss first some qualitative measures of reverberation. Then, we review the results of the work of Neely and Allen in designing filters for minimum phase impulse responses, and the work of Morjopolous concerned with filters for mixed phase impulse responses. We then discuss the "mechanics" of these design methods, and present examples of inverse filter designs for simulated room impulse responses.

4.1 Measures of Success of Inverse Filtering

The goal of the filtering process is to remove the displeasing effect of the reverberation upon the speech. As explained in Chapter 1, reverberation at short delays may in some cases be beneficial to speech perception in a binaural environment; however, we are working with a single channel, and it will be assumed that it is desirable to eliminate all reverberation.

Certain mathematical measures may be used to estimate in some way the degree of reverberation and hence the success of the processing. The simplest measures the ratio of direct and reverberant energy. For a room impulse response $h(n)$ of duration

n_1 samples, we define the direct to reverberant ratio [44], in decibels, as

$$R = 10\log\left\{\frac{h^2(0)}{\sum_{n=1}^{n_1-1} h^2(n)}\right\} \text{ (dB)} \tag{4.1}$$

With this definition, we make the approximation that only the first impulse at $h(0)$ corresponds to the direct energy.

It is useful to compare the direct to reverberant ratio before and after filtering. If a "spiking" filter $g(n)$ of duration n_2 is applied the goal is that

$$f(n) = h(n) * g(n) \approx \delta(n) \, , n = 0, \ldots, n_1 + n_2 - 1 \tag{4.2}$$

The "processed" direct to reverberant energy is defined as

$$R_p = 10\log\left\{\frac{f^2(0)}{\sum_{n=1}^{n_1+n_2-1} f^2(n)}\right\} \text{ (dB)} \tag{4.3}$$

We can now define the amount of reverberation "removed" from the signal as

$$I_r = R_p - R \text{ (dB)} \tag{4.4}$$

When this quantity is large, the processed speech has less reverberant compared with direct energy than the original speech.

Although the measure I_r describes the reduction of reverberant energy it tells us little about the frequency domain characteristics. Since spectral colouration, or modification of the short time spectrum, is an important factor in the perceptual effect of reverberation it is desirable to quantify it in some way. One way of quantifying spectral colouration is to measure the degree of spectral "ripple" or deviation from the mean value, and several methods to do this have been proposed. It was found [16] that the average value of the second moment of the room spectral response correlates with the subjective perception of colouration. A second, similar measure

is the standard deviation of the room energy spectral log magnitude response. (The energy spectral magnitude is found from the squared spectral magnitude of the room impulse response.) A theoretical relationship between the room "critical distance" (the distance at which the direct and reverberant energies become equal) and the standard deviation of the normalized energy magnitude response was found in [45]. Although this relationship is true in theory only under certain assumptions (the independence of direct and reverberant sounds, and a uniform frequency response of the reflecting surfaces), good agreement was found from both room simulations and real room measurements. Because of the above findings we adopt this measure as the "spectral ripple" measure S. It is calculated from

$$S = \overline{[(H_n(k) - \overline{H_n(k)})^2]}^{0.5} \text{ (dB)} \tag{4.5}$$

where the normalized magnitude squared coefficients are defined as

$$H_n(k) = 10 \log \left\{ \frac{|H(k)|^2}{\overline{|H(k)|^2}} \right\} \text{ (dB)} \tag{4.6}$$

Again we may wish to compare the spectral ripple before and after processing. We define S_p from the net impulse response $f(n)$ as

$$S = \overline{[(F_n(k) - \overline{F_n(k)})^2]}^{0.5} \text{ (dB)} \tag{4.7}$$

and the improvement gained by processing, is

$$I_s = S - S_p \text{ (dB)} \tag{4.8}$$

When this quantity is large, the spectral ripple measure has been reduced, indicating a decrease in spectral colouration.

While these measures are useful insofar as they provide a quantitative basis for comparing the "badness" of reverberation before and after processing, they can not replace listening tests for this determination. The energy ratio measure R does not take into account the distribution in time or the frequency pattern of the reverberant energy, while the spectral ripple measure S can not, for example, distinguish between minimum and maximum phase impulse responses of the same spectral magnitude. Neither the I_r nor the I_s test indicates if in reducing the reverberation with the application of the filter, smaller amplitude, longer delay echo has been introduced, which may be perceptually worse than the original reverberation.

4.2 Review of Previous Work: Linear Filters for Dereverberation

The enhancement of reverberant speech using linear filters is a problem which is not covered extensively in the literature, probably because the impulse response which must be inverted is not usually known. Of the studies described here, one investigated the feasibility of enhancement with zero delay causal filters; the other used two sided filters with delay for causality.

4.2.1 Zero Delay Filters

The inverse of a minimum phase impulse response is causal and stable. Neely and Allen [14] investigated the feasibility of designing such inverse systems for minimum phase room impulse responses by simple Fourier inverse techniques. With the minimum phase impulse response $h(n)$, the inverse filter $g(n)$ is found by

$$g(n) \;=\; \mathrm{DFT}^{-1}\left\{\frac{1}{H(k)}\right\} \tag{4.9}$$

Although the true inverse sequence is infinite in length, $g(n)$ is effectively truncated by choice of DFT length. Using minimum phase simulated room responses generated by the image source simulation model [17], it was found in [14] that the reverberant speech, after being filtered with $g(n)$, sounded identical to the original speech.

Next, the application of such inverse filters was considered for the general mixed phase impulse response case. Mixed phase impulse responses may be considered as the convolution of a minimum phase component h_{min} having all z-plane zeroes inside the unit circle and and an all-pass (unity magnitude) component $h_{ap}(n)$ having zeroes both inside and outside the unit circle. The stable Fourier inverse system for $h_{min}(n)$ is causal but the stable inverse of $h_{ap}(n)$ is two-sided. The authors thus factored $h(n)$ into these two components with cepstral techniques and applied the Fourier inverse of $h_{min}(n)$ to the reverberant speech. Although the "room effect" had been removed, the resulting filtered speech exhibited a chime-like distortion. This distortion was due to narrowband peaks in the group delay of the remaining all pass component.

4.2.2 Two Sided Inverse Filters

The design of two sided inverse filters for mixed phase impulse responses is discussed in [46]. The authors considered two techniques: least squares inversion and homomorphic inversion. The least squares technique seeks the inverse filter $g(n)$ which produces a unit impulse at a specified delay time, with the least possible time domain squared error. The homomorphic inversion factors the impulse response into minimum and maximum phase components using the complex cepstrum, inverts each with either least squares or Fourier inversion, and convolves the component filters

with each other to form the filter $g(n)$. No advantage was found for the homomorphic technique to justify the increase in complexity.

Morjopolous [44] performed further tests with speech signals recorded in real rooms, by first estimating the impulse response from exciting signals, then by designing least squares filters from the resulting estimate. The amount of delay to specify in the least squares design was determined by an optimization procedure but it was found that the result was relatively insensitive to the amount of delay over a broad range. It was found that by using relatively short inverse filters (1024 points), most of the reverberant energy was removed from speech. Residual distortion remained, however, and was attributed to the inability of the short inverse filters used to remove fine spectral irregularities. In general, the improvement in direct to reverberant energy was greatest for those situations in which the source to microphone distance was large.

4.3 Design of Inverse Filters

The cepstral identification procedures described in Chapter 3 produce an estimate of the reverberation impulse response. We consider here the design of a corresponding inverse filter. The estimated impulse response is generally expected to be mixed phase, and we therefore require a two sided filter. In light of the results of [14] and [44] the ensuing delay between the reverberant and processed speech seems unavoidable for mixed phase impulse responses.

There are at least two avenues of design open for the filter: either Discrete Fourier Transform or least squares inversion of the impulse response estimate $h(n)$. The former requires less computation but the latter is guaranteed to be optimal in the

mean square error sense for the impulse response estimate. It should be noted that the mean square error criterion is probably not related to perceptual criteria.

The DFT inverse $g(n)$ for $h(n)$ is obtained from equation 4.9. We must note that by calculating $g(n)$ in this manner, we do not calculate the true Fourier inverse sequence of $h(n)$. The true inverse sequence is of infinite duration. In choosing a finite length DFT for our calculations, the true inverse sequence is aliased to produce $g(n)$. It is therefore necessary to choose a DFT length longer than the length of the desired inverse filter. In this way most of the aliasing distortion which occurs does so in regions outside the range of the computed inverse filter. Then, we can be confident that the distortion which occurs with this inverse method is due mainly to the effect of truncating the inverse filter rather than the effect of aliasing distortion.

If $h(n)$ is minimum phase, then its inverse sequence $g(n)$ is causal (again excepting for any aliasing distortion present). If $h(n)$ is mixed phase then its inverse sequence $g(n)$ has components in both causal and anticausal locations. In practical terms, for an N point DFT, the causal filter values $g(0),\ldots,g(\frac{N}{2}-1)$ are returned in DFT inverse locations $n = 0,\ldots,\frac{N}{2}-1$. The anticausal filter values $g(-1),\ldots,g(\frac{-N}{2})$ are returned in DFT inverse locations $n = N-1,\ldots,\frac{N}{2}$.

The least squares inverse $g(n)$ of length M for the impulse response $h(n)$ of length N is found [47] by solving an $M \times M$ system of equations

$$\mathbf{R}_{hh}\mathbf{g} = \mathbf{r}_{hd} \tag{4.10}$$

where $\mathbf{R}_{hh}$ is the $M \times M$ autocorrelation matrix for $h(n)$ and $\mathbf{r}_{hd}$ is the cross correlation between $h(n)$ and $d(n)$, the desired response. The desired response for inversion is a unit impulse at $n = 0$ for a zero delay system or at $n = l$ or a system with

delay l. Because $\mathbf{R}_{hh}$ is a Toeplitz matrix the system may be solved in a number of computations proportional to M^2 with, for example, the Levinson-Durbin algorithm.

4.3.1 Factors Governing Selection of Filter Parameters

In order to select the filter parameters such as filter length and delay general characteristics of the impulse response must be known. In general, a "long" impulse response will require a long filter. The computational burden both for calculating the filter coefficients and for filter execution increases as the filter length increases. For real time applications especially, it is desirable to use smaller filter lengths. The delay required depends upon the phase characteristics, but the results of [46] indicate that the choice of this parameter over a broad range is not critical. For example, with 1024 points of a mixed phase impulse response, delay values between about 50 and about 500 samples achieved roughly the same squared error performance. For real time applications, it is desirable to use as small a delay time as possible to minimize the time mismatch between reverberant and filtered speech. In this sense a delay within the above mentioned range (between 6.25 and 62.5 ms at a sampling rate of 8000 Hz) would probably be acceptable.

4.4 Inverse Filter Design: Experimental Results

In order to evaluate the suitability of the above design methods to room impulse response dereverberation, a series of experiments was performed using 3 simulated room impulse responses representing 3 different phase conditions: $h_1(n)$: minimum phase, $h_2(n)$: "close" to minimum phase (some z-plane zeroes just outside the unit

circle), and $h_3(n)$: mixed phase (with z-plane zeroes farther outside the unit circle). These impulse responses are shown in Figures 4.2, 4.3, and 4.4 and their characteristics are summarized in Table 4.1. In the table, γ_{min} refers to the exponential weighting value required to bring all zeroes inside the unit circle; hence it is a measure of the maximum zero distance from the unit circle. The number of such zeroes exterior to the unit circle is calculated from the linear phase factor of the impulse response (see Chapter 2). (If the linear phase factor indicates that there are no zeroes exterior to the unit circle, the impulse response is minimum phase.) Fourier and least squares inverse filters of various lengths and delays were designed and speech reverberated from the impulse responses was compared before and after filtering.

IMPULSE RESPONSE	PHASE	LENGTH	R (dB)	S (dB)	γ_{min}	# ZEROES OUTSIDE UNIT CIRCLE
h_1	MIN	2077	7.55	3.29	-	0
h_2	MIXED	989	5.35	3.90	0.9995	2
h_3	MIXED	1024	1.70	5.40	0.9963	32

Table 4.1 Characteristics of impulse responses

4.4.1 Fourier Inverse Filters

In this section we report the results of designing inverse filters with the DFT inverse method. In order to ensure that aliasing did not represent a serious source of error, very long DFT lengths (8,192 or 16,384) were chosen in order to calculate the inverse sequences. These DFT lengths were chosen by repeating the experiments a number of times with increasing DFT lengths, and reporting the results for the value that did not produce a perceptible difference in the filtered speech.

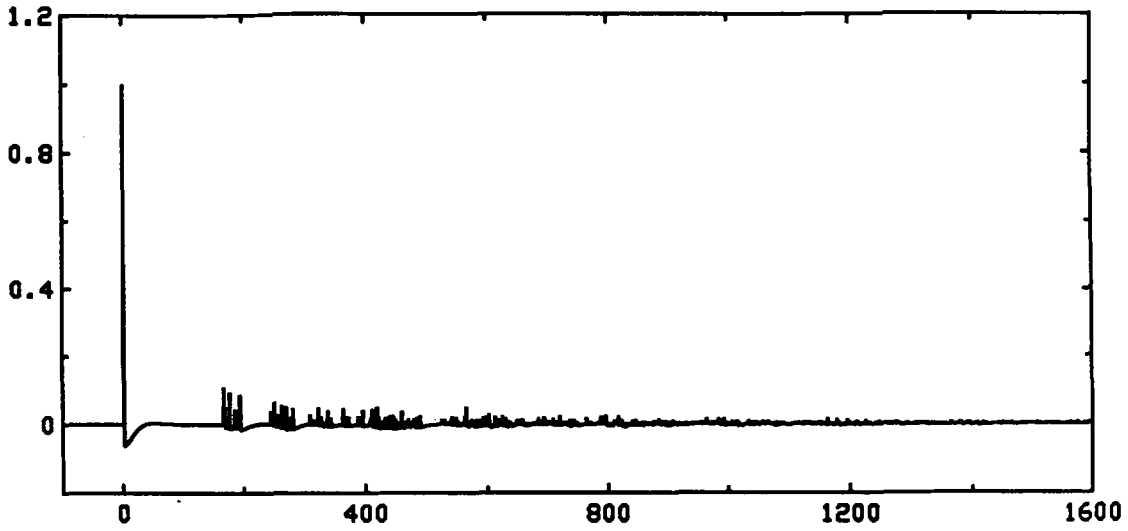

Fig. 4.2 Minimum phase impulse response $h_1(n)$

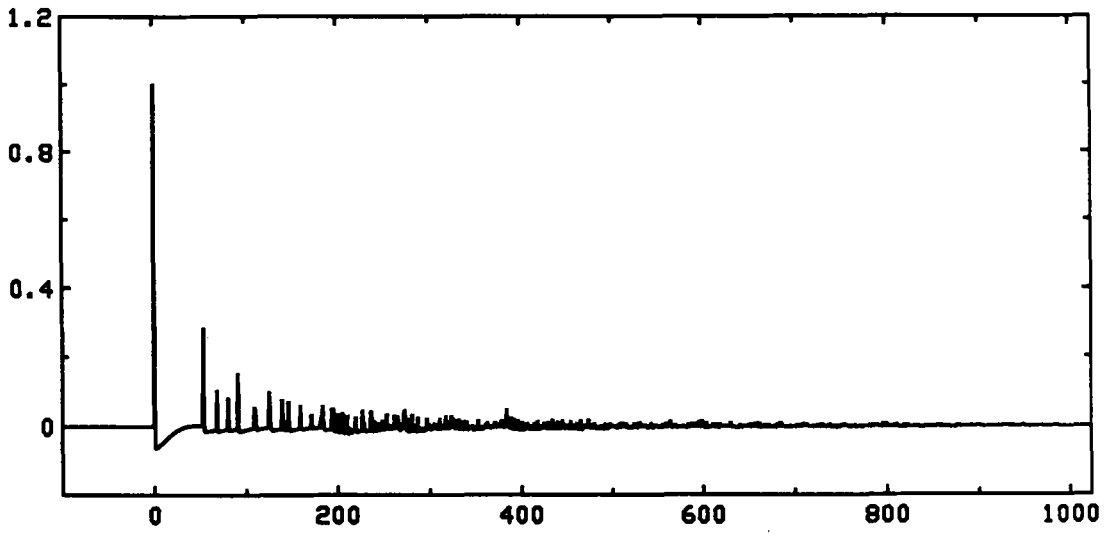

Fig. 4.3 Mixed phase impulse response $h_2(n)$

The Fourier inverse of $h_1(n)$ is shown in Figure 4.5. It can be seen that this inverse is entirely causal (with the exception of aliasing error), as would be expected for a minimum phase impulse response. The application of this filter to the corresponding reverberant speech produces a reduction in reverberant sound with a slight audible echo in the background. It was found that by increasing the length of the causal

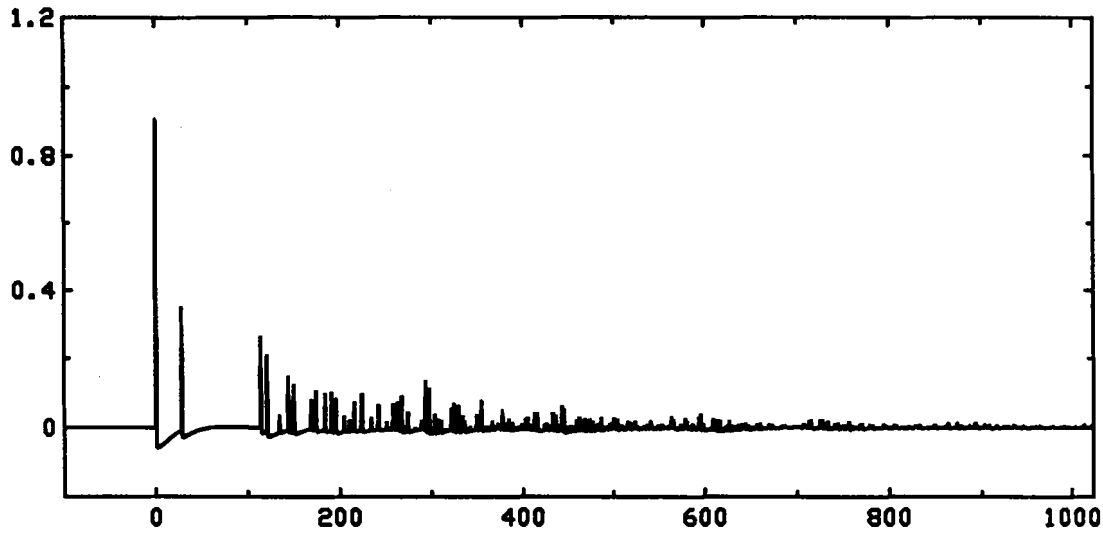

Fig. 4.4 Mixed phase impulse response $h_3(n)$

portion of the inverse filter, almost complete removal in reverberation was achieved, and the audible echo was decreased in amplitude.

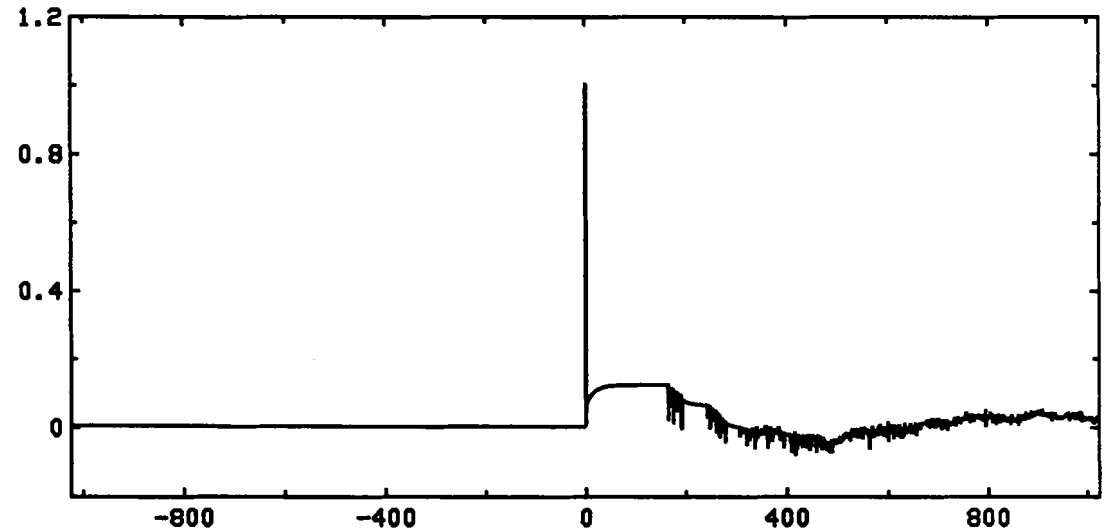

Fig. 4.5 Fourier inverse of $h_1(n)$

By contrast, Figure 4.6 shows 1024 anticausal and 1024 causal points of the Fourier inverse of $h_2(n)$. In order to obtain a realizable filter, the speech would be delayed by 1024 samples so that the filter term at $n = -1024$ would be applied at

$n = 0$. This filter would thus produce filtered speech at a delay of 1024 samples or $\frac{1}{8}$ seconds with respect to the reverberant speech. We may notice several interesting features from this inverse. First, most of the "energy" of the filter is in the causal portion, confirming that the impulse response has most of its zeroes inside the unit circle. Second, the causal portion of the filter decays faster than the anticausal portion. This may be a result of the proximity of the zeroes outside the unit circle to the unit circle, and the fact that the inverses of such zeroes decay very slowly in time.

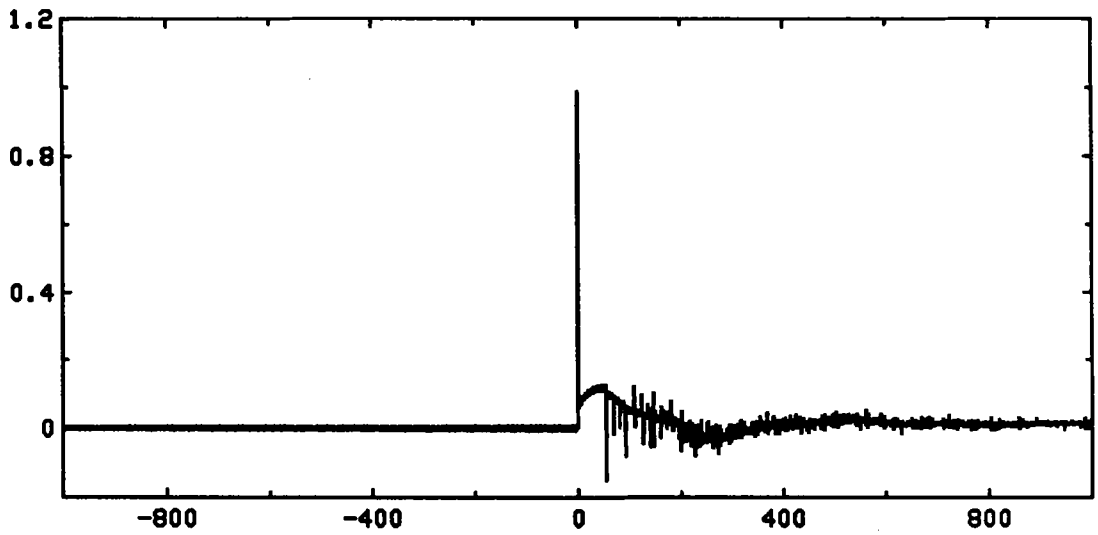

Fig. 4.6 Fourier inverse of $h_2(n)$

Figure 4.7 shows the convolution of the $h_2(n)$ and the Fourier inverse filter. The error measures before and after filtering showed improvements of $I_r = 15.02$ dB and $I_s = 3.30$ dB. The nonzero portions of Figure 4.7 aside from the impulse at $n = 0$ indicate error due to truncation of the inverse filter at 2048 points. Although most of the reverberant energy has been removed from the impulse response, filter truncation error has produced significant output error at regions below about

$n = -500$ and above $n = 1024$. Since the original impulse response $h_2(n)$ was truncated at 980 samples, the net impulse response after convolution with the filter has become extended in duration. Listening tests showed that the filtered speech had very little "room effect" remaining from the reverberation, but severely annoying distortion had been introduced from the filter truncation error. This error sounded as high pitched, distorted copies of the speech before and after the main speech.

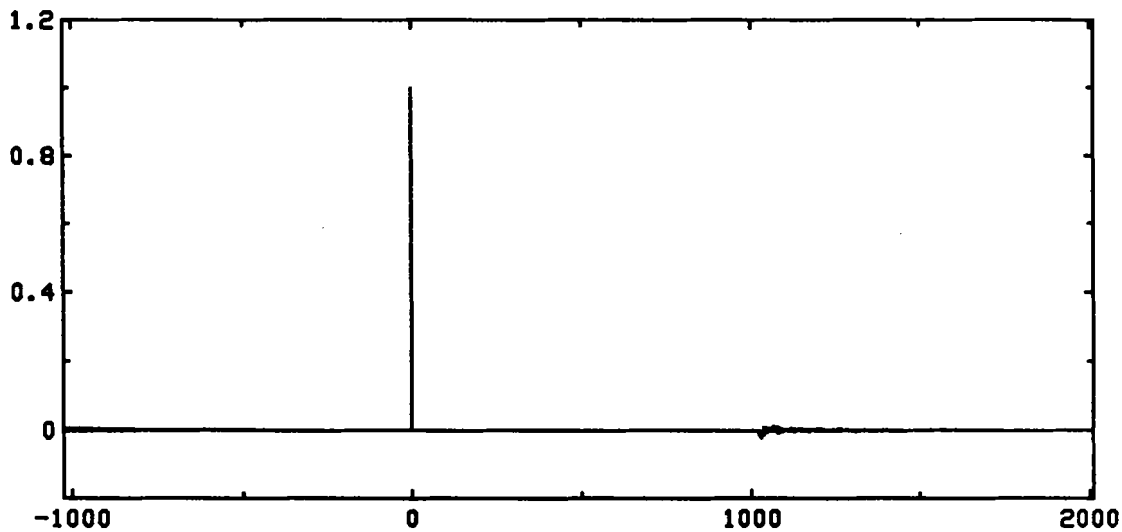

Fig. 4.7 Fourier inverse convolved with $h_2(n)$

Finally, the Fourier inverse of $h_3(n)$, shown in Figure 4.8 is substantially more "two sided" than the inverse of $h_2(n)$. The application of this filter to reverberant speech again produced a reduction in "room effect" but at a cost of annoying ringing distortion similar in form to that caused by the filter for $h_2(n)$.

4.4.2 Least Squares Inverse Filters

The least squares filter method was applied with various delays and filter lengths to the three impulse responses. Each filter was then convolved with the corresponding

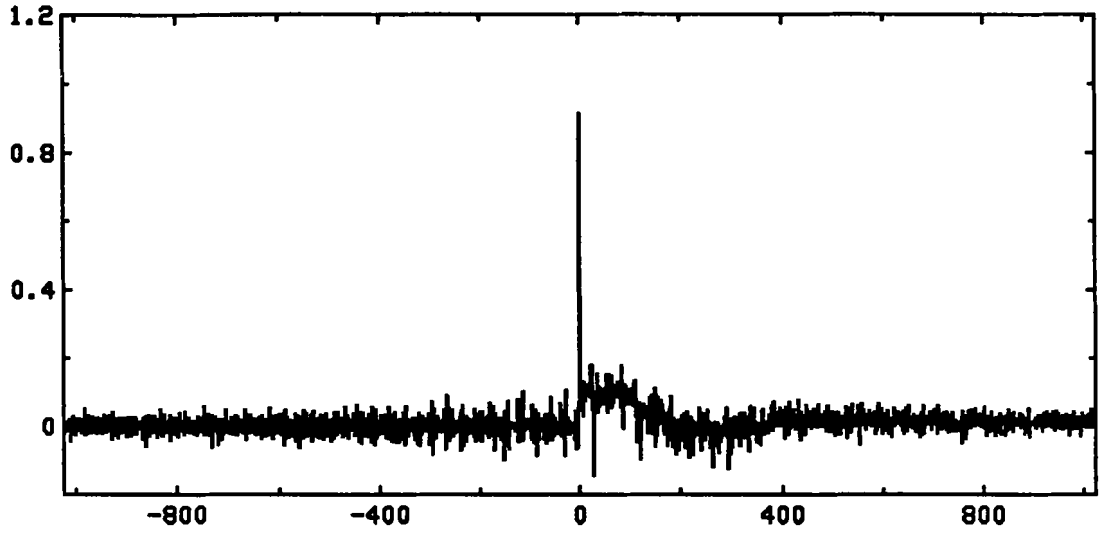

Fig. 4.8 Fourier inverse of $h_3(n)$

impulse response and the resulting measures of improvement in reverberation with respect to the original impulse response were calculated. The results for $h_1(n)$, $h_2(n)$, and $h_3(n)$ are summarized in Tables 4.2, 4.3, and 4.4. The rightmost two columns in these tables show a subjective description by the author of the effect of the filtering on the reverberant speech: the amount of reverberant sound or "room effect" compared with the un-filtered speech; and the nature of any distortions introduced by the filtering. These are further described below.

The impulse response $h_1(n)$ is minimum phase and has significant components extending to 250 ms. For this impulse response, least squares filters of length less than 2048 samples (about 250 ms) left a relatively large part of the room effect and induced an audible echo or copy of the speech into the background. For filters of 2048 samples and more, these effects diminished. For all filter lengths, the best sound and least distortion, and largest quantitative improvement as measured by I_r and I_s was achieved by using zero delay. Furthermore, in every instance both measures improved

FIL LENGTH	DELAY	I_r (dB)	I_s (dB)	REMAINING REV	DISTORTION
1024	0	12.30	2.61	significant	echo
1024	256	9.56	2.33	significant	echo
1024	512	6.80	1.94	significant	echo
1536	0	16.67	2.86	significant	echo
1536	256	14.68	2.75	significant	echo
1536	512	12.40	2.58	significant	echo
1536	768	9.64	2.28	significant	echo
2048	0	20.88	3.03	little	faint echo
2048	512	16.76	2.86	significant	echo
2048	1024	12.46	2.56	significant	echo
4096	0	30.03	3.18	v. little	v. faint echo
4096	1024	26.49	3.12	v. little	v. faint echo
4096	2048	21.14	2.99	little	faint echo

Table 4.2 Least Square Filter Results: $h_1(n)$

as filter length was increased. With a filter of length 4096 samples the filtered speech sounded very close to the original clear speech; in this case the remaining distortion was a very low amplitude copy of speech.

In contrast to $h_1(n)$, both $h_2(n)$ and $h_3(n)$ have zeroes outside the unit circle. As shown in Tables 4.3 and 4.4 least squares filtering for these impulse response was more problematic than for the minimum phase $h_1(n)$. Although the bulk of the reverberant energy was easily removed, the ensuing distortion was often of the form of an annoying ringing sound. One might expect that as the filter length was increased, a better result would be obtained as was the case with the minimum phase impulse response. As far as the error measures I_r and I_s indicated, this was true; however, the ringing distortion did not follow this rule. Ringing distortion was least, in both cases, for relatively short filter lengths with small delays. Furthermore, there was not a clear relationship between filter delay and this distortion; however for the longest filter lengths, the ringing effect was least with longer delays.

FIL LENGTH	DELAY	I_r (dB)	I_s (dB)	REMAINING REV	DISTORTION
1024	0	17.22	3.45	v. little	slight ringing
1024	256	15.33	3.19	little	slight ringing
1024	512	12.19	2.58	little	v. slight ringing
1024	768	8.14	1.99	worse	-
1024	1024	-8.64	-2.00	much worse	-
1536	0	19.69	3.65	v. little	ringing
1536	256	19.54	3.52	v. little	slight ringing
1536	512	18.36	3.30	little	ringing
1536	768	15.94	2.83	little	slight ringing
2048	0	20.78	3.73	none	ringing
2048	512	21.29	3.50	v. little	ringing
2048	1024	19.08	3.23	little	ringing
2048	1536	12.59	2.68	considerable	ringing
4096	0	21.88	3.85	none	ringing
4096	1024	25.49	3.53	none	ringing
4096	2048	26.54	3.67	none	ringing
4096	3072	20.09	3.41	little	slight ringing

Table 4.3 Least Square Filter Results: $h_2(n)$

In Figure 4.9 we show the least squares inverse filter of length 2048 and of delay 1024 for $h_2(n)$. This figure may be compared to the Fourier inverse of Figure 4.6. The convolution of the least squares inverse with $h_2(n)$, shown in Figure 4.10 for comparison with Figure 4.7, distributes the time domain error more evenly throughout the net impulse response, whereas the Fourier inverse filter concentrates the error in bursts at both ends heard as audible echoes.

4.4.3 Summary

The experiments with inverse filter design indicate that the removal of minimum phase impulse responses through inverse filter design is achieved relatively easily, and that the distortion introduced, an audible echo of small amplitude, decreases as the

FIL LENGTH	DELAY	I_r (dB)	I_s (dB)	REMAINING REV	DISTORTION
1024	0	8.44	4.35	little	ringing
1024	256	9.99	3.13	little	v. slight ringing
1024	512	9.67	2.17	little	v. slight ringing
1024	768	7.44	1.22	considerable	audible echo
1024	1024	-1.11	-0.15	worse	audible echo
1536	0	9.15	4.52	none	ringing
1536	256	11.67	3.58	v. little	ringing
1536	512	12.93	3.35	little	slight ringing
1536	768	12.75	2.84	little	slight ringing
2048	0	9.41	4.68	v. little	ringing
2048	512	14.23	3.72	v. little	slight ringing
2048	1024	14.66	3.31	little	ringing
2048	1536	11.28	2.77	considerable	ringing
4096	0	9.72	5.05	v. little	ringing
4096	1024	18.60	4.12	v. little	slight ringing
4096	2048	20.47	4.20	v. little	slight ringing
4096	3072	16.70	4.15	little	slight ringing

Table 4.4 Least Square Filter Results: $h_3(n)$

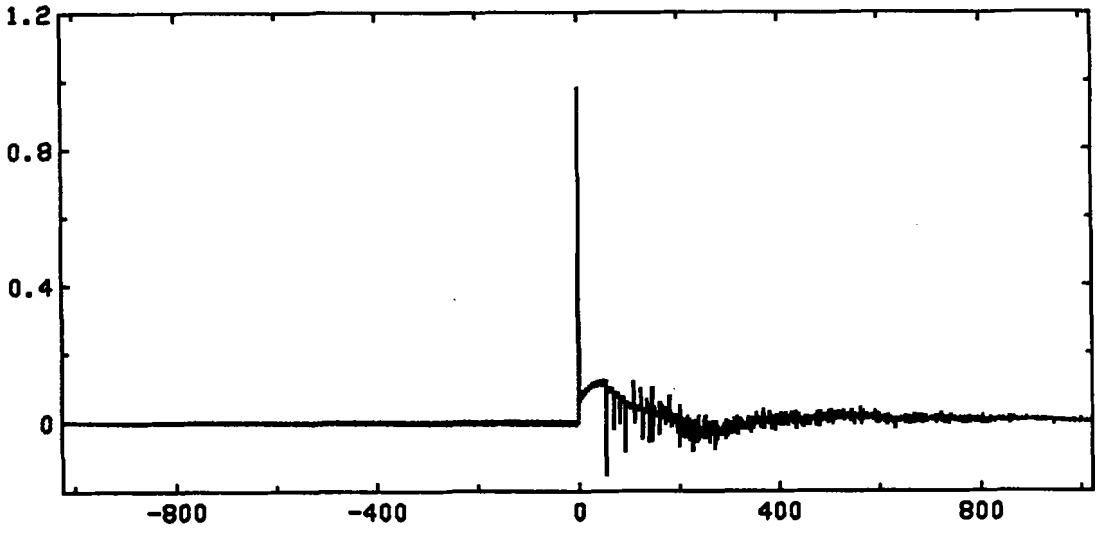

Fig. 4.9 Least squares inverse of $h_2(n)$

filter length increases. For the same filter length, the least squares filter removes more of the reverberant sound and introduces less distortion than Fourier inverse filters.

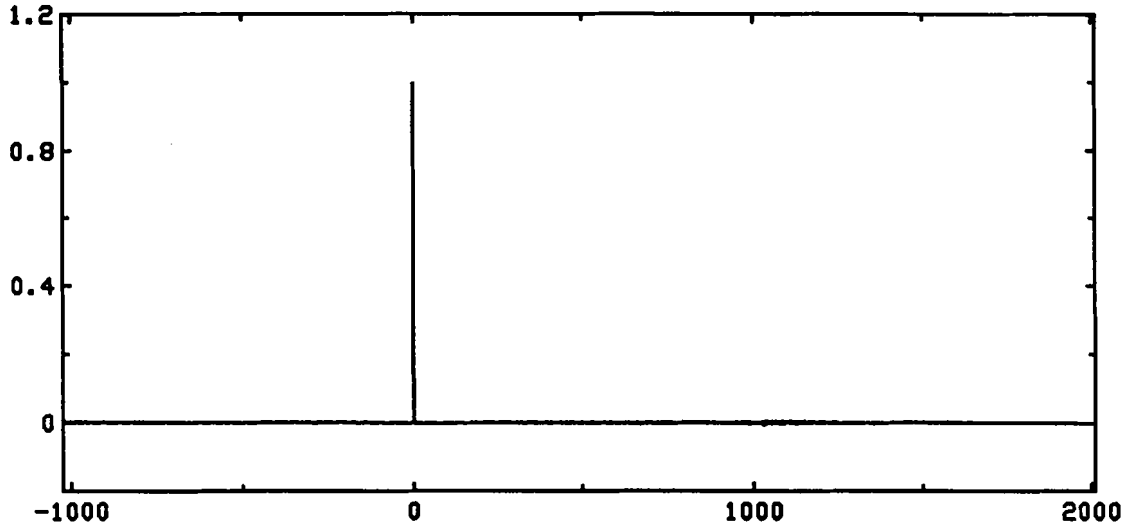

Fig. 4.10 Least squares inverse convolved with $h_2(n)$

For minimum phase impulse responses no filter delay is required.

The results also show that the removal of mixed phase impulse responses is a more difficult problem. The Fourier inverse technique was markedly inferior in the introduction of distortion than the least squares technique. This result was not surprising in view of the fact that simple truncation of the Fourier inverse sequence makes no attempt to optimize the use of the available filter coefficients. The least squares method chooses the filter coefficients according to the mean squared error criterion. However, the least squares filters left, in most cases, a ringing noise in the background of the filtered speech, and this noise often increased as the filter length was increased to moderately long values. The best least squares results seemed to be a compromise between the reduction of reverberant energy and "room effect", and the introduction of this ringing distortion. This compromise was best achieved with filter lengths on the order of the length of the impulse response and with short delays. Alternately, least squares filters with very long lengths and delays removed the room

effect well and kept the ringing noise low in amplitude.

Chapter 5 Experimental Results

The approaches to dereverberation of speech through the cepstral identification of the impulse response and its removal through linear filtering were presented in the previous chapters. In this chapter we present the experimental results of applying these methods.

The experiments described here test the system described in Figure 3.6. In this system, reverberant speech is segmented with segment starts defined to begin after silent gaps; segments are multiplied by an exponentially weighted window function; the cepstrum is calculated using the spectral magnitude only; cepstra from several segments are averaged together; a cepstral estimate of the impulse response is made from the cepstral average; a time domain impulse response estimate is made by inverse transforming the cepstral estimate and de-weighting; the inverse filter is designed from the time domain impulse response estimate using the least squares technique; and the filter is applied to the reverberant speech.

For these tests, a sentence of speech digitized at 8000 Hz was convolved with sequences which represented various reverberation impulse responses, and the resulting speech sentences were the inputs to the cepstral filtering procedure. The sentence used, of duration approximately 10 seconds, was "The software described in this doc-

ument is furnished under a license and may only be used or copied in accordance with the terms of such license". A male speaker was used. To simplify the tests, in each experiment the impulse response was kept constant over the entire sentence. The cepstral average was taken over each entire sentence, and the resulting filter was then convolved with the sentence to form the enhanced speech. This method thus simulates the operation of the system in Figure 3.6 after a "training" period. To further simplify the tests, the silent gap detection was made by examination of the speech waveforms on a display terminal.

5.1 Description of Experiments

In the first experiment, a single echo of delay 37.5 ms was added to the speech. The object of this experiment was to verify that the segmentation, windowing, and averaging procedures allow an accurate estimation of the impulse response in the case of a simple echo.

The second experiment tested the ability of the system to identify and remove a more complicated sequence of echoes which combine to form a minimum phase impulse response. Both identification and filtering tests were performed.

In the third experiment, the image source model of [17] was used to simulate a minimum phase room impulse response. The object of this test was to simulate the operation of the dereverberation system in the case of a relatively small speaker-microphone distance. Also in this experiment, investigations were made which provided insight into the sources of error in the estimation procedure.

The object of the fourth experiment was to test the dereverberation of speech under the conditions of a mixed phase impulse response. Such an impulse response,

which may correspond to a larger speaker-microphone separation or a room with more reflective surfaces, requires a 2-sided filter for best results. In this experiment, the exponential weighting factor chosen was sufficient to convert the impulse response to minimum phase. This impulse response had components near the speech pitch period and one of the objects of this test was to determine if this components could be identified.

The object of the fifth experiment was to evaluate the result of an inappropriate choice of exponential weighting factor in the case of a mixed phase impulse response. In a real scenario, it is probable that the degree of exponential weighting required to ensure conversion of the impulse response to minimum phase would be unknown. Thus, it is important to understand the consequences of choosing a value of exponential weighting insufficient to achieve this conversion.

5.2 Experiment 1: Simple Minimum Phase Echo

For the first experiment, an echo of amplitude 0.8 and delay 300 samples (37.5 ms) was applied to the speech. The object of this test was to verify that the proposed identification step of the cepstral filtering procedure was valid in this simple case. From the reverberant speech, 11 segments were chosen by examination of the waveform. The segment starts were chosen to correspond to the estimated locations of beginning of speech activity after visible pauses, as explained in Chapter 3. For example, Figure 5.1 shows one such segment. In this figure, the segment start was defined at sample number 4550.

For this experiment the impulse response was minimum phase and the complication of choosing an exponential weighting sufficient to convert it to minimum phase

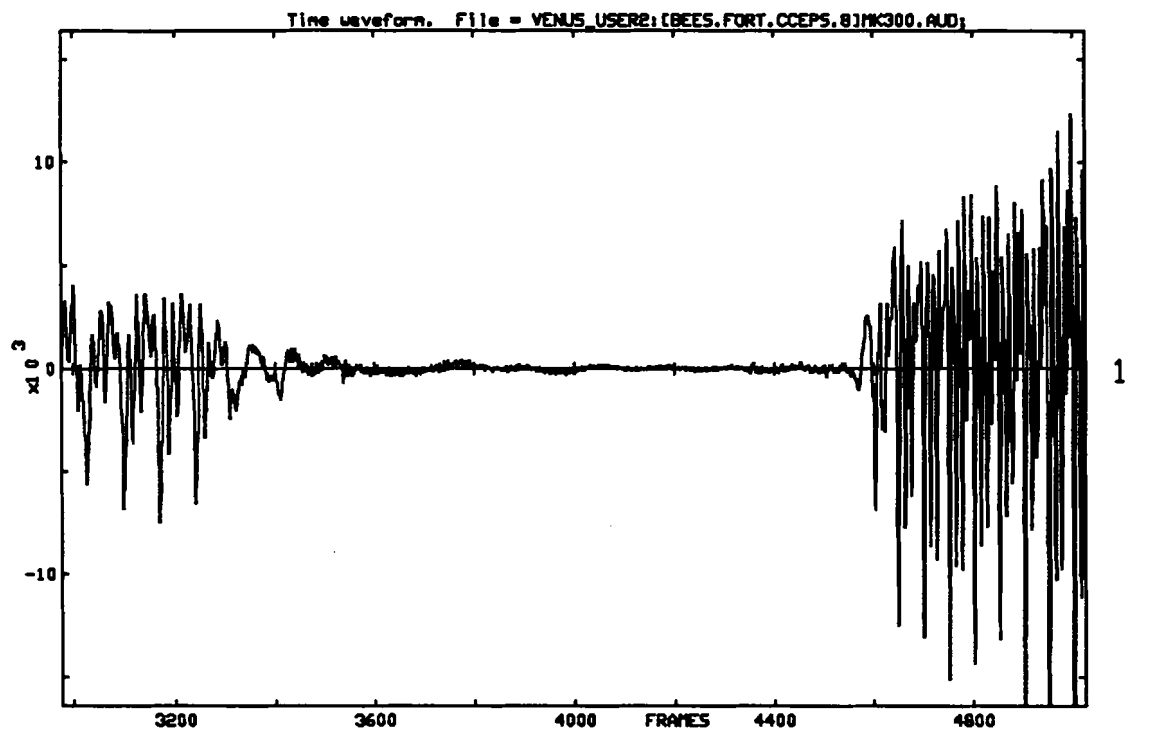

Fig. 5.1 Ex 1- Section of reverberant speech

was not present. A moderate exponential weighting factor of 0.9995 was chosen with a large data buffer size of 8192 and DFT length of 16384. The cepstrum was computed from the log magnitude of the DFT samples, and the first 1024 quefrency samples of the 11 cepstra, with their average at the bottom of the figure, are shown in Figure 5.2. For comparison, the cepstra computed with the segment start times selected at random are shown in Figure 5.3.

Examination of Figure 5.2 shows the success of the computation of the complex cepstrum of the echo. The echo cepstral peaks, which correspond very closely to those expected from theory, are clearly distinguished from the background noise, and peak-picking can easily be used to separate the two. This is also true for the random

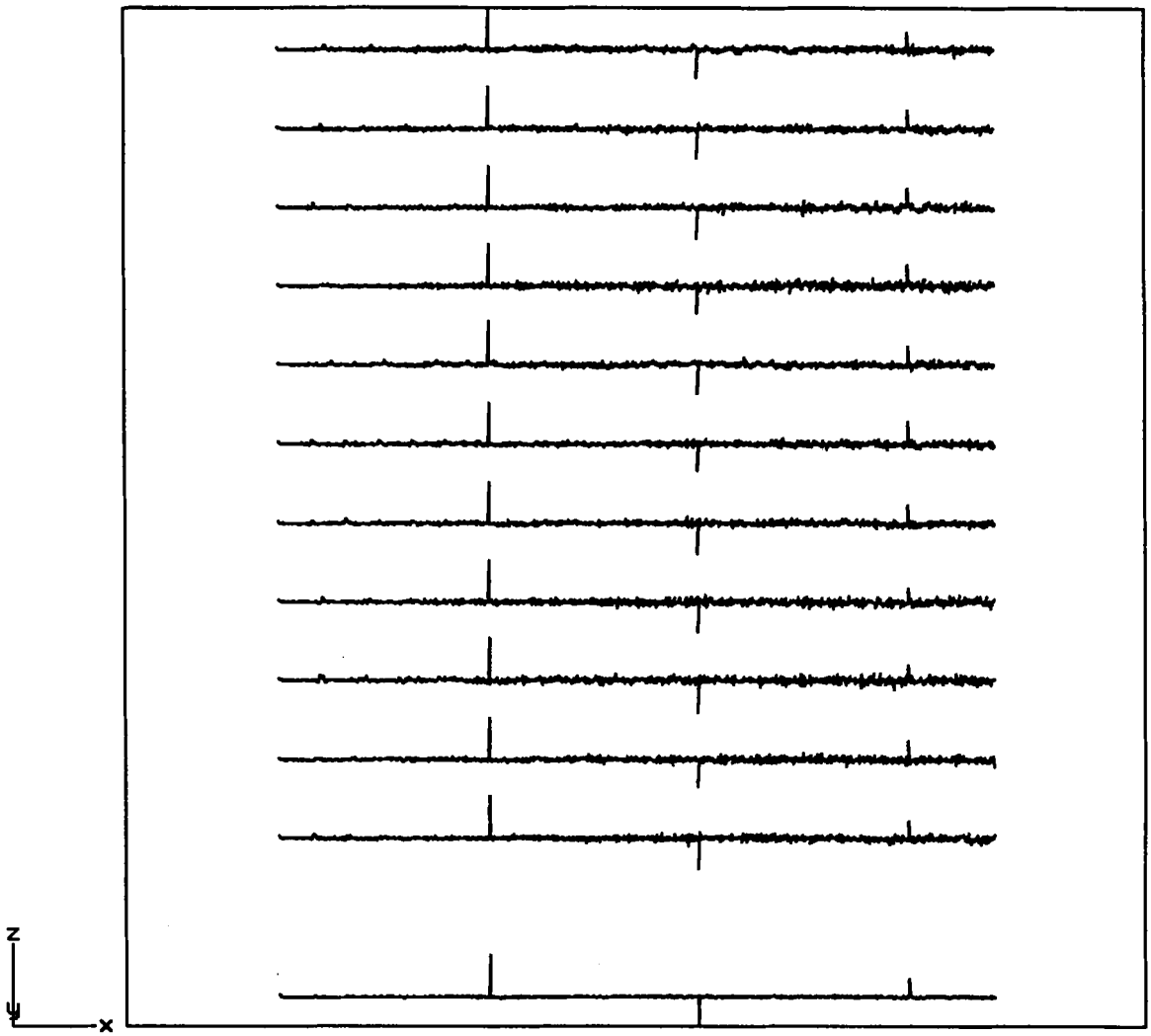

Fig. 5.2 Ex 1- Cepstra ($\hat{x}(n), 0 \leq n < 1023$) of speech with single echo (segment starts selected)

segment start times of Figure 5.3, but the echo peaks in this case are quite variable and their averages are much lower in magnitude than the theoretical values.

Figures 5.4 and 5.5 show the estimated impulse responses from the averaged cepstra of Figures 5.2 and 5.3, respectively. These were calculated by truncating the cepstra at 800 samples (corresponding to 100 ms) and by peak picking so as to preserve only the echo peaks. Both resemble the actual impulse response but the value of the echo in Figure 5.4, 0.785, is much closer to the true value, 0.8, than

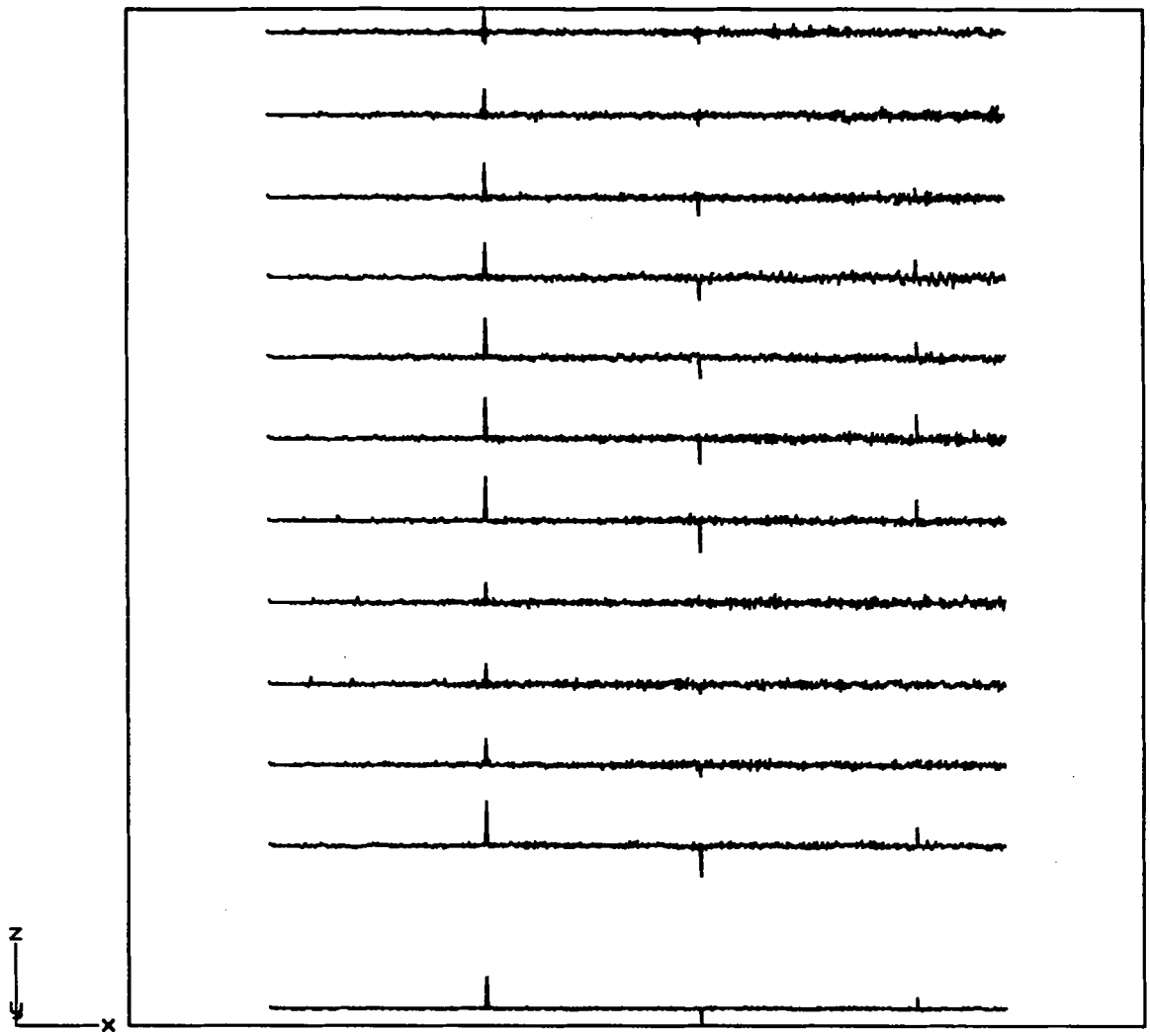

Fig. 5.3 Ex 1- Cepstra ($\widehat{x}(n), 0 \leq n < 1023$) of speech with single echo (random segment starts)

that for Figure 5.5, 0.572. The design of a linear filter from the estimate of Figure 5.4 is straightforward and obviously a more accurate removal of the echo is possible than with a filter designed from the estimate of Figure 5.5. This experiment shows that for the case of simple echoes which are minimum phase the proposed cepstral identification procedure is capable of very accurate impulse response estimation.

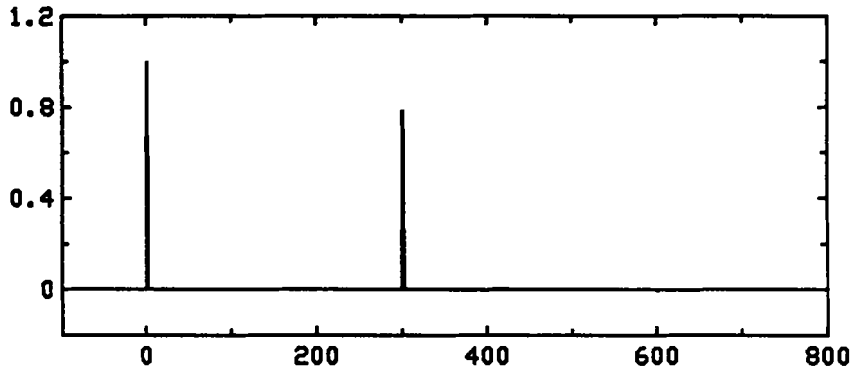

Fig. 5.4 Ex 1- Estimated impulse response (segment starts selected)

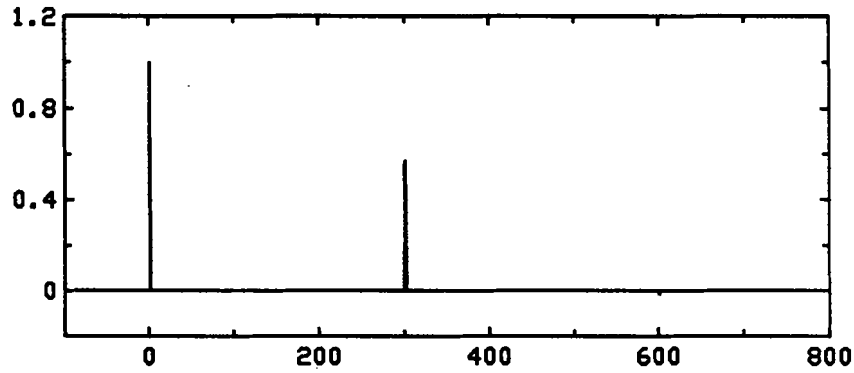

Fig. 5.5 Ex 1- Estimated impulse response (random segment starts

5.3 Experiment 2: Multiple Echoes

In this experiment the object was to investigate the ability of the proposed system to identify and remove a minimum phase impulse composed of many echoes of varying amplitude over a wide range of delays. Such an impulse response is shown in Figure 5.6. It was verified to be minimum phase using the method described in Section 4.4. The duration of this impulse response is slightly over 200ms, and it has a direct to reverberant energy ratio of -1.94 dB. The reverberant speech is extremely distorted and unpleasant to listen to, with a harsh, metallic sound. Using the same processing

parameters as in experiment 1, the cepstral average was calculated and is shown in Figure 5.7. After cepstral peak picking with a threshold of 5.0 after scaling, the resulting estimated impulse response is shown in Figure 5.8. From this estimate, truncated at 1200 samples, a linear filter of 1500 taps, shown in Figure 5.9, was designed. Since the impulse response was known to be minimum phase, a least squares filter of zero delay was specified. The application of the filter to the reverberant speech produced a striking improvement in quality. All of the harsh reverberant sound has been removed; the residual echo remaining is of low amplitude and produces a pleasant "full" sound. The convolution of the filter and the impulse response is shown in Figure 5.10.

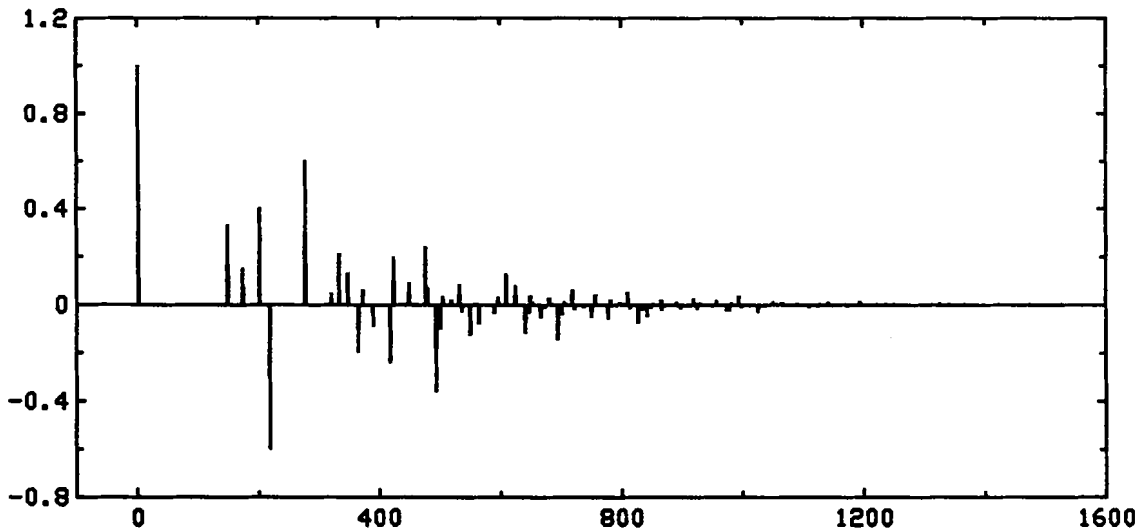

Fig. 5.6 Ex 2- Impulse response

For this case, the direct to reverberant energy ratio has been improved by $I_r = 16.91$ dB, and the spectral colouration gain is $I_s = 6.65$ dB. To investigate the effect of the background cepstral noise, which had been largely removed through peak picking, the cepstral average was again transformed to the time domain, but this time

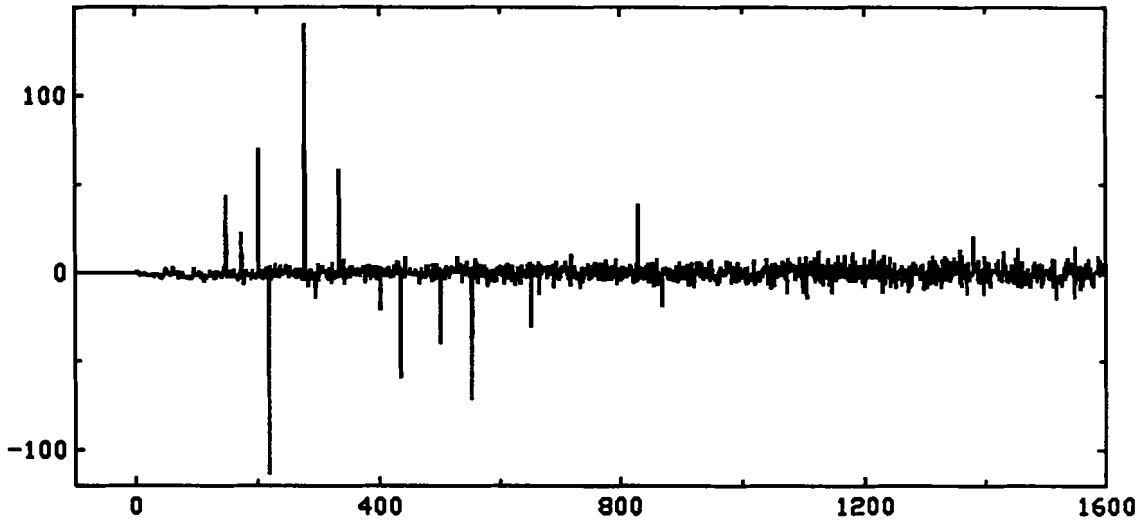

Fig. 5.7 Ex 2- Cepstral average of reverberant speech

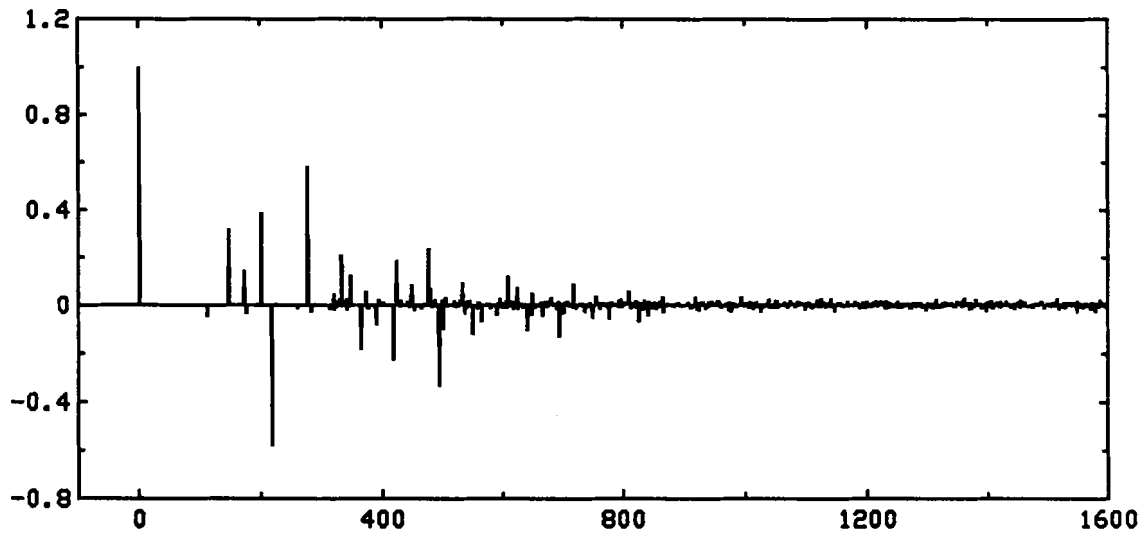

Fig. 5.8 Ex 2- Estimated impulse response from peak-picked averaged cepstrum

a high pass window with a cut-off of 125 samples was applied to the cepstrum and no peak picking was performed. The high pass window allows the echo cepstrum to pass, but blocks the cepstral components of speech inside the first pitch period. The resulting time domain impulse response estimate is shown in Figure 5.11, the least

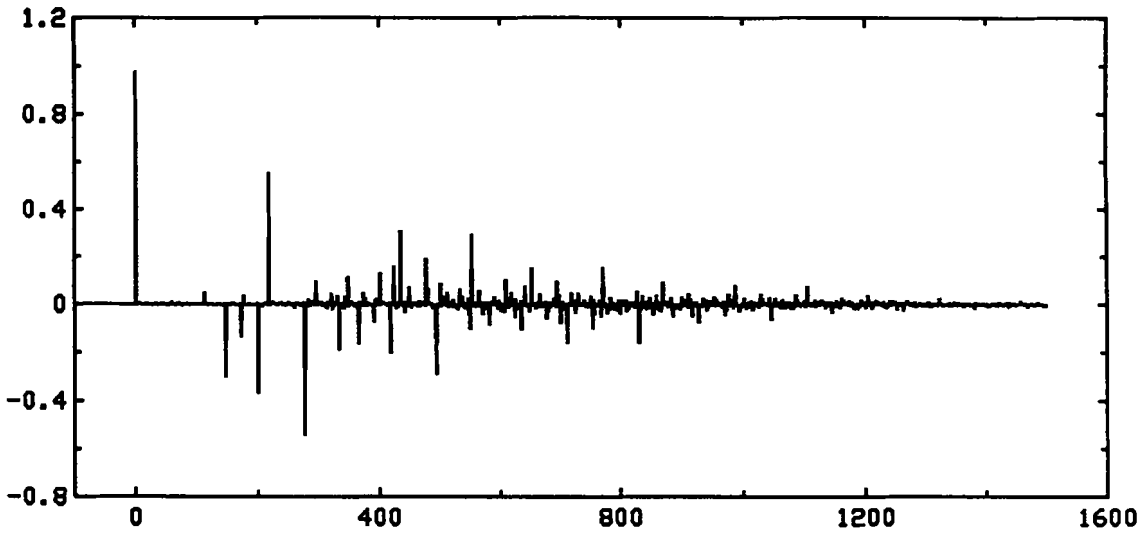

Fig. 5.9 Ex 2- Least squares filter designed from impulse response
estimated with peak picking

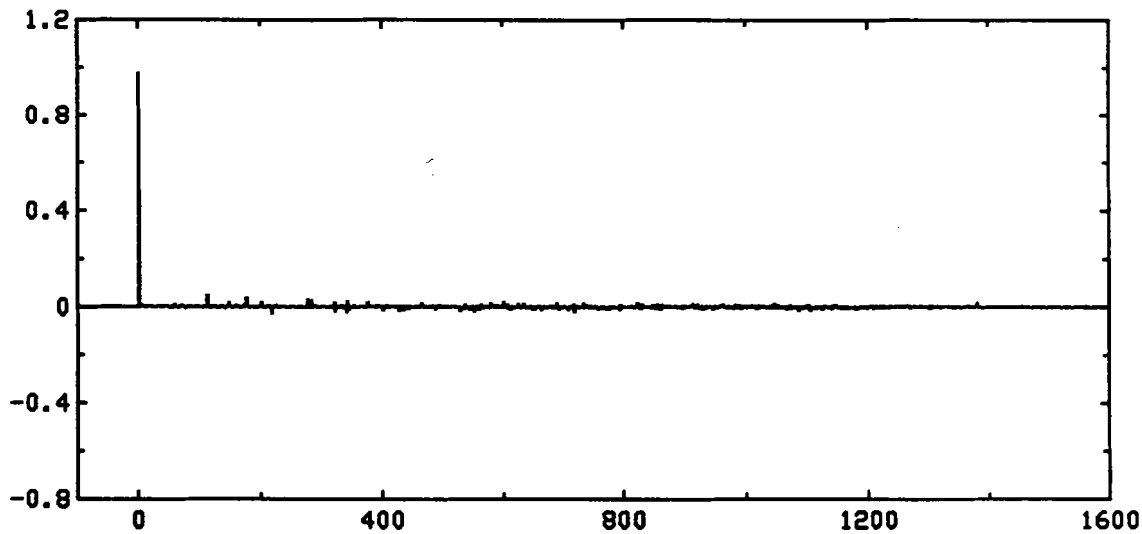

Fig. 5.10 Ex 2- Convolution of filter and actual impulse response

squares filter designed from it in Figure 5.12, and the convolution of the filter with the actual impulse response in Figure 5.13. The increased noise in the impulse response estimate is apparent; however, the noise remains small in comparison the magnitude of the echo components up until about 1000 samples. Low level noise is also apparent

in the least squares filter, and results in low level, noisy echo components in the convolution of the filter with the impulse response. The direct to reverberant processing gain is $I_r = 13.12$ dB, and the spectral colouration processing gain is $I_s = 6.10$ dB. Although these values are lower than those achieved with the peak picked case, the improvement in sound with this filter is almost as great. It thus appears that with the large echo components of this impulse response, the background cepstral noise does not play a seriously degrading role in the enhancement. This experiment confirmed the ability of the proposed system to perform effective enhancement in the presence of multiple, discrete echoes forming a minimum phase impulse response.

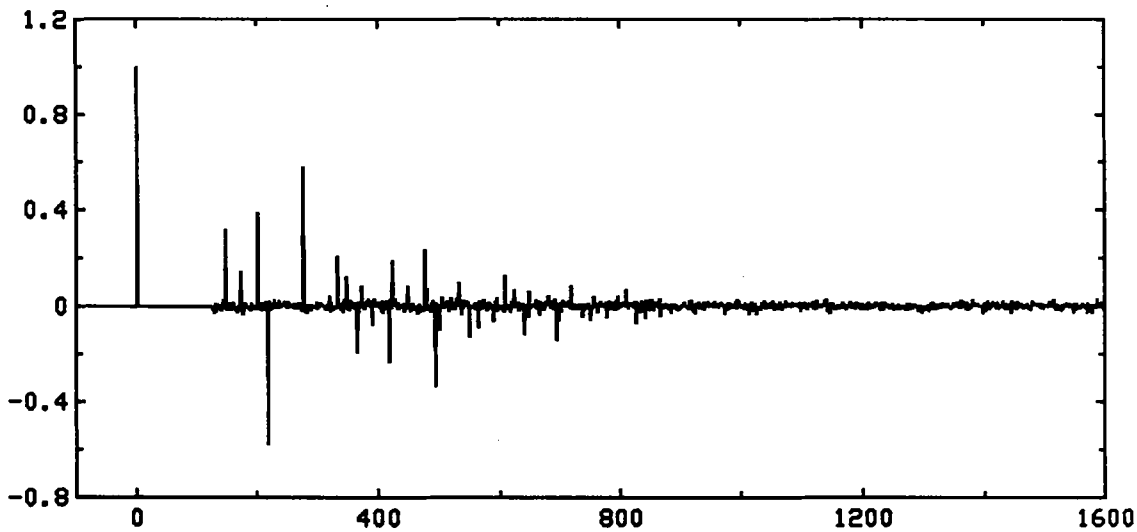

Fig. 5.11 Ex 2- Estimated impulse response from high-passed cepstral average

5.4 Experiment 3: Minimum Phase Room Impulse Response

The previous two experiments have shown that echo series composed of sharply defined, discrete peaks, which are minimum phase, may be accurately identified using

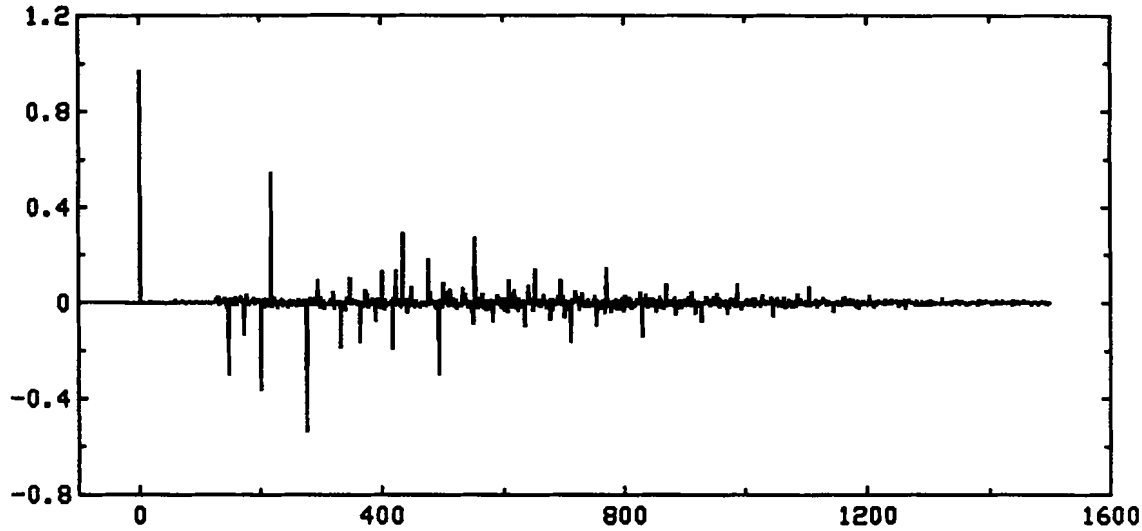

Fig. 5.12 Ex 2- Least squares filter designed from impulse response
estimated with high-passed cepstral average

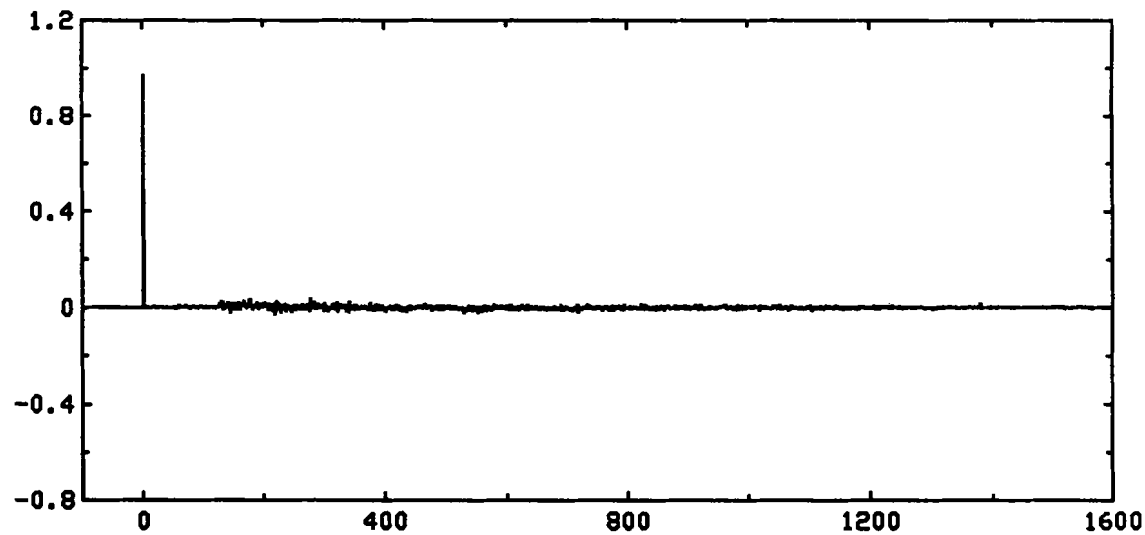

Fig. 5.13 Ex 2- Convolution of filter and actual impulse response

the cepstral techniques. In these cases there was a sufficiently large difference between the magnitude of the echo and the distortion caused by the cepstral noise that the two could be separated with peak picking. Although in some cases room acoustic responses can be minimum phase, it can not be expected that they will be composed of such

sharply defined echoes. In this experiment, we select a minimum phase simulated room response generated by the image model. This response, shown in Figure 4.2, is composed of echoes which are smaller in amplitude and dispersed in time to a greater degree than the discrete echoes considered above. The speech subjected to this reverberation sounds hollow, but no audible copies of the speech are present. The effect of this impulse response is not as disagreeable as that of the previous experiment, but the reverberation could be considered irritating.

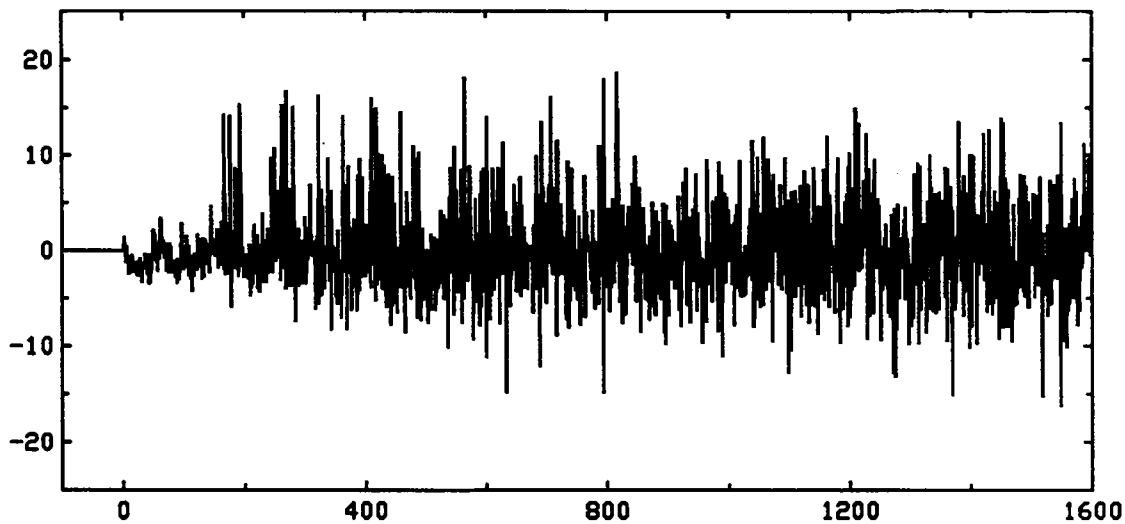

Fig. 5.14 Ex 3- Reverberant speech cepstral average

The averaged cepstrum, calculated as in previous experiments with a weighting of 0.9995, is shown in Figure 5.14. For comparison, the actual cepstrum of the impulse response is shown in Figure 5.15. It can be seen that in this case, the cepstral distortion is significant with respect to the impulse response cepstrum. Since the impulse response cepstrum, after linear scaling, decays with delay while the cepstral distortion slowly increases, the distortion becomes worse in effect as delay increases.

The impulse response estimate obtained from the high passed averaged cepstrum

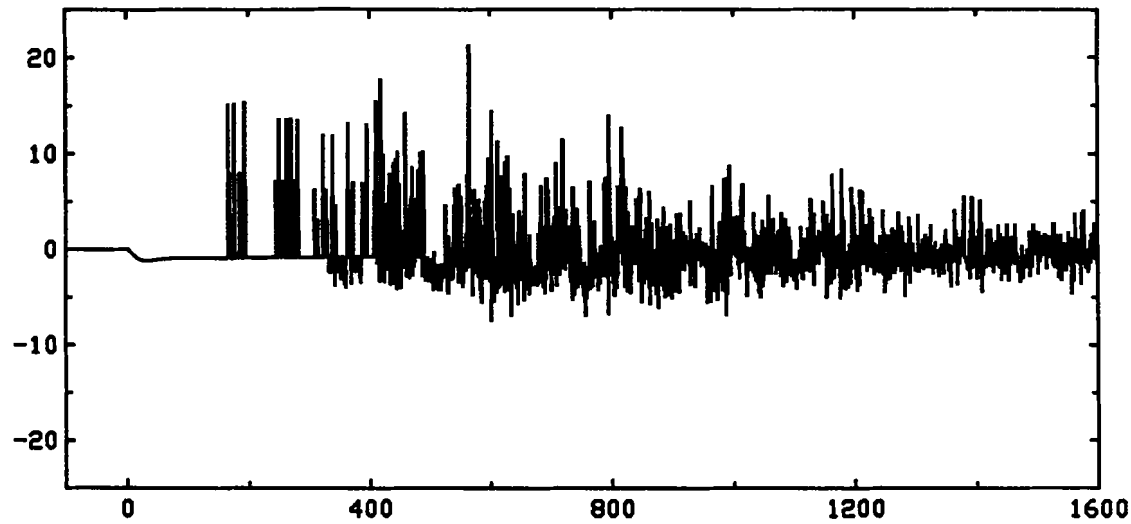

Fig. 5.15 Ex 3- Impulse response actual cepstrum

is shown in Figure 5.16, where the inverse transformation followed by exponential de-weighting was performed. Designing an inverse filter from this estimate is problematic; if the estimated impulse response is truncated too early, error is introduced by ignoring later impulse response terms, but truncation too late will include terms where the distortion is greater in magnitude than the impulse response. As an example of the filter design which can be performed with this estimate, the convolution of a linear least squares filter designed by truncating the estimate at 1000 samples and the actual impulse response is shown in Figure 5.17. The larger echo peaks have clearly been attenuated by the filter but the price paid is greater distortion at large delay values. The direct to reverberant energy ratio has been improved by 2.08 dB, and the spectral colouration by 0.51 dB. A somewhat better filter can be designed in this case by peak-picking the earlier regions of the impulse response estimate or the cepstrum to separate the sharper peaks from the noise.

It is useful to examine this case further to gain understanding of the sources of

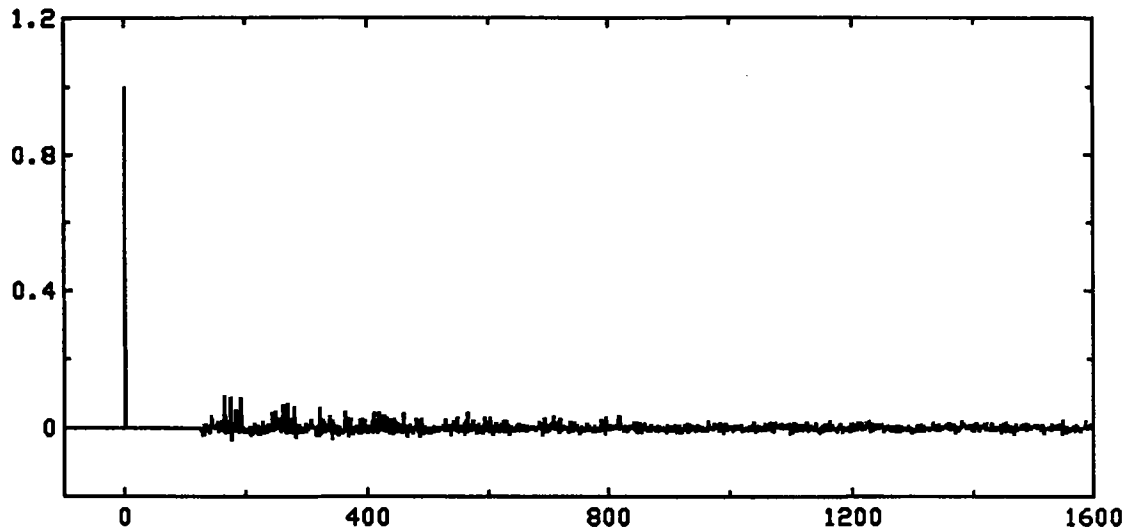

Fig. 5.16 Ex 3- Impulse response estimate from high passed cepstrum

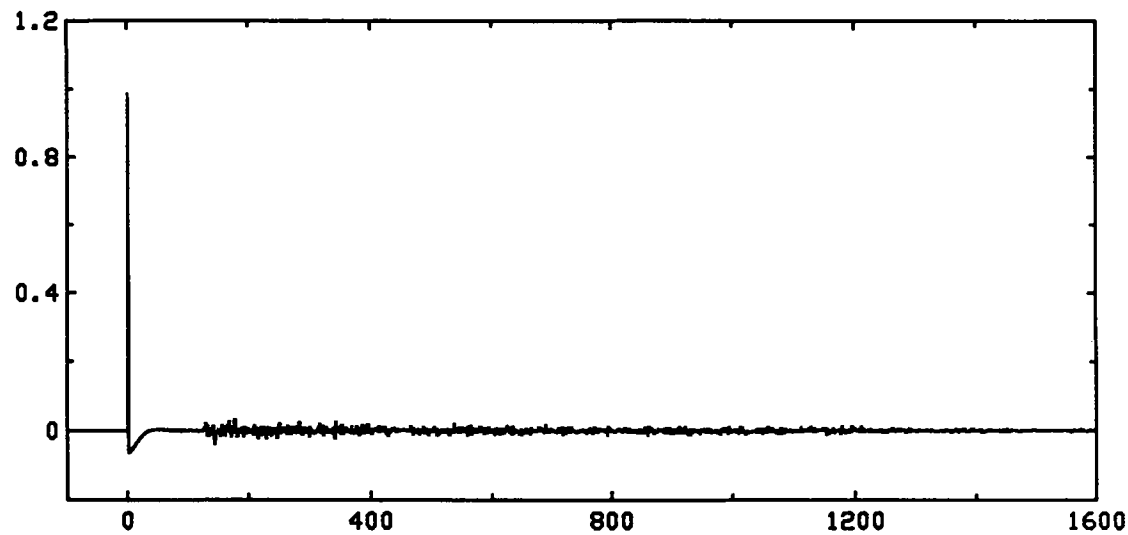

Fig. 5.17 Ex 3- Convolution of filter and actual impulse response

the cepstral distortion. Since it is known that the impulse response was minimum phase, two sources of error were investigated: residual speech cepstrum left after averaging and residual segmentation error. Since the original speech was available the first source of error is easy to evaluate. To do so, the cepstrum of the original

speech was calculated in exactly the same locations as for the reverberant speech, with the same processing parameters. The averaged speech cepstrum is shown in Figure 5.18. It matches closely the apparent additive cepstral distortion in Figure 5.14. The time domain, de-weighted counterpart to Figure 5.18 is shown in Figure 5.19, and represents the distorting function which would be convolved with the actual impulse response in our estimate. This function, surprisingly, appears to match the *additive* time domain distortion. This is explained by the nature of the two convolved impulse responses, both of which are formed of a large peak at $n = 0$ and a series of much smaller components at $n > 0$. For functions of this type, their convolution is similar to their sum.

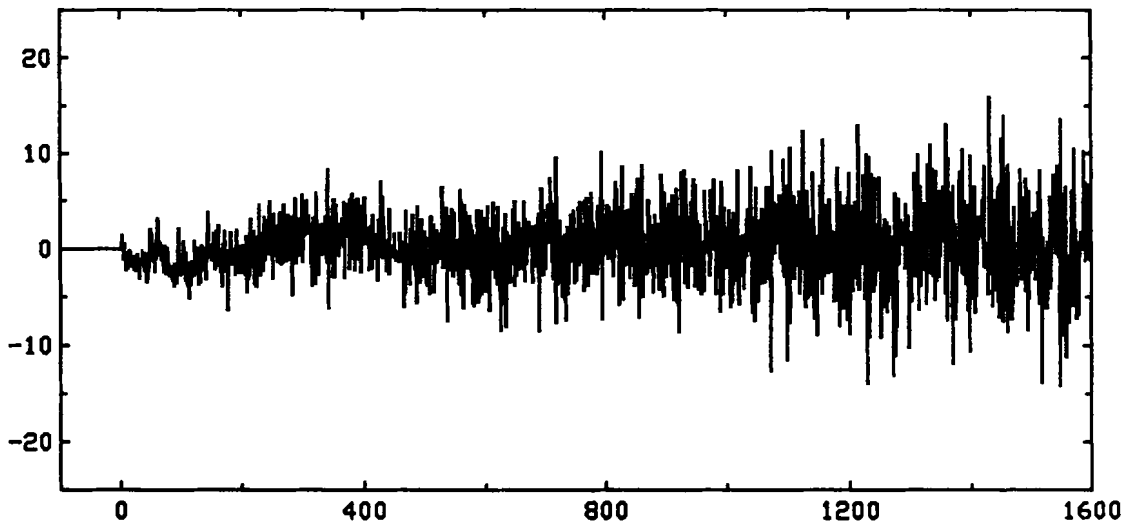

Fig. 5.18 Ex 3- Averaged speech cepstrum

The above plots indicate that the residual speech cepstrum in this example is a primary source of error. To confirm this, the averaged speech cepstrum was subtracted from the reverberant averaged cepstrum, and the resulting time domain de-weighted sequence is shown in Figure 5.20. Clearly, a much better estimate is now obtained.

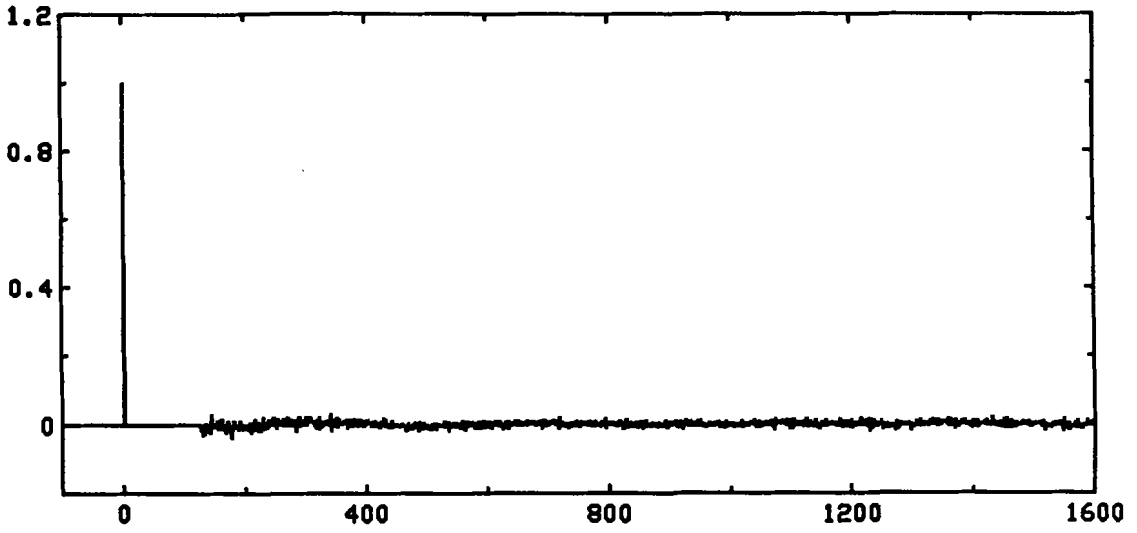

Fig. 5.19 Ex 3- Time domain representation of averaged speech cepstrum

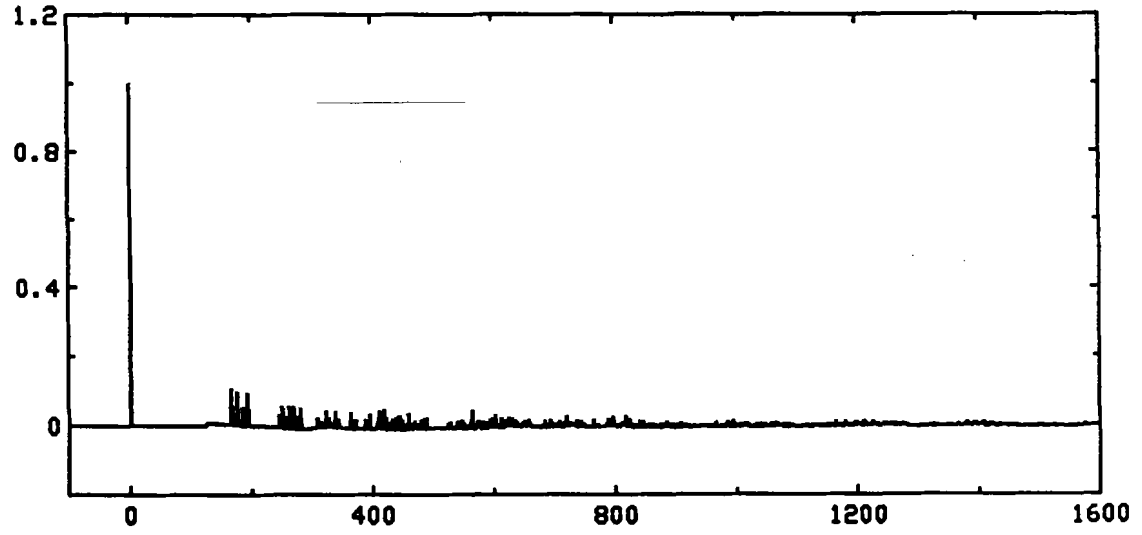

Fig. 5.20 Ex 3- Impulse response estimate from high passed reverberant speech cepstrum − speech cepstrum

Next we investigated the degree to which residual segmentation error contributed to cepstral distortion and impulse response estimation error. With the exponential window, as shown in Chapter 3, the segmentation error at the end of the segment may be neglected, but we expect any error at the start of the segment caused by intrusion

of the echo of the previous segment to cause distortion. We saw in experiment 1 that the segment placement selection dramatically improved the cepstral estimation for the discrete echo. This echo, however, was of relatively short delay time, and it might be expected that with the longer impulse response of experiment 3 that the segmentation error remaining would be more severe. That is, since the echo duration is greater than 200 ms, a speech pause of greater length than this is required to define a segment with no segmentation error at the start. Figure 5.20 shows that this in spite of this, the impulse response has been well estimated except for the distortion caused by the residual speech cepstrum. This may be due to the fact that the echo components at longer delay times are of small amplitude, and most of the echo energy is concentrated within 100ms.

We then postulated a relationship between heavy exponential weighting and segmentation error: a heavier weighting, which causes the window to fall off more sharply, shortens the effective buffer length, and the intrusion error at the segment start may become magnified in relative importance. In Figure 5.16 is shown the estimated impulse response of the reverberant speech processed cepstrally with a weighting of 0.997, and Figure 5.22 shows the estimate for this speech when the speech cepstrum is subtracted from the reverberant speech cepstrum prior to conversion to the time domain and de-weighting. It is now seen that with the heavier weighting, the residual speech cepstrum also plays a large role in the distortion but significant distortion apart from this effect is present. It seems likely that this extra distortion is due to the shortened exponential window amplifying remaining segmentation error.

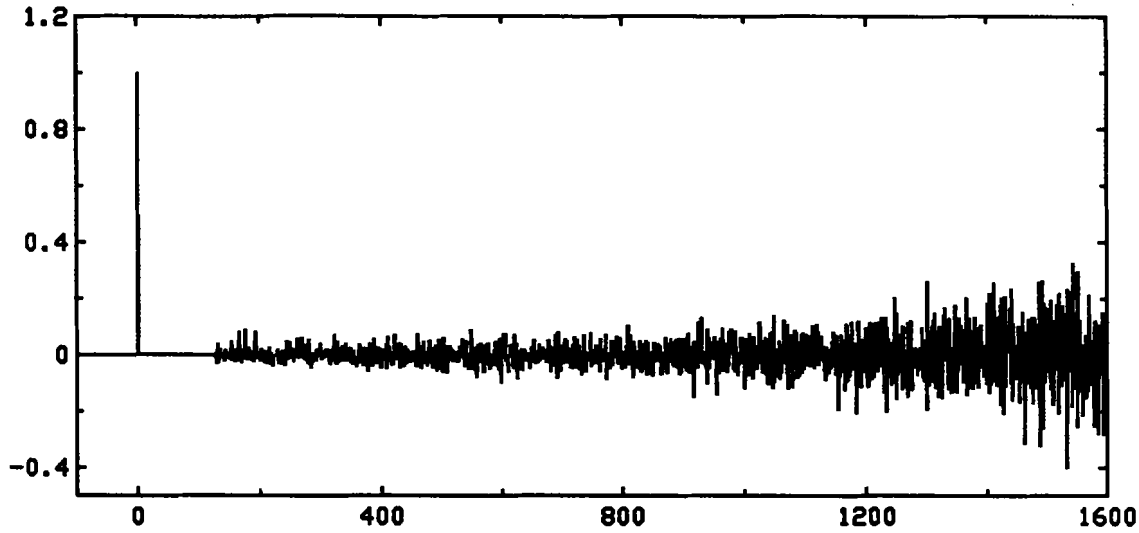

Fig. 5.21 Ex 3- Impulse response estimate from high-passed cepstrum with heavy exponential weighting

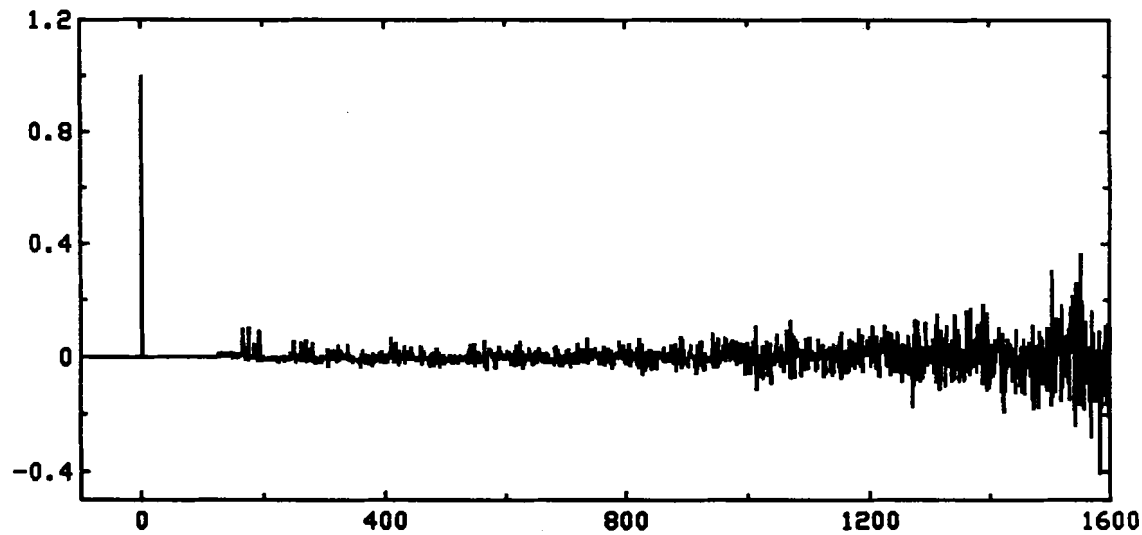

Fig. 5.22 Ex 3- Impulse response estimate from high-passed reverberant speech cepstrum − speech cepstrum with heavy exponential weighting

5.5 Experiment 4: Mixed Phase Room Impulse Response

The object of the fourth experiment was to test the dereverberation system in the case of a mixed phase impulse response with echo components near the pitch

period. The impulse response $h_2(n)$, shown in Figure 4.3, which is mixed phase, and which has components at delay times within a range corresponding to common pitch periods, was used. The exponential weighting factor of 0.999 was used in the calculation of the reverberant speech averaged cepstrum, and we see from Table 4.1 that this weighting factor converts $h_2(n)$ to minimum phase.

We conjecture that within a cepstral range extending from $n = 0$ to some cut-off value, say $n = 150$, that by peak-picking we may identify cepstral components due to the impulse response and reject much of the cepstral "noise" due to speech pitch components. In Figure 5.25 the first 150 samples cepstrum of each segment is displayed together with the average over the 11 segments. By careful examination of Figure 5.25 it is possible to see that in each segment, the peaks due to the echo cepstrum are constant. The peaks due to speech voicing components, which dominate in any individual segment the echo peaks, move around in location from segment to segment. For example, in the first 4 traces (from the bottom) the voicing peaks move gradually from about 50 to about 60 samples. The net effect of the variability of the speech cepstrum is that the averaged cepstrum isolates relatively effectively the echo cepstrum.

The first 300 averaged cepstral samples before and after peak-picking with a threshold of 5.0 (after scaling) are shown in Figure 5.23 and Figure 5.24, respectively. By comparison of the actual cepstrum of $h_2(n)$, shown in Figure 5.26, with the peak picked average of Figure 5.24, we see that the averaging and peak-picking strategy is successful to a large extent. However, some of the cepstral peaks and all of the low level cepstral components due to $h_2(n)$ are lost, and some spurious peaks are introduced.

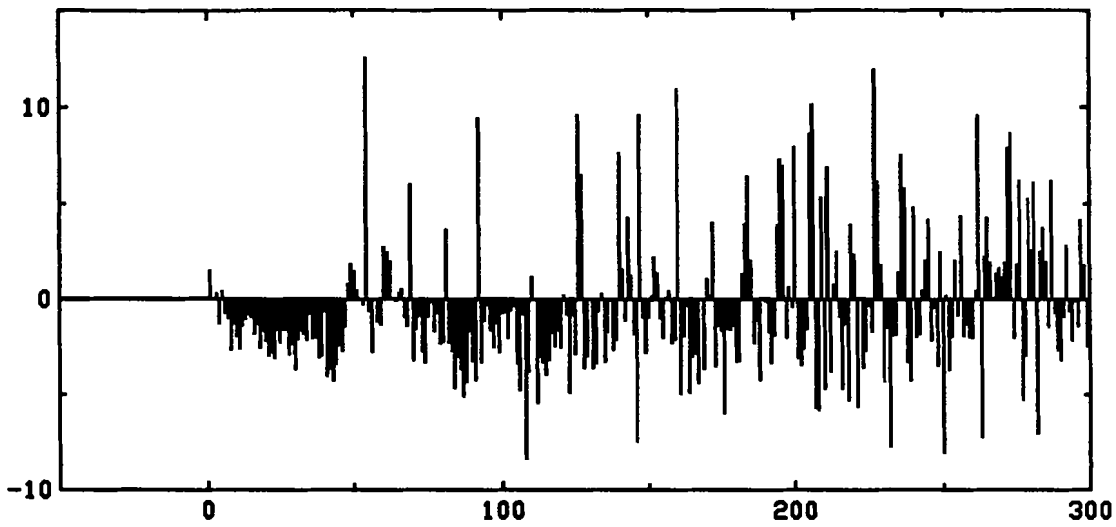

Fig. 5.23 Ex 4- Reverberant speech averaged cepstrum ($h_2(n)$)

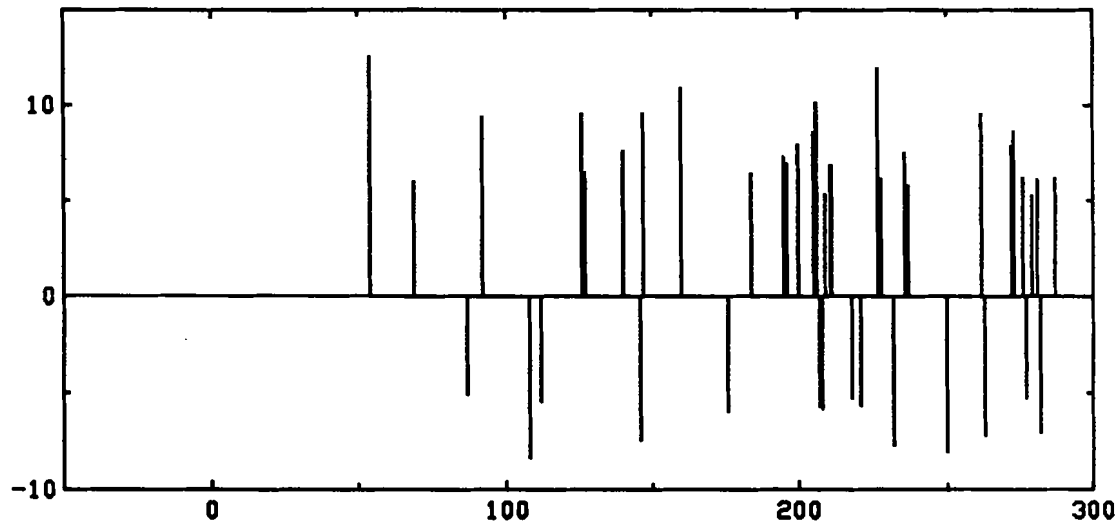

Fig. 5.24 Ex 4- Reverberant speech peak-picked averaged cepstrum ($h_2(n)$)

The estimated impulse response, calculated from the combined peak-picked average (from $n = 0$ to $n = 150$) and averaged (above $n = 150$) cepstrum is shown in Figure 5.27. The first and largest echo component, at $n = 56$, has been identified quite accurately as have been a large proportion of the other echo components in the

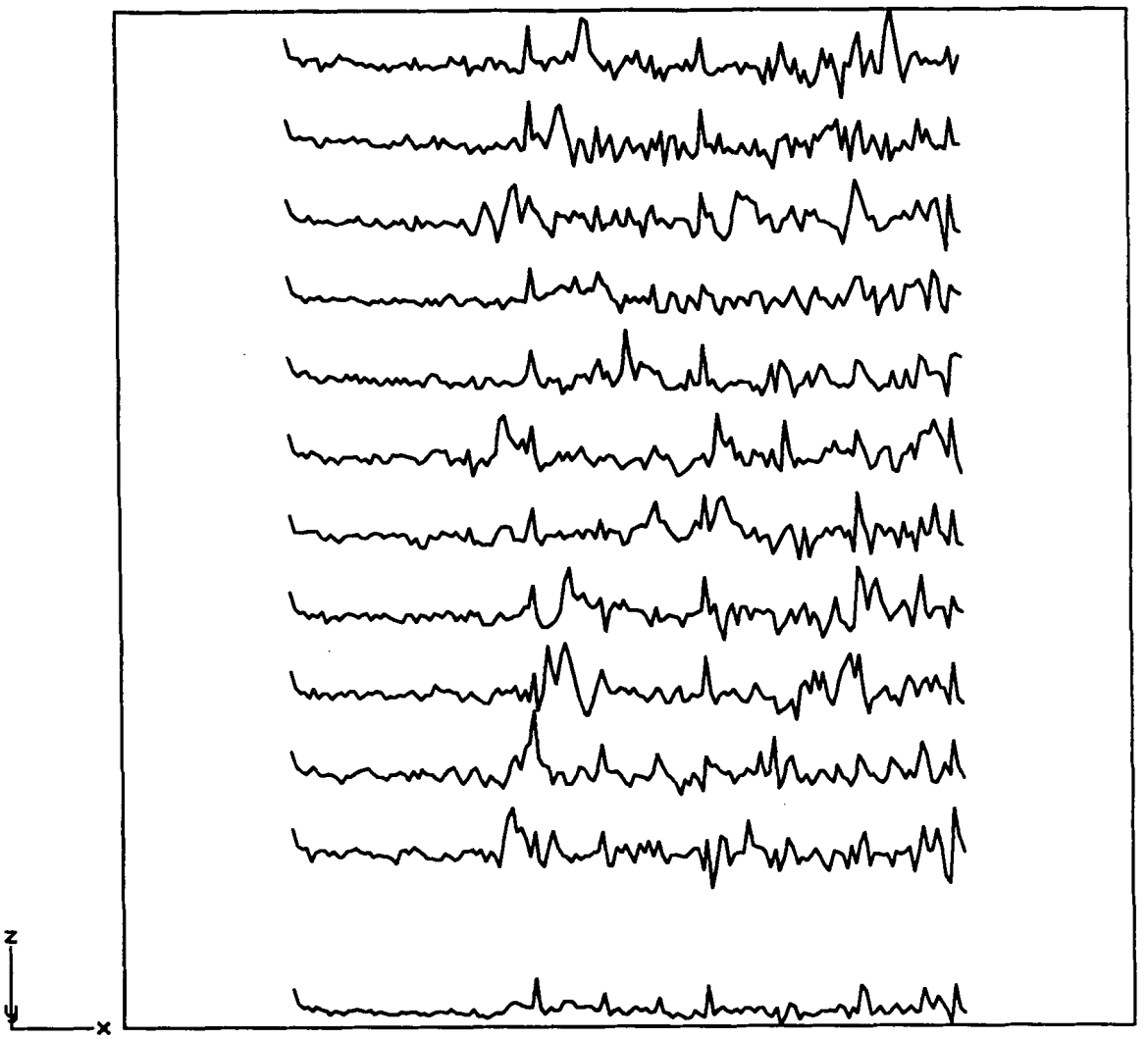

Fig. 5.25 Ex 4- First 150 cepstral samples in each segment

first 150 samples. As in previous experiments, the noise caused by residual speech cepstrum begins to dominate the estimate at later samples. A least squares filter of length 700 samples with a 200 sample delay was designed from the first 500 samples of this estimate. The net impulse response is shown in Figure 5.28. In listening tests the reverberant room effect has been diminished to some extent and no apparent distortion has been introduced. The direct to reverberant energy ratio shows an improvement of $I_r = 3.50$ dB and the colouration measure shows an improvement of

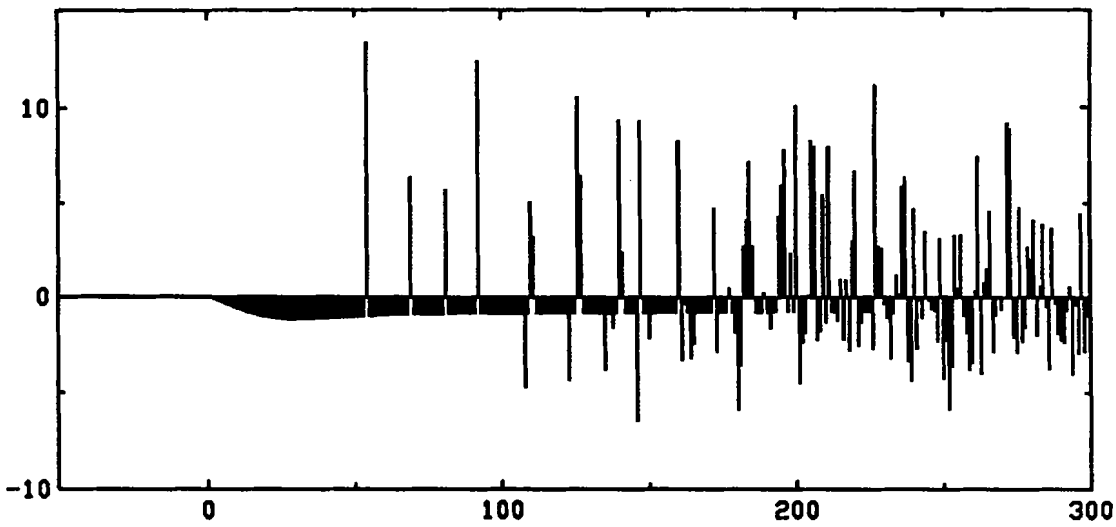

Fig. 5.26 Ex 4- Impulse response cepstrum ($h_2(n)$)

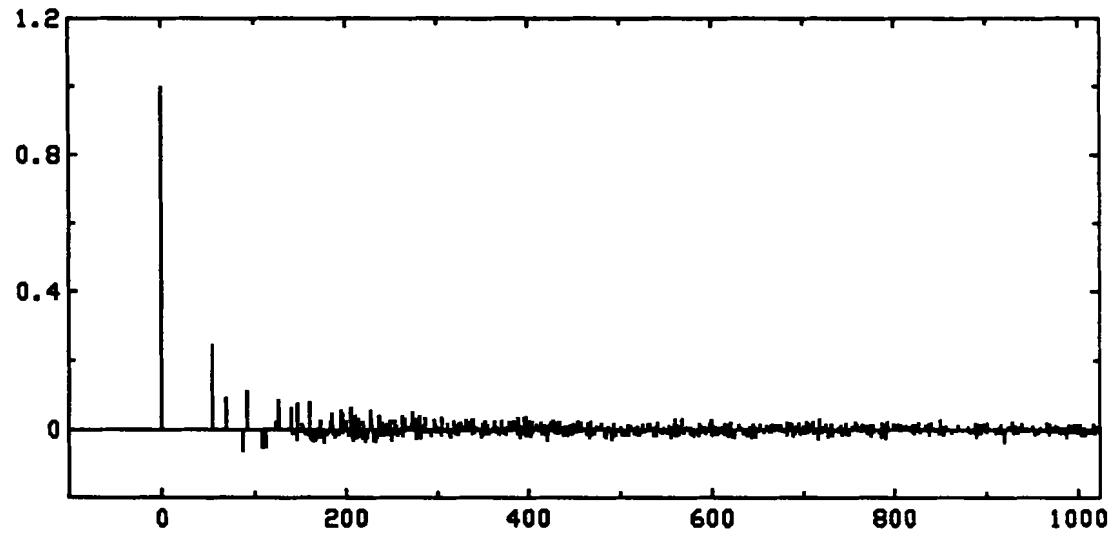

Fig. 5.27 Ex 4- Estimated impulse response

I_s = 0.97 dB.

5.6 Experiment 5: Mixed Phase Room Impulse Response

In a real application it is unlikely that the exponential weighting factor required to bring the reverberation impulse response z-plane zeroes inside the unit circle will

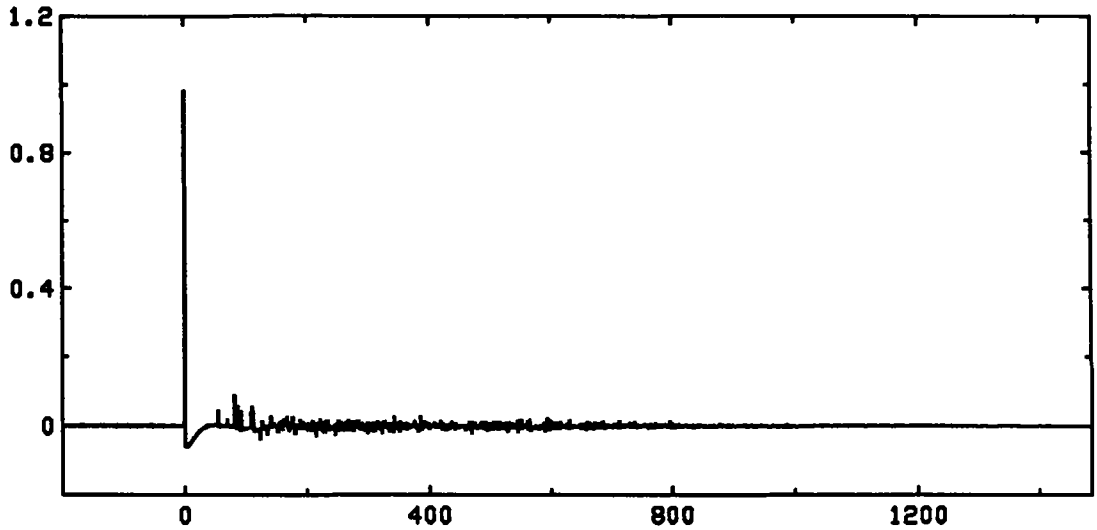

Fig. 5.28 Ex 4- Filter convolved with impulse response

be known. In this experiment we consider an example for which the exponential factor applied is insufficient. The impulse response is $h_3(n)$ and is shown in Figure 4.4. As shown in Table 4.1, a weighting of $\gamma = 0.9963$ or less is required to is required to convert $h_3(n)$ to minimum phase. We apply an exponential weighting of 0.999. The time domain impulse response was estimated as shown in Figure 5.29 from the reverberant speech averaged cepstrum (with peak picking in the first 150 samples). A filter with 800 taps and delay of 200 taps was designed from the estimated impulse response truncated at 600 samples. The convolution of the filter and the actual impulse response is shown in Figure 5.30, in which it can be seen that all of the echo peaks have been attenuated but low amplitude echoes have been introduced at delays up to 1000 samples. Improvements of $I_r = 4.30$ dB and $I_s = 1.95$ dB have been effected by the filtering. The resulting speech sounds less "boomy" and a harshness caused by this impulse response has been removed; however a feint "squeaky" distortion has been induced. It must be noted that in comparison to the

ringing distortions induced by the least squares filters of Chapter 4 designed from the actual impulse response, the distortion induced is, surprisingly, much less annoying.

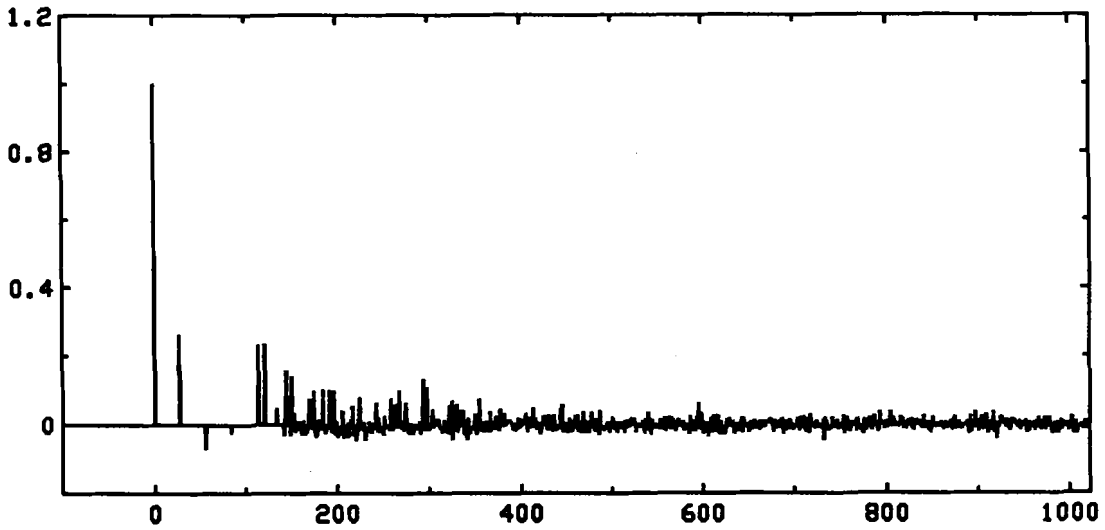

Fig. 5.29 Ex 5- Estimated impulse response

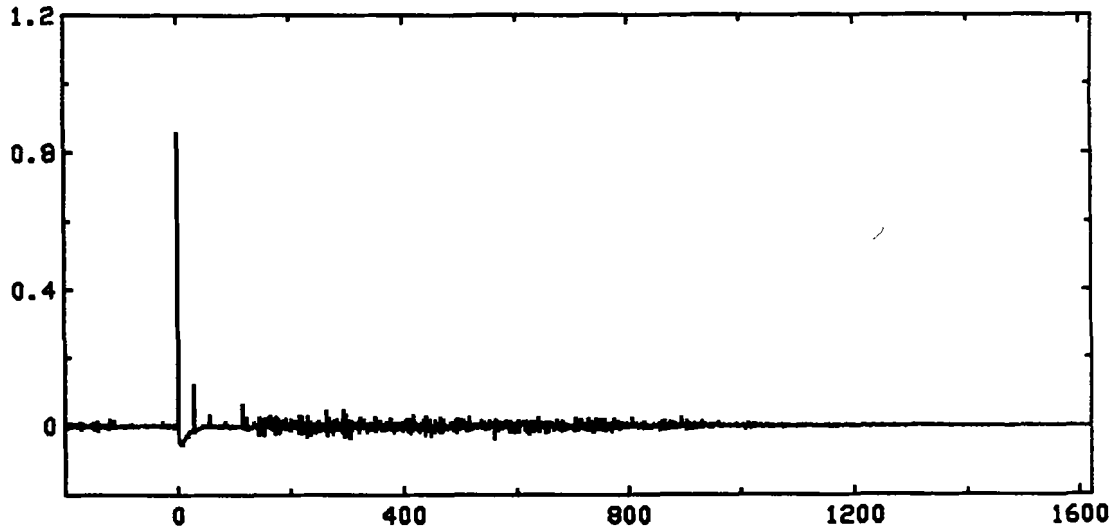

Fig. 5.30 Ex 5- Filter convolved with impulse response

5.7 Discussion of Experimental Results

The results described above appear to validate the ideas developed in Chapter 3 for the cepstral identification of the reverberant impulse response. Specifically, the definition of the segment start time as beginning after silent periods, combined with an appropriate degree of exponential weighting, reduces the segmentation error to a significant degree. Furthermore, cepstral averaging allows the exploitation of variability of the speech cepstrum to improve the isolation of the impulse response. The filtering step, using least squared error filter design as described in Chapter 4, was effective in most cases. For one of the mixed phase impulse responses, the filtered speech had a distorted sound. Whether this was due to the filter design or to a mis-identified impulse response is not clear.

The identification step was most successful when the impulse response was of a "peaky" nature. In these cases, as in experiments one and two, peak picking in the cepstrum allowed almost total rejection of the averaged speech cepstrum while preserving most of the impulse response cepstral components. Furthermore, peaky cepstral components due to the impulse response which occurred at delay times around the pitch period were identified quite accurately by exploiting the variability of the voice pitch from segment to segment.

It was clear, however, that for the impulse responses which were not composed of artificially discrete echoes, as in experiments three, four, and five, that some estimation error remained. There were several sources of error. Experiment three revealed that with moderate exponential weighting, most of the estimation error was due to residual speech cepstrum. Thus, it appears that this residual cepstrum represents a

limit to the accuracy of this technique for impulse responses for which peak picking in the cepstrum is not acceptable.

It was also shown that heavy exponential weighting in itself leads to estimation error. We speculate, with the evidence of the experiments performed in Chapter 3, that this error is due to residual segmentation error at the segment start. When a heavy weighting is applied, the effective length of the exponential window is shortened, and the weight of the window moves towards the front end. Any segmentation error not removed by the segment start location choice is thus amplified in relative importance.

The degree of exponential weighting applied is thus seen to be an important factor affecting performance. We wish to choose a weighting value which is sufficient to ensure that the impulse response is converted to minimum phase. Also, the exponential window must be of low value at the segment finish to ensure that the truncation error is reduced, but this may also be ensured by increasing the window length. Balanced against the reasons for increasing the exponential weighting factor is the concern about causing additional segment start error.

In experiment four, the exponential weighting factor was sufficient to convert the mixed phase impulse response to minimum phase, while in experiment five it was not. In comparing these two results, we note that experiment four led to enhanced speech with no appreciable distortion, while experiment five led to speech with a feint "squeaky" sound in the background. However, the direct to reverberant energy ratio and the spectral colouration measures were both improved more substantially in experiment five.

Chapter 6 Conclusions

6.1 Summary of Research

Reverberation has been shown by many researchers to degrade the intelligibility of speech, in particular under monaural conditions. It has also been noted that reverberation affects the naturalness of speech, lending it a "hollow" sound. The goal of this research was to develop a method of enhancement which could be applied to speech recorded or transduced with one microphone in reverberant conditions. A particular goal was that the enhancement method would be applicable when specific knowledge of the acoustical impulse response was unavailable. A second goal was that the enhancement technique would be effective over the the range of delays with respect to direct path speech commonly encountered in typical rooms, namely from zero to several hundred milliseconds.

In view of the results of previous approaches to speech dereverberation, these were ambitious goals. In general, previous two-microphone techniques had had some success in the absence of specific impulse response knowledge, but no such one-microphone techniques applicable to all ranges of delays were known. However, deconvolution using complex cepstral techniques had been proposed by Schafer in the

1960's as a method for the removal of simple echoes from speech. These techniques require only one microphone and no *apriori* knowledge of the echoes. It was thus natural to investigate whether advances in computing power would make possible the application of cepstral deconvolution to the more general problem of enhancement of speech recorded in reverberant rooms.

When we applied the techniques of cepstral deconvolution as described in [27] to reverberant speech, the results were unsatisfactory. The primary reason for this failure was that by segmenting the reverberant speech into finite length blocks for cepstral processing, one of the fundamental assumptions upon which cepstral deconvolution is predicated was violated. That is, the transformation of time domain signals into the complex cepstral domain is a process which maps convolution into addition. The cepstra of convolved time domain components appear as additive sequences in the quefrency domain; deconvolution can then proceed if these additive components are located in different areas of the cepstral (quefrency) axis. Reverberant speech can be expressed as a convolution of "clean" speech with an impulse response, but the segmented speech can not be expressed as a convolution of a segment of clean speech with that impulse response. The result is that the calculation of the cepstrum, and in particular of the impulse response cepstrum which we seek to identify and remove, is distorted. The deconvolution accordingly fails to enhance the speech.

We attempted to ameliorate these problems by several steps. The first was to develop a segmentation and windowing strategy which would allow us to preserve convolution between the segments of clean speech and the reverberation impulse response as accurately as possible. We found that a window function other than a rectangular window was necessary in order to introduce taper at segment boundaries

and therefore reduce the impact of truncation on the cepstral calculation. However, commonly used windows such as Hamming windows which achieve taper also have the effect of distorting the convolutional combination of signals. We found that the cepstral distortion resulting from these windows was as severe as that resulting from the rectangular window function. We then evaluated the exponential window, which had been suggested by Schafer for its property of conversion of mixed phase to minimum phase sequences. We found that a direct application of cepstral processing with exponential windows was also unsatisfactory. Although the exponential window achieved taper at the segment finish, and although the exponential window applied to convolved sequences indeed preserves their convolutional combination, the truncation error at the segment start boundary was found to cause a large distortion.

We then found that by judiciously choosing the segment start boundaries, this source of distortion could be largely eliminated. This was accomplished by defining segment starts to begin only after speech silence intervals. In this manner, the echo of the original speech of the previous segment, which is the source of the truncation error at the segment start, was greatly reduced in amplitude. Thus, by combining exponential weighting with this segment definition, convolutional combination between speech and impulse response was largely preserved.

Exponential weighting is also useful because z-plane zeroes of the weighted sequence may be moved inwards. In our technique, we assumed that all of the zeroes of the reverberation impulse were moved inside of the unit circle by the exponential weighting. By thus converting the impulse responses to minimum phase sequences, we could bypass the computationally expensive step of phase unwrapping in the cepstral calculation, and also avoid ambiguities in the time alignment of the impulse response

estimate which would occur with mixed phase sequences. However, this assumption led to difficulties in cases where the z-plane zeroes were not moved sufficiently far to accomplish the conversion to minimum phase.

The second major step in our proposed technique was to average the cepstra of a number of segments of reverberant speech together. We found that the cepstrum of the impulse response could be identified much more accurately with such averaging under the condition that the impulse response remain constant over the averaging interval. In this way, the variability of the speech cepstrum and its tendency to average towards zero, which we had noted in experiment, could be exploited.

Using these techniques in combination, we found that a relatively accurate estimate of the reverberant impulse response could be made. From such an estimate, we proposed to design a linear inverse filter to be applied to the reverberant speech. The cepstral identification and the filtering steps would thus be de-coupled and the cepstral processing delays would not affect the delay between reverberant and enhanced speech. We felt, therefore, that the proposed technique had potential for real-time operation.

We investigated the available techniques for designing linear filters for the removal of mixed phase impulse responses. We found that few choices were available, but that the least squared error design criterion could be used to design a filter of acceptable performance. We made the empirical observation that given perfect knowledge of an impulse response generated from an acoustical simulation program, a perceptually acceptable filter could be designed with the filter length on the order of the length of the impulse response.

6.1.1 Evaluation of Results

The speech enhancement techniques described above were evaluated using a digitized sentence of speech subject to convolution with various synthetic impulse responses. We found that when the impulse response was composed of one or more well defined discrete echoes of arbitrary pattern, and was minimum phase, that a very substantial enhancement of the reverberant speech was possible. We found that this enhancement was much more effective than the direct application of the techniques described in [27]. This we attribute to the improvements gained by segment start selection and by cepstral averaging.

Furthermore, we found that when the impulse response was generated from a program which simulates the acoustical impulse response of a room [17], that substantial reduction in reverberation could be effected. In some cases, this was achieved at the cost of distortion. We found this distortion was caused by two factors.

First, we found that when a particularly heavy degree of exponential weighting was applied, an increase in distortion ensued. We attributed this distortion to the fact that a heavy value of exponential weighting effectively shortens the length of the analysis window, because the window function falls off more rapidly. Components near the beginning of the window thus become magnified in relative importance as other components are multiplied by very small values. Any segmentation error remaining at the segment start therefore causes distortion; such segment start error is always present unless the previous segment speech is truly silent. These findings forced us to choose relatively light exponential weighting which falls off gradually, and long analysis windows to allow for segment taper. In cases where the impulse

response was minimum phase or close to it, this had no adverse impact. In cases of impulse responses with z-plane zeroes far outside the unit circle, distortion resulted from calculating a mixed phase sequence without the phase unwrapping step. Even in the latter case, substantial reduction in reverberation was noted, although at the price of a tone-like distortion in the enhanced speech.

Second, we found that residual speech cepstrum which remained after the cepstral averaging process contributed to distortion introduced by our techniques. We noted that the level of speech cepstrum, and in particular the speech cepstrum clustered about the quefrency origin, decreased with averaging. However, we found when averaging on the order of ten speech segments that such residual speech cepstrum was still significant in comparison to the level of cepstrum due to some of the impulse responses. We found that the presence of residual speech cepstrum produced a low, fairly constant level error in the reverberation impulse response estimate. Best enhancement was therefore achieved by truncation of the estimate at a point before the true level of the impulse response had fallen below the error level from the speech cepstrum. We have found no rigorous method to make this decision.

It should be noted that these studies were performed on speech free of additive noise. We do not know what the effect of additive noise will be upon the enhancement technique.

6.2 Directions for Future Study

We discuss here several directions of research which may lead to a wider range of application for the technique which we have developed. We have, as discussed above, identified two sources of distortion with our method. First, we found that

application of heavy exponential weighting to the speech segments amplifies residual segmentation error. A possible solution to this problem would be to adopt a *closed-loop* filtering structure. That is, the cepstral identification step, which of course includes the exponential weighting, would be performed *after* the linear filtering stage rather than before. Rather than designing the entire filter in each "iteration" as is the case with our open-loop approach, we could view a closed loop approach as *correcting* a previous filter estimate. The advantage of this approach would be that the exponential weighting step is performed upon the enhanced rather than reverberant speech. Assuming that a previous filter estimate is approximately correct, most of the energy of the reverberation and hence most of the segmentation error would be removed in the enhanced speech. This approach may therefore permit a heavier value of exponential weighting to be applied.

Second, we point out the need for further study and understanding of the level of speech cepstrum to be expected after cepstral averaging. Such residual speech cepstrum appears to be a limiting factor in the accuracy of this technique.

Furthermore, more rigorous techniques for the choice of exponential weighting factor, and for the optimal truncation length of the impulse response estimate, are required. Finally, we have noted that few results are available in the literature from which one may design the best *perceptual* linear filter for enhancement given knowledge of a mixed phase impulse response. We have no reason to think that the least-squared-error criterion is the perceptually optimum design.

6.3 Conclusions

To summarize, we feel that the techniques which we have developed achieve very

significant enhancement of speech subject to reverberation by discrete echoes. We furthermore feel that this method has significant promise in the application to speech enhancement degraded by acoustical room reverberation. However, further study is required in order to refine the cepstral estimation process. We feel that the greatest potential for enhancement of reverberant speech is in cases of severe reverberation, where the ratio of direct to reverberant speech is low. Such severe reverberation tends to be associated with heavily mixed phase impulse responses, for which our technique breaks down. However, we feel that by techniques similar to those described in the previous section, such problems may be overcome.

References

1. D. O'Shaughnessy, *Speech Communication*, Addison-Wesley, 1987.
2. A. Oppenheim, R. Schafer, *Digital Signal Processing*, Prentice-Hall, 1975.
3. D. Berkley, O. Mitchell, "Seeking the ideal in hands-free telephony", *Bell Laboratories RECORD*, Nov., 1974.
4. H. Hass, "The influence of a single echo on the audibility of speech", *Journal of the Audio Engineering Society*, Vol. 20, No. 2, 1972.
5. A. Nabelek, J. Pickett, "Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners", *Journal of Speech and Hearing Research*, Vol. 17, No. 4, 1974.
6. A. Koenig, J. Allen, D. Berkley, T. Curtis, "Determination of masking-level differences in a reverberant environment", *Journal of the Acoustic Society of America*, Vol. 61, No. 5, 1977.
7. H. Kurtovic, "The influence of reflected sound upon speech intelligibility", *Acustica*, Vol. 33, No. 1, 1975.
8. J. Lochner, J. Burger, "The intelligibility of speech under reverberant conditions", *Acustica*, Vol. 11, No. 4, 1961.
9. F. Santon, "Numerical prediction of echograms and of the intelligibility of speech in rooms", *Journal of the Acoustic Society of America*, Vol. 59, No. 6, 1976.
10. S. Gelfand, S. Silman, "Effects of small room reverberation upon the recognition of some consonant features", *Journal of the Acoustic Society of America*, Vol. 66, No. 1, 1979.
11. A. Nabelek, T. Letowski, F. Tucker, "Reverberant overlap and self-masking in consonant identification", *Journal of the Acoustic Society of America*, Vol. 86, No. 4, 1989.
12. R. Cox, G. Alexander, C. Gilmore, "Intelligibility of average talkers in typical listening environments", *Journal of the Acoustic Society of America*, Vol. 81, No. 5, 1987.
13. C. Stockbridge, T. Curtis, "Audio teleconference rooms - general guidelines" (Abstract), *Journal of the Acoustic Society of America*, Vol. 61(S), 1977.
14. S. Neely, J. Allen, "Invertibility of a room impulse response", *Journal of the Acoustic Society of America*, Vol. 66, No. 1, 1979.
15. R. Bolt, A. Macdonald, "Theory of speech masking by reverberation", *Journal of the Acoustic Society of America*, Vol. 21, No. 6, 1949.
16. T. Curtis, "Characterization of room coloration by moments of room spectral response" (Abstract), *Journal of the Acoustic Society of America*, Vol. 58(S), 1975.
17. J. Allen, D. Berkley, "Image method for efficiently simulating small-room acoustics", *Journal of the Acoustic Society of America*, Vol. 65, No. 4, 1979.

18. O. Mitchell, D. Berkley, "Reduction of long-time reverberation by a center-clipping process" (Abstract), *Journal of the Acoustic Society of America*, Vol. 47, No. 1, 1970.

19. O. Mitchell, G. Yates, T. Bateman, "Reduction of long-time reverberation by center clipping" (Abstract), *Journal of the Acoustic Society of America*, Vol. 58(S), 1975.

20. J. Mourjopoulos, J. Hammond, "Modelling and enhancement of reverberant speech using an envelope convolution model", *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, Boston, 1983.

21. J. Mourjopoulos, P. Clarkson, J. Hammond, "Dereverberation of speech using optimum control", *Digital Signal Processing - 84*, V. Cappelini and A. Constantinides, Eds., Elsevier Science Publishers, 1984.

22. H. Wang, F. Itakura, "Recovering of reverberated speech using a narrow band envelope estimation method", *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, 1989.

23. J. Flanagan, R. Lummis, "Signal processing to reduce multipath distortion in small rooms", *Journal of the Acoustic Society of America*, Vol. 47, No. 6, 1970.

24. J. Allen, D. Berkley, J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals", *Journal of the Acoustic Society of America*, Vol. 62, No. 4, 1977.

25. P. Bloom, "Evaluation of a dereverberation process by normal and impaired listeners", *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, Denver, 1980.

26. Y. Ephraim, D. Malah, "Adaptive speech signal dereverberation", *Proc. of IEEE Convention in Israel*, 1980.

27. R. Schafer, "Echo removal by discrete generalized linear filtering", *MIT Technical Report #466*, 1969.

28. B. Bogert, M. Healy, J. Tukey, "The quefrency alanysis of time series for echoes: cepstrum, psuedo-autocovariance, cross-cepstrum and saphe cracking", *Time Series Analysis*, M. Rosenblatt, Ed., Wiley, 1963.

29. L. Rabiner, R. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

30. A. Oppenheim, R. Schafer, "Homomorphic analysis of speech", *IEEE Trans. Audio and Electroacoustics*, Vol. AU-16, No. 2, 1968.

31. A. Noll, "Cepstrum pitch detection", *Journal of the Acoustic Society of America*, Vol. 41, No. 2, 1967.

32. J. Tribolet, "A new phase unwrapping algorithm", *IEEE Trans. Acoust., Speech, and Signal Processing*, Vol. ASSP-25, No. 2, 1977.

33. J. Bednar, T. Watt, "Calculating the complex cepstrum without phase unwrapping or integration", *IEEE Trans. Acoust., Speech, and Signal Processing*, Vol. ASSP-33, No. 4, 1985.

34. R. Kemerait, D. Childers, "Signal detection and extraction by cepstrum techniques", *IEEE Trans. Inf. Theory*, Vol. IT-18, No. 6, 1972.

35. J. Hassab, R. Boucher, "Analysis of signal extraction, echo detection and removal by complex cepstrum in presences of distortion and noise", *Journal of Sound and Vibration*, Vol. 40, No. 3, 1975.

36. D. Childers, D. Skinner, R. Kemerait, "The cepstrum: A guide to processing", *Proceedings of the IEEE*, Vol. 65, No. 10, 1977.

37. J. Tribolet, T. Quatieri, "Computation of the complex cepstrum", in *Programs for Digital Signal Processing*, IEEE Press, 1979.

38. T. Ulrych, "Application of homomorphic deconvolution to seismology", *Geophysics*, Vol. 36, No. 4, 1971.

39. P. Stoffa, P. Buhl, G. Bryan, "The application of homomorphic deconvolution to shallow-water marine seismology - part 1: models", *Geophysics*, Vol. 39, No. 4, 1974.

40. J. Tribolet, T. Quatieri, A. Oppenheim, "Short-time homomorphic analysis", *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, Hartford, 1977.

41. W. Verhelst, O. Steenhaut, "On short-time cepstra of voiced speech", *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, New York, 1988.

42. M. Picheny, N. Durlach, and L. Braida, "Speaking clearly for the hard of hearing II: acoustic characteristics of clear and conversational speech", *Journal of Speech and Hearing Research*, Vol 29, No. 4, 1986.

43. B. Atal, "Automatic speaker recognition based on pitch contours", *Journal of the Acoustic Society of America*, Vol. 52, No. 6, 1972.

44. J. Mourjopoulos, "On the variation and invertibility of room impulse response functions", *Journal of Sound and Vibration*, Vol. 102, No. 2, 1985.

45. J. Jetzt, "Critical distance measurement of rooms from the sound energy spectral response", *Journal of the Acoustic Society of America*, Vol. 65, No. 5, 1979.

46. J. Mourjopoulos, P. Clarkson, J. Hammond, "A comparitive study of least-squares and homomorphic techniques for the inversion of mixed phase signals", *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, Paris, 1982.

47. J. Proakis, D. Manolakis, *Introduction to Digital Signal Processing*, Macmillan, 1988.