AN INVESTIGATION INTO GLOTTAL WAVEFORM BASED SPEECH CODING

by

Chnstopher J Bleakley, B Sc (Hons)

Submıtted ın fulfilment of the requırements for the degree
Doctor of Phılosophy

Supervısed by
Dr Ronan Scaıfe

School of Electronıc Engıneerıng
Dublın Cıty Unıversıty
Dublın 9, Ireland

September, 1995

DECLARATION

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph D is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work

Signed C J Bleakley
Candidate

ID No 91700523

Date 25/9/95

DEDICATION

To Mum and Dad

ACKNOWLEDGEMENTS

As I complete this thesis I am acutely aware that there are several people to whom I owe a debt of gratitude

Firstly, I extend sincere thanks to my supervisor, Dr Ronan Scaife His expert advice and guidance have ensured this project's completion Much thanks is also due to Dr Sean Marlow and the technical staff at the university, particularly Dave Condell Their assistance, patience and resourcefulness were of great benefit and are much appreciated

I am grateful to my colleagues in the speech lab, Billy, Jeeva and Albert, and to my friends in college, for supporting and encouraging me, especially during times of pressure

The contribution of my parents and sister, Honor, must be recorded It is said that there are two lasting bequests we can give our children - one is roots, the other wings I thank my parents for giving me both (plus some cash!) My sister's ready wit and good humour lightened those moments when problems made the task seem more arduous

Last but not least, I acknowledge the important part played by my girlfriend, Eileen Burns She encouraged me to begin this research (even completing my application form!) and her support and loyalty have helped to make all this possible She endured my ups and downs and always inspired me to press on towards my goal

The work was funded by the Programme in Advanced Technology (Telecommunications), by ESPRIT (BRA 6975) Speech Maps and by the School of Electronic Engineering, Dublin City University

TABLE OF CONTENTS

AN INVESTIGATION INTO GLOTTAL WAVEFORM BASED SPEECH CODING

C.J. BLEAKLEY

ABSTRACT

Coding of voiced speech by extraction of the glottal waveform has shown promise in improving the efficiency of speech coding systems This thesis describes an investigation into the performance of such a system

The effect of reverberation on the radiation impedance at the lips is shown to be negligible under normal conditions Also, the accuracy of the Image Method for adding artificial reverberation to anechoic speech recordings is established

A new algorithm, Pre-emphasised Maximum Likelihood Epoch Detection (PMLED), for Glottal Closure Instant detection is proposed The algorithm is tested on natural speech and is shown to be both accurate and robust

Two techniques for glottal waveform estimation, Closed Phase Inverse Filtering (CPIF) and Iterative Adaptive Inverse Filtering (IAIF), are compared In tandem with an LF model fitting procedure, both techniques display a high degree of accuracy However, IAIF is found to be slightly more robust.

Based on these results, a Glottal Excited Linear Predictive (GELP) coding system for voiced speech is proposed and tested Using a differential LF parameter quantisation scheme, the system achieves speech quality similar to that of U S Federal Standard 1016 CELP at a lower mean bit rate while incurring no extra delay

CHAPTER 1

INTRODUCTION

1.1 AIMS OF THE THESIS

Glottal waveform processing has shown promise in increasing the efficiency of speech coders [Hedelin, 1984, 1986], improving the naturalness of speech synthesisers [Rosenberg, 1971, Holmes, 1973] and increasing the accuracy of recognition systems [Blomberg, 1991] This thesis aims to investigate the performance of existing techniques and proposes new methods for glottal waveform processing In particular, the thesis focuses on the application of glottal processing techniques to the problem of low bit rate speech coding

To this end, four main studies have been undertaken Firstly, models for representing the effects of reverberation on the recorded speech signal are investigated This study is made with special reference to the impact of reverberation on the accuracy of glottal waveform estimation by inverse filtering Secondly, an improved method for identification of the Glottal Closure Instant (GCI) from the speech signal is proposed and tested Thirdly, two algorithms for glottal estimation, Closed Phase Inverse Filtering (CPIF) [Berouti, 1976, Wong et al , 1979, Deller, 1981] and Iterative Adaptive Inverse Filtering [Alku, 1992a,b,c], are evaluated Fourthly, a Glottal waveform Excited Linear Prediction (GELP) coding system for voiced speech is developed The performance of the system, in comparison with that of three standard conventional coders, is assessed experimentally

It is believed that the research described herein will aid system developers in algorithm selection and will give direction to future research in the area Although the thesis focuses on techniques for speech coding, the results of the investigation are, nevertheless, of relevance to the related areas of speech synthesis and recognition

1 2 PLAN OF THE THESIS

The thesis contains eight chapters and three appendices Broadly speaking, Chapters two and three provide background material concerning speech coding and glottal processing, Chapters four through seven present original research into glottal processing and Chapter eight concludes the thesis

Chapter 2 Background Theory

This chapter presents an overview of the principles of speech coding It is intended to provide a context for the research described in the remainder of the thesis Three main topics are covered - the physiology of the human speech production system, models for the system and techniques for speech coding

The anatomy of the speech production system is described briefly and the physiological processes involved in generating speech are explained Modelling the system is discussed with special reference to the source-filter theory and the Linear Prediction vocal tract model Current speech coding techniques are detailed, together with the standards based upon them Lastly, the state of the art in speech coding is assessed, focusing on methods for low bit rate transmission

Chapter 3 Focal Theory

Theory directly concerned with glottal waveform based speech coding is dealt with in Chapter 3 The chapter covers five areas - the fundamentals of glottal processing, methods for glottal waveform estimation, models for the voice source, techniques for determining the GCI and a summary of glottal processing applications

The principles and assumptions underlying glottal processing are examined The two main approaches to glottal waveform estimation, inverse filtering and joint source-tract estimation, are studied with reference to the literature Similarly, dynamic and flow models for the voice source are reviewed Methods for identification of the GCI are considered in three sub-sections These cover electroglottography, epoch detection and closed phase detection Lastly, the application of glottal processing techniques to the problems of speech synthesis, recognition and coding is described In particular, the literature regarding glottal based speech coding is reviewed in detail

Chapter 4 Reverberation Modelling

Chapter 4 studies models for representing the effects of reverberation on speech recordings made in a typical room Two effects are considered - the variation of the radiation impedance at the lips and the inclusion of sound reflections in the signal received at the microphone

The variation of the lip radiation impedance is studied by developing theory for predicting the effect of reverberation on the radiation impedance at a vibrating piston set in an infinite baffle The theory is confirmed by comparing the results of Monte Carlo simulations with measurements of the radiation impedance variation at a loudspeaker The verified theory is applied to the problem of predicting the lip radiation impedance variation which occurs in normal enclosures

Reverberant speech material can be generated by convolving anechoic speech signals with typical room impulse responses One convenient method for creating these responses is the Image Method of Allen and Berkley [Allen and Berkley, 1979] The accuracy of the Image Method is studied by comparing artificially generated room impulse responses with measured responses The Image Method is employed throughout the remainder of the thesis for the generation of reverberant speech material

Chapter 5 Glottal Closure Detection

A new and improved method for GCI identification is proposed in Chapter 5 The inadequacies of the previous method, Maximum Likelihood Epoch Detection (MLED) [Cheng and O'Shaughnessy, 1989], are illustrated and the increased reliability of the new system, Pre-emphasised Maximum Likelihood Epoch Detection (PMLED), is established The performance of the new technique is evaluated across a variety of speech material, both male and female, and assessed under conditions of noise and reverberation

Chapter 6 Glottal Waveform Extraction

The performance of two existing algorithms for glottal waveform estimation, CPIF and IAIF, is evaluated in Chapter 6. The two methods are tested in conjunction with an LF model [Fant et al., 1985] fitting procedure for glottal waveform parameterisation. The accuracy and robustness of the techniques are assessed in experiments involving the processing of voiced speech recorded by subjects of both sexes under noisy and reverberant conditions. The LF parameters extracted in this way are compared with the results of previously published studies and are discussed in this context.

Chapter 7 Glottal Excited Speech Coding

Chapter 7 proposes a GELP coding system and evaluates its performance compared with that of three standard conventional coders.

The GELP system models voiced speech as an LF glottal waveform excitation applied to a Linear Prediction vocal tract filter. The GELP encoder detects the GCI by the new PMLED method and estimates the glottal excitation during voiced speech by inverse filtering. The LF model is fitted to the estimated waveform and the LP synthesis filter is determined by an ARX procedure [Astrom and Eykhoff, 1971]. Two configurations of the coder are tested - one using CPIF and the other employing IAIF. A variable rate quantisation scheme is developed whereby the LF parameters are encoded differentially on a period-by-period basis.

The speech quality, bit rate and robustness provided by the GELP system is compared empirically to those of LPC-10, CELP and GSM. The test data used in the investigation consists of continuous all-voiced male and female speech with added white noise and reverberation. The speech quality provided by the systems is evaluated using an objective quality measure, the Bark Spectral Distortion (BSD) [Wang et al., 1992]. Lastly, the performance of the GELP system is discussed and contrasted to that of the standard techniques.

Chapter 8 Conclusion

A brief summary of the thesis is given in Chapter 8. Following this, the contributions made by the investigation are assessed and the thesis concludes with a number of suggestions for future research.

Appendices

Three appendices are included at the end of the document. The first two are copies of papers which were written by the author and published during the course of this investigation. Although they do not directly pertain to the main thrust of the thesis, they are relevant and have been included for this reason. The third appendix concerns the capture of the speech test data used in the investigation.

Appendix A reprints the paper "The variation of the lip radiation impedance in a reverberant enclosure" which was originally published in the *Proceedings of the European Signal Processing Conference (EUSIPCO) 1994*, vol. 3, pp. 1689-1692.

Appendix B reprints the paper "New formulae for predicting the accuracy of acoustical measurements using the averaged *m*-sequence correlation technique" which was originally published in the *Journal of the Acoustical Society of America*, 1995, vol. 97, no. 2, pp. 1329-1332.

Appendix C describes the recording of the speech material and supplies an explanation of how the noisy and reverberant test data was created

CHAPTER 2

BACKGROUND THEORY

2 1 INTRODUCTION

This chapter presents an overview of the principles and techniques of digital speech coding Three main elements are considered - the anatomy and physiology of the human speech production system, digital models for that system and techniques for speech coding

The anatomy of the human speech production system is described, together with an explanation of the acoustic processes involved in speech production In particular, the two main types of excitation, voiced and unvoiced, are detailed

In this context, the source-filter theory of speech production, which underlies most of today's speech processing systems, is presented The main discrete-time models for the human speech production system are covered Special attention is paid to the lossless tube and Linear Prediction vocal tract models As will be seen, the Linear Prediction model is the foundation of the vast majority of low rate speech coding systems

Finally, the main methods of speech coding are presented, together with the standards based upon them The present state of the art in speech coding is assessed, especially those coding techniques currently under investigation by the speech research community

The chapter is divided into five sections - this introduction, a description of the physiology of the human speech production system, a description of the acoustic theory of speech production, an overview of speech coding and a conclusion

2.2 PHYSIOLOGY OF THE SPEECH PRODUCTION SYSTEM

For the development of efficient speech processing systems, an understanding of the human speech production system is essential The means by which we generate speech sounds is determined by the anatomy of our vocal mechanisms In turn, the nature of the sounds generated is determined by the acoustic processes within those mechanisms To explain these items clearly, this section is divided into four sub-sections The first describes the anatomy of the speech production system, the second describes how speech is actually produced by the system, the third covers the various types of excitations used in producing speech sounds and the fourth sub-section looks in detail at the voiced excitation

For further information see [Fant, 1970, Flanagan, 1972, Ladefoged, 1975, Rabiner and Schafer, 1978, O'Shaughnessy, 1987, Deller et al , 1993]

2.2.1 Anatomy

By definition, the speech signal is an acoustic sound pressure wave that is generated by the movements of the anatomical structures which make up the human speech production system

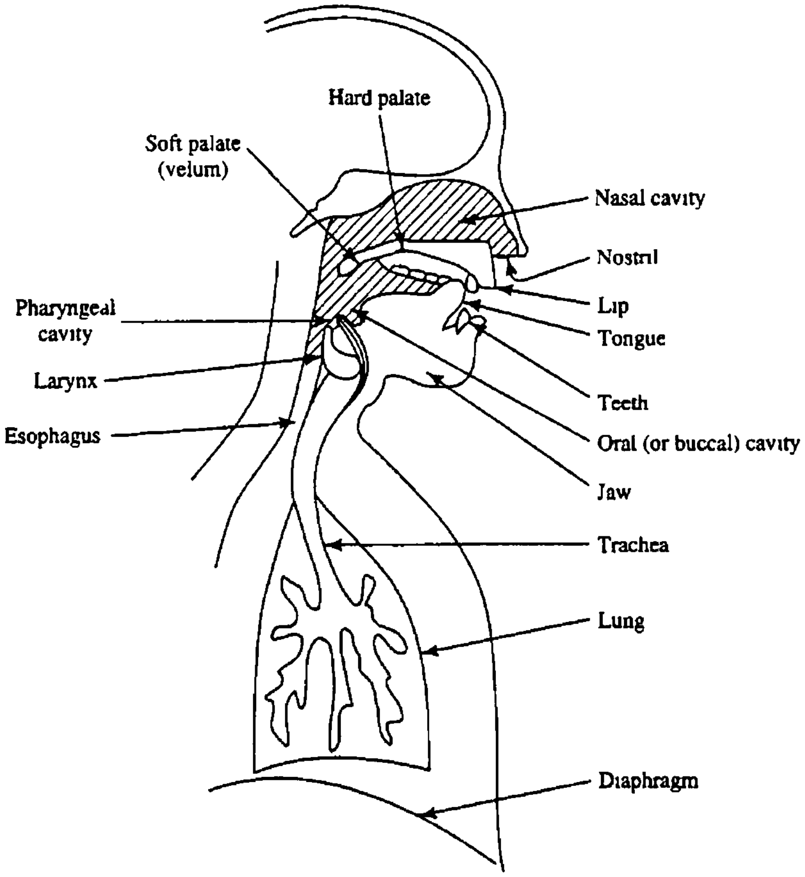


Fig 2 1 Schematic diagram of the human speech production system [after Deller et al, 1993]

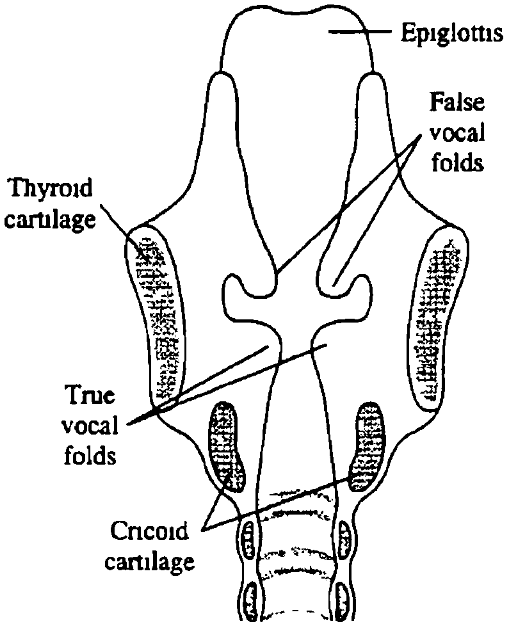


Fig 2 2 Cross-section of the larynx as viewed from the front [after O'Shaughnessy, 1987]

Fig 2 1 shows the anatomy of the human vocal mechanism The organs of the system, that is, the larynx, vocal tract and nasal tract, generate the speech signal by modulating air flow from the lungs The largest organ of the system, the vocal tract, has an average length of about 17 cm in an adult male and roughly 14 cm in an adult female Movements of the vocal apparatus or articulators (e g the lips, tongue, jaw, etc) effect changes in the shape of the vocal tract, the cross-sectional area of which can vary between zero (closure) and 20 cm^2 or more The cross-sectional area of the nasal tract is fixed, with a length of roughly 12 cm in an adult male Acoustic coupling between the vocal and nasal tracts is controlled by the velum The velum can be opened or closed to permit or prevent sound propagation through the nasal tract.

A cross-section of the larynx is shown in Fig 2 2 The function of the larynx is to provide a periodic or voiced excitation to the rest of the speech production system It achieves this by repeatedly opening and closing the vocal folds, alternately permitting and preventing air flow into the vocal tract The folds themselves are a pair of elastic bands of muscle and mucous membrane that stretch over the trachea or windpipe from the thyroid cartilage in the front to the arytenoid cartilage at the back The thyroid cartilage can be seen at the front of the neck and is commonly known as the Adam's apple The cartilage of the larynx is held together by a network of ligaments and membranes that control the positioning of the vocal folds during speech This positioning determines the setting of the vocal folds, i e open, closed or vibrating, and the mode of their vibration, i e frequency and timing

2.2.2 Speech Production

To produce speech, air flow from the lungs is converted into an excitation signal This signal excites the acoustic resonances of the vocal tract cavity and, if the velum is open, those of the nasal cavity The nature of the vocal tract resonances is determined by the shape of the cavity Thus, the resonances can be controlled by the positioning of the articulators The resonances or formants are perceived by the listener as high concentrations of acoustic energy at certain frequencies It is by the pattern of these formants that the listener determines the phonetic content of the utterance So, for example, the vowel [i] of the word "he" can be distinguished from the vowel [æ] of the word "had" because their formant patterns are different In turn, the formant patterns for the two vowels differ because they are generated using dissimilar vocal tract configurations The [i] sound is produced by a high tongue position at the front of the mouth, while the [æ] sound is generated using a low tongue position Note that the International Phonetic Alphabet (IPA) symbols used herein represent a speaker of British English

The nasal cavity, if acoustically coupled to the vocal tract, introduces anti-resonances into the speech signal This is perceived by the listener as nasalisation and leads to a specific class of speech sounds, known as the nasals (e g [m] and [n])

In summary, the speaker provides linguistic information to the listener by controlling the excitation to, and the position of, the articulators of the speech production system

2.2 3 Excitation Types

There are two fundamental excitation types, voiced and unvoiced, and three lesser types, mixed, whisper and plosive

Voiced sounds are produced by forcing air through the larynx or glottis The tension in the vocal cords is adjusted so that oscillations are set up and maintained This periodic interruption of the air flow from the lungs results in quasi-periodic puffs of air that excite the vocal tract This process is known as phonation and occurs during all vowels and some consonants (e g [w], [l] and [m])

Unvoiced sounds are formed by forcing air through a constriction in the vocal tract This causes turbulence in the air flow, which is perceived by the listener as a "hissing" or noisy sound Unvoiced excitation is used for all the fricatives (e g [s] and [f])

The two fundamental excitations can be combined to produce a mixed excitation This involves simultaneous use of the voiced and unvoiced excitations, leading to a class of sounds known as the voiced fricatives (e g [z])

A whisper is created by forcing air through a partially open glottis to excite an otherwise normally articulated utterance The air flow constriction at the glottis creates a turbulent or noisy excitation which replaces the normal voiced excitation to give a glottal fricative A related form of voicing is the breathy voice type This occurs during voiced speech when the glottis undergoes incomplete closure Thus, a hissing excitation is produced during what would normally be the closed phase

Plosive sounds are produced by closing the vocal tract completely, allowing air pressure to build up behind the closure and suddenly releasing it Plosives can be further categorised based on whether the sound following them is voiced or unvoiced For example, the sound [t] is an unvoiced plosive, while [b] from "boot" is a voiced plosive

2.2.4 Voice Production

The myoelastic/aerodynamic theory of phonation describes how the voiced excitation is generated [van den Berg et al , 1957, van den Berg, 1958] The process is illustrated by the sequence of laryngeal cross-sections in Fig 2 3 Initially, the vocal cords are closed and air pressure builds up below the larynx due to the contraction of the lungs (Fig 2 3 (a)) This pressure forces the cords apart (Fig 2 3 (b) and (c)) and air flows through the slit-like opening (Fig 2 3 (d)) Due to the glottal constriction, the air flow has a large velocity As a result of the Bernoulli effect, a negative pressure is generated This force coupled with the elastic tension of the cords, pulls the folds back together again (Fig 2 3 (e)) The glottis is again closed (Fig 2 3 (f)) and, as before, air pressure builds up below the vocal folds (Fig 2 3 (a)) In this way, vocal fold vibration is set-up and maintained The air flow increases and decreases periodically in sympathy with the opening and closing vocal folds

The rate and timing of vocal fold vibration is controlled by the tension and spreading of the vocal cords plus the air pressure in the lungs The fundamental frequency of the vibration lies, typically in the range 50-250 Hz for an adult male and 120-500 Hz for an adult female [Deller et al , 1993] The timing of the vibration affects the overall shape of the glottal volume velocity waveform This includes the skew

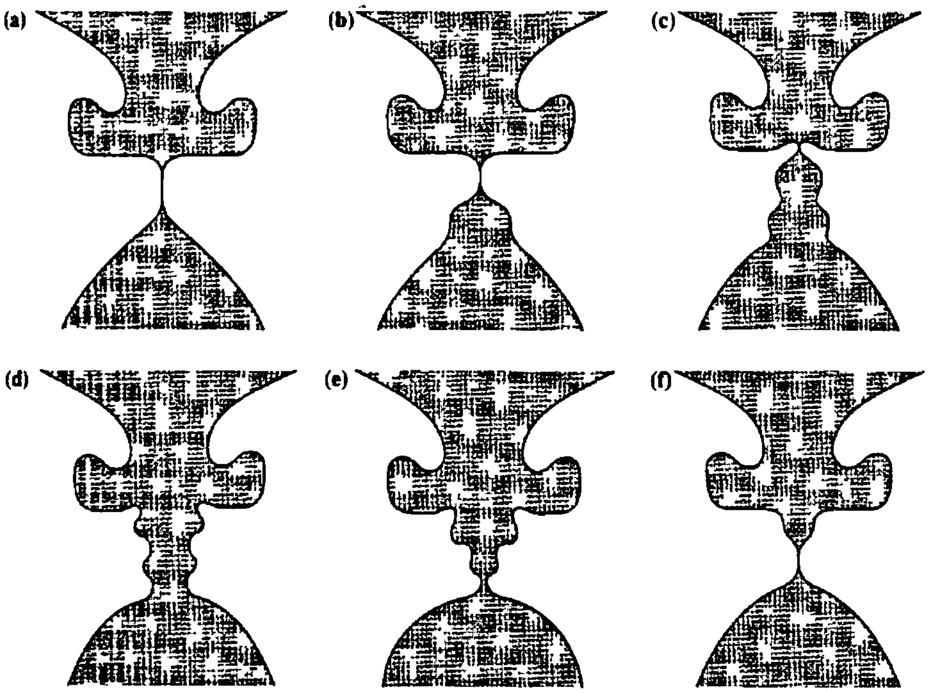


Fig 2 3 Sequence of cross-sections of the larynx illustrating a complete glottal cycle [after Deller et al, 1993]

of the pulse, the rate of glottal opening, the rate of closure, the maximum flow and the ratio of the glottal closed phase to open phase In combination, these factors control the excitation waveform and provide the listener with information on the speaker's meaning, identity and emotional state [Cummings and Clements, 1990, 1992, Childers and Lee, 1991]

2.3 MODELLING SPEECH PRODUCTION

For effective machine processing of speech, mathematical models of the speech production systems must be developed These models attempt to represent the speech production process in an accurate and efficient manner allowing for automatic synthesis, recognition and coding

The section is split into four sub-sections In the first, the basic acoustic theory of speech production, which is the cornerstone of most speech processing systems, is described In the remaining three sub-sections, the main components of most speech production models are detailed These are the excitation, vocal tract and lip radiation models

2 3 1 Acoustic Theory of Speech Production

As can be seen from the previous section, the human speech production mechanism is a complex system incorporating a large number of component parts To fully represent such a system would require a large set of equations describing the physical process of air propagation within the vocal mechanism [Sondhi, 1974] Such a universal theory would require the characterisation of such elements as the time-varying vocal tract shape, the time-varying vocal folds, nasal coupling, subglottal coupling, viscous friction, heat conduction, wall loss and low-viscosity compressible fluid mechanics Such a theory has not yet emerged

The most commonly used approximation to the speech production system is the so-called source-filter theory [Fant, 1970] In this, the excitation signal, the vocal tract filter (incorporating the effects of nasalisation) and the lip radiation function are considered to be separable linear systems which are short-time invariant. Thus, in the z-domain, the speech pressure signal $S(z)$ can be calculated from the volume velocity excitation $G(z)$, the vocal tract filter $H(z)$ and the lip radiation $R(z)$

$$S(z) = G(z)H(z)R(z) \tag{2 1}$$

The model assumes that there is no coupling between the sub-systems and that there is planar sound propagation within the vocal tract Neither of these assumptions is, in fact, valid for the real speech production process [Teager and Teager, 1990] However, the simplifications which they allow have facilitated the development of computationally feasible techniques for speech modelling, coding and synthesis In general, systems built on the source-filter theory have shown good performance in everyday applications

2.3.2 Excitation Modelling

The two fundamental types of excitation, voiced and unvoiced, are generally represented using models of the excitation waveforms

In the case of the voiced excitation, the quasi-periodic glottal signal has been represented by very simple and very complex models The simplest model is a train of impulses at the fundamental period of phonation [Tremain, 1982] More accurate models, such as the LF model, parameterise the glottal waveshape and reproduce the timing details of the glottal excitation as well as the pitch period [Rosenberg, 1971, Fant et al , 1985] Still more complex models attempt to simulate the movement of the vocal cords and the fluid flow through them, often including vocal tract and subglottal coupling effects [Ishizaka and Flanagan, 1972, Ananthapadmanabha and Fant, 1982, Titze, 1989] Studies in speech perception suggest that accurate modelling of voiced speech is crucial for natural sounding coding and synthesis systems [Borden and Harris, 1980]

The turbulent unvoiced excitation is most commonly represented by a white noise source with controllable gain Since the human ear is insensitive to the details of the unvoiced excitation, this simple model is effective for synthesising unvoiced speech

The remaining excitations, mixed, plosive and whisper, are frequently ignored in the speech production model When they are included, they are normally represented by a combination of the voiced and unvoiced excitations

2.3 3 Vocal Tract Modelling

One of the most intuitive vocal tract representations is the lossless tube model [Chiba and Kajiyama, 1941, Dunn, 1950, Stevens and House, 1955, 1961, Fant, 1970, Lindbolm and Sundberg 1971] In this, the vocal tract is considered as a series of concatenated lossless acoustic tubes, see Fig 2 4 The cross-sectional areas of these tubes are chosen so as to approximate the cross-section of the real vocal tract Assuming that one-dimensional planar propagation of sound occurs within the tract, an

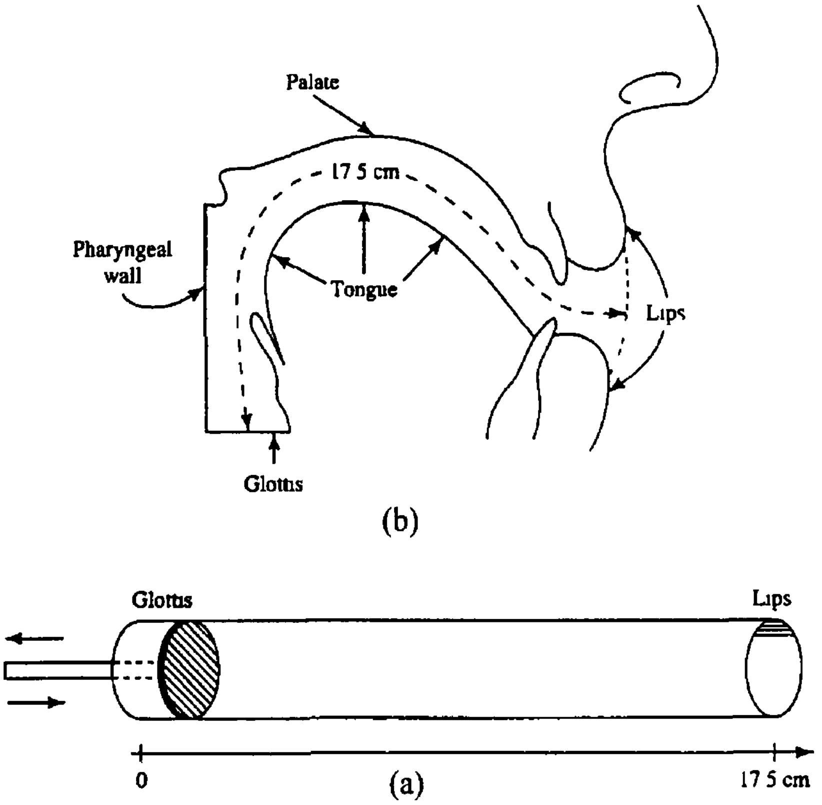


Fig 2 4 (a) Straightened cross-section of the vocal tract, (b) lossless tube vocal tract model

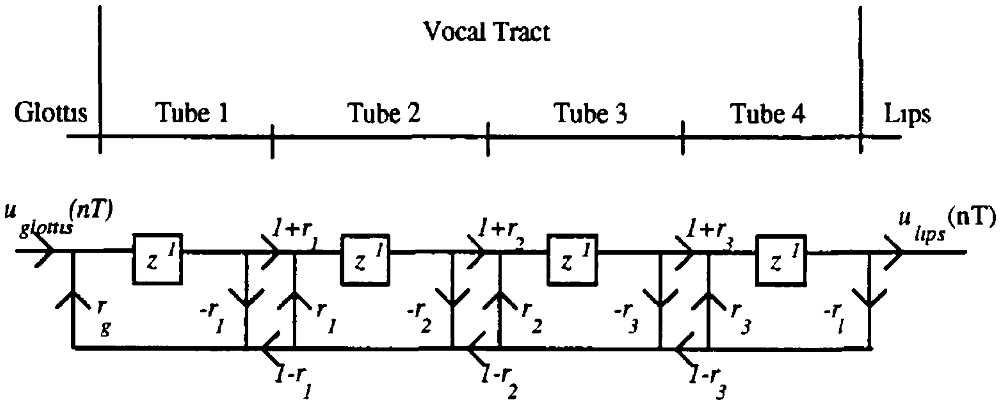


Fig 2 5 Discrete time signal flow vocal tract model assumption which is reasonable for low frequencies (< 4 kHz), then the movement of air within the tubes is governed by the Law of Continuity of Mass and Newton's Force Law. Using these relations, the pressure and volume velocity in each of the tubes depends only on the pressure and volume velocity of the air in the neighbouring tubes and on the ratio of the cross-sectional area of the tube to that of its neighbours. In this way, the pressure and flow in each of the tubes can be calculated as a function of time. Thus, the sound pressure radiated at the lips can be determined by applying a glottal volume velocity signal to the lossless tube model and calculating the air flow from the glottis to the lips. The flow at the lips is radiated as a sound pressure wave.

The basic lossless tube model has shown good results in reproducing human vowels from vocal tract area functions measured in X-ray pictures [Fant, 1970; Flanagan, 1972]. However, various modifications have been made to the basic model in order to improve the quality of the speech obtained from it. These modifications have included vibration loss, thermal loss, conduction loss and an extra cavity to facilitate the production of nasal sounds.

The basic lossless tube model has proven extremely suitable for implementation in a discrete-time system [Kelly and Lochbaum, 1962]. To achieve this, the vocal tract can be divided into a number of tubes of a common fixed length. Analysis of wave propagation proceeds as before, except that each tube incurs the same time delay. Using this simplification, the overall lossless tube can be represented using a discrete-time signal flow model as in Fig. 2.5. Sound takes half a sampling period z^{-1} to traverse each tube and the flow between tubes k and k+1 is determined by the reflection coefficient of the junction r_k. The reflection coefficient is simply calculated from the cross-sectional area of the tubes A_k and A_{k+1}.

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \tag{2.2}$$

For simplicity, and without loss of accuracy, the half sample delays in the feedback path can be moved up into the forward path. The resulting transfer function for a two tube vocal tract model is then of the form

$$\begin{aligned} H_{2\text{-tube}}(z) &= \frac{U_{lips}(z)}{U_{glottis}(z)} \\ &= \frac{\left[(1+r_g)/2\right](1+r_1)(1+r_l)z^{-1}}{1+(r_1 r_g + r_1 r_l)z^{-1} + r_g r_l z^{-2}} \end{aligned} \tag{2.3}$$

Removing the overall delay of the system and generalising the coefficients, the N-section lossless model has the form

$$H(z) = \frac{H_o}{1 - \sum_{k=1}^{N} a_k z^{-k}} \tag{2.4}$$

where H_o is the gain and a_k are the coefficients of the system. This expression corresponds to the transfer function of an all-pole filter. The poles of $H(z)$ define the resonant or formant structure of the vocal tract. In general, an eight section model is required to represent speech at a sampling frequency of 8 kHz. This corresponds to an assignment of two poles per formant.

This Linear Prediction model of the vocal tract has proven extremely useful in speech processing [Markel and Gray, 1976]. It is the basis of most low rate speech coding systems, has been used in many speech synthesis applications and is a common representation in speech recognition systems [Rabiner and Schafer, 1978, Deller et al , 1993] ,The assumptions and approximations intrinsic to the model are more than made up for by its flexibility and computational tractability

2.3 4 Lip Radiation Modelling

The volume flow signal at the lips is radiated in the form of an acoustic pressure wave At low frequencies, the radiating area of the mouth can be assumed to have a velocity distribution that is uniform and co-phasic [Flanagan, 1972] Therefore, the radiating area is roughly equivalent to a vibrating piston set in a baffle corresponding to the head

The most accurate representation of the lip radiation function is the piston in a spherical baffle model [Morse and Ingard, 1962] Unfortunately, the mathematical expression for this function is complex and cannot be expressed in closed form More commonly used are the first terms of the series expansion for the radiation impedance of a piston in an infinite baffle [Flanagan, 1972]

The gross effect of the radiation function is to apply a +6 dB/octave emphasis to the flow signal at the lips In the digital domain, this can be represented by a first order differentiation

$$R(z) = 1 - a_0 z^{-1} \qquad (2\ 5)$$

where $a_0 \approx$ 0 9 Although more complex z-domain models, in which the radiation function depends on the lip area do exist [Laine, 1982], the simple differentiation model is by far the most common

Frequently in studies of the glottal excitation, the differentiation effects of the lip radiation function are applied directly to the glottal volume velocity signal [Fant et al , 1985] This entails representing the glottal excitation by the differentiated volume velocity and removing the radiation function altogether In this study, the differentiated glottal volume velocity signal is referred to as the glottal excitation or glottal waveform

2.4 SPEECH CODING

Due to world-wide demand for advanced telecommunication systems, speech coding remains an area of intense research activity [Jayant, 1990, Gersho, 1994, Rabiner, 1994] The continuing goal of this research is to transmit high quality speech at a low bit rate

Coding systems improve the efficiency, and hence reduce the cost, of speech transmission and storage However, coding incurs extra costs, in terms of the once-off purchase of the encoding and decoding units The two main applications for speech coding are voice messaging, where disk space must be conserved, and cellular telephony, where bandwidth must be conserved Today's commercial coding systems range from low complexity, high bit rate waveform coders to high complexity, low bit rate vocoders Some of the most common systems are described below

Virtually all speech coding systems are lossy, that is, the decoded speech waveform is not the same as the encoded waveform The key to efficient coding is to preserve only the sound information which is perceptually important to the listener As well as this, coders are optimised so as to reproduce high quality speech Other sounds, such as music, need not be so well represented Nevertheless the coding process should under no circumstances introduce annoying artefacts

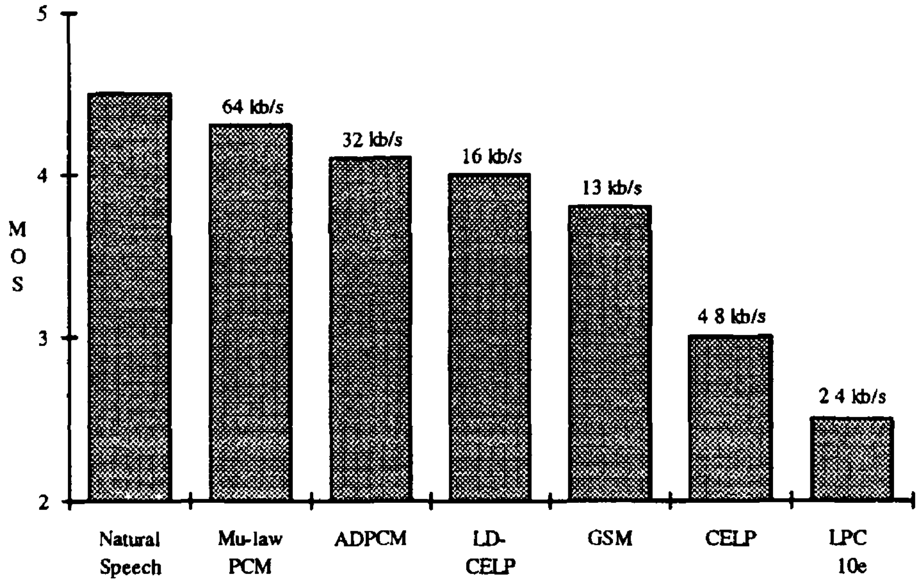


Fig 2 6 Histogram of Mean Opinion Scores achieved by standard speech coding systems [after Rabiner, 1994]

Speech coding systems generally operate at a sampling frequency of 8 kHz This allows the reproduction of speech within the telephone bandwidth of 200 Hz - 3 4 kHz This bandwidth contains the first four formants and is satisfactory for the transmission of speech

Systems are normally assessed in terms of their transmission rate and speech quality Simple objective quality measures, such as the Signal to Noise Ratio (SNR), show poor performance in predicting the quality of the coded speech [Quackenbush et al , 1988] For coding purposes, it is not important that the re-synthesised waveform matches the original signal What is important is that the re-synthesised speech sounds like the original To this end, subjective quality measures such as the Mean Opinion Score (MOS), are more effective in assessing the quality of a coding system The Mean Opinion Score is a subjective evaluation of speech quality based on listening tests [IEEE, 1969] In these tests, subjects are asked to rate coder performance as 1 (bad), 2 (poor), 3 (fair), 4 (good) or 5 (excellent) The mean rating is taken as the MOS The MOS of some of the coding systems described in this section are shown in Fig 2 6

Recent, more complex objective quality measures, such as the Bark Spectral Distortion [Wang et al , 1992], which model the properties of the human auditory system, have shown promise in predicting the quality of speech coders However, these measures have not yet been widely used due to the lack of standards Doubtless, an accurate standard for objective quality assessment will eventually emerge

2.4.1 Waveform Coders

Waveform coders transmit speech data by encoding the speech waveform on a sample-by-sample basis Although numerous waveform coding schemes have been proposed, the two most common approaches are Pulse Code Modulation (PCM) and Adaptive Differential Pulse Code Modulation (ADPCM)

The simplest of these systems is PCM In this, the amplitude of the speech waveform is sampled at 8 kHz and the level is quantised to an 8 bit integer Non-uniform quantisation (or companding) is commonly used to improve the quality of PCM The Telecommunications Standardisation Sector of the International Telecommunication Union (ITU-T) defined a PCM standard in 1972 [Jayant, 1990] This standard provides high quality speech 4 3 MOS at a transmission rate of 64 kb/s

Adaptive Differential PCM achieves quality comparable to that of PCM at a lower rate by exploiting some of the redundancies in the speech signal ADPCM uses a Linear Prediction filter to predict the next sample from previous samples The coefficients of the Linear Predictor are calculated by the autocorrelation method applied over a block of samples The autocorrelation method determines the Linear Prediction coefficients which give the minimum mean squared error in predicting the speech signal from previous samples The coefficients obtained are quantised and transmitted to the receiver, together with the difference between the actual signal and the predicted signal The overall transmission rate of the system is lower than PCM because, relative to the raw speech signal, the difference signal has a reduced dynamic range and so requires fewer bits for transmission The ITU-T standard for ADPCM is G 721 [Jayant, 1990] Set in 1984, it has been rated as providing a MOS of 4 1 at 32 kb/s

2.4.2 Vocoders

Vocoders, or voice coders, attempt to characterise speech in terms of a speech production model This requires frame-by-frame analysis of the speech signal and extraction of the parameters of the model In general, low rate vocoders ignore the details of the speech waveform and reproduce the perceptually important short-term spectral magnitude information The most common forms of vocoder in use today utilise a Linear Prediction vocal tract filter and a waveform excitation model, as shown in Fig 2 7 Four of the major coder types are described below - LPC-10, Residual Excited Linear Prediction (RELP), Multipulse LPC (MP-LPC), Regular Pulse Excitation (RPE) and Code Excited Linear Prediction (CELP)

LPC-10 is based on an extremely simplistic model of the speech production system The system makes a decision as to whether a frame of speech is voiced or unvoiced In the case of voiced speech, an impulse train at the pitch period is used to excite a 10th order Linear Prediction filter This filter represents not only the vocal tract, but also the spectral contribution of the glottal excitation For unvoiced speech, a white noise excitation is applied to a 4th order filter The system operates at a very low bit rate but produces poor quality speech The system is particularly susceptible to voicing decision and pitch errors These errors cause considerable distortion of the speech signal and lead, in part, to the poor subjective quality rating for the system In addition, during voiced regions the re-synthesised speech has a "buzzy" or synthetic quality due, in part, to the use of an impulse excitation The U S Federal Standard 1015 LPC-10e developed in the mid-70s, achieves a MOS of 2 5 at 2 4 kb/s [Tremain, 1982]

Multipulse LPC [Atal and Remde, 1982] introduced two key features into speech coding systems - analysis-by-synthesis and perceptually weighted error measures The basic MP-LPC system uses an LP filter, calculated by conventional methods over a frame of speech samples A number of candidate multipulse excitations are passed through the LP filter to re-synthesise the speech signal The difference

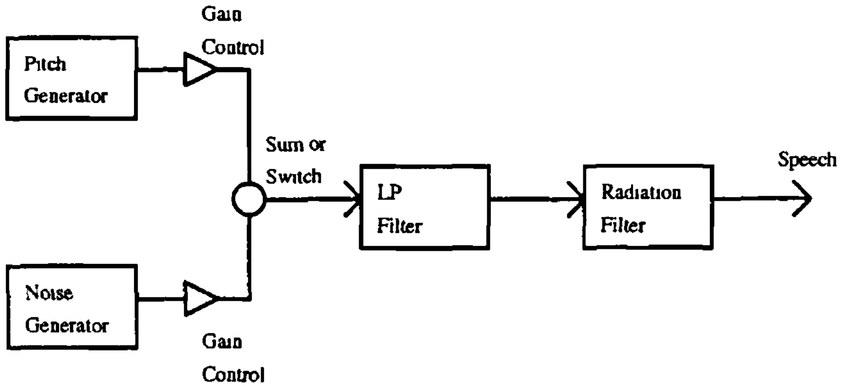


Fig 2 7 Generalised schematic diagram of a Linear Predictive speech coding system [after Rabiner and Schafer, 1978]

between the re-synthesised and original signals is calculated and applied to a perceptual weighting filter This filter de-emphasises errors at frequencies which are present in the speech signal and emphasises errors at frequencies which are lacking in the speech signal Thus, errors which would be heard by a human listener are emphasised and those masked from a listener by the speech signal itself are de-emphasised The energy of the perceptually weighted error signal is minimised by iteratively adjusting the multipulse excitation The overall process of selecting the best excitation by re-synthesising the speech signal and comparing it to the original is referred to as analysis-by-synthesis The analysis-by-synthesis approach avoids the need for hard decisions, such as voicing or pitch, and ensures good quality speech under most conditions The multipulse excitation is normally a sparse sequence of pulses separated by zeros This leads to a low overall transmission rate for the system An MP-LPC algorithm at 9 6 kb/s was recently adopted as a standard for aviation satellite communications [Gersho, 1994]

Inspired by MP-LPC, Regular Pulse Excitation coding uses regularly spaced pulse patterns for the excitation to a conventional LP filter The RPE sub-system operates in combination with a Long Term Prediction sub-system which removes redundancy in the speech signal due to the slowly changing pitch period RPE-LTP-LP was selected in 1988 as the standard for digital cellular telephony by the Global System for Mobile telecommunications (GSM) sub-committee of the European Telecommunications Standards Institute (ETSI) [ETSI, 1989] The system achieves a MOS of 3 8 at a transmission rate of 13 kb/s [Rabiner, 1994]

Currently, the most important form of vocoding system is CELP This approach uses a conventional LP synthesis filter excited by innovation sequences from a stochastic and an adaptive codebook. The stochastic codebook contains sparse random pulse sequences, while the adaptive codebook holds time lagged versions of previous excitations Each codebook is exhaustively searched to find the two entries which minimise the perceptual error between the re-synthesised speech and the original The optimum codebook indices and gains, together with the quantised LP coefficients, are transmitted to the receiver The search procedure is highly computationally expensive - a CELP encoder requires roughly 30 MIPS [Rabiner, 1994] CELP systems achieve good quality speech at a low bit rate Standardised CELP coders include the U S Federal Standard 1016, dating from 1989, which attains a quality rating of 3 0 MOS at 4 8 kb/s [Campbell et al , 1991]

The principles of CELP have been further developed to produce ITU-T standard G 728, Low Delay CELP [ITU, 1993] Standardised in 1991, the system achieves high quality speech MOS 4 0 at a medium rate 16 kb/s with a coding delay comparable to that of ADPCM The system uses only a stochastic codebook, and achieves a low transmission rate by backward adaption of the gain and LP synthesis filter As in conventional CELP, an exhaustive codebook search with a perceptually weighted error criterion is employed This makes LD-CELP extremely computationally complex - a LD-CELP encoder needs approximately 50 MIPS [Rabiner, 1994]

Note that the GSM standard and the U S Federal Standards 1015 LPC-10e and 1016 CELP are described in more detail in Section 7 4 1

2 4 3 State of the Art

In recent years, developments in speech coding have been driven by two main factors Firstly, the availability of low cost, high speed Digital Signal Processors (DSPs) has allowed the implementation of increasingly complex coding algorithms Secondly, improved knowledge of the human speech production and auditory mechanisms has allowed the removal of further redundancy from the speech signal Essentially, systems no longer allocate bandwidth to sounds which cannot be produced by the speaker or which cannot be heard by the listener The current state of the art can be summarised by examining the performance achievable at a given transmission rate

The quality of a good connection in the Plain Old Telephone System (POTS), i e toll quality, can now be achieved at 16 kb/s with LD-CELP G 728 The coder offers low delay and high quality speech

Speech coding at around 8 kb/s is currently under standardisation To this end, CELP type systems are under investigation for the half-rate GSM, North American half-rate digital cellular and ITU-T standards [Gersho, 1994]

At 4-6 kb/s, the best CELP algorithms introduce noticeable coding noise, although intelligibility, naturalness and identifiability of the speaker's voice are retained The quality at this rate is often referred to as digital cellular

At 2-3 kb/s the performance of CELP is further degraded, the speech quality becoming noisy and "hoarse" The quality at these rates is described as communications quality, that is, the speech is intelligible but distorted Research is currently under way to develop algorithms which provide better quality at these rates Amongst the most promising algorithms are Sinusoidal, Mixed-Excitation, Prototype Waveform Interpolation (PWI) and Glottal Excited Linear Prediction (GELP)

Sinusoidal coders operate by parameterising the short-term spectrum of the speech signal [Hedelin, 1981, Almeida and Tribolet, 1982, Marques et al , 1990, McAulay and Quatieri, 1986, 1992, Brandstein et al , 1990, Nishiguchi et al , 1993] In particular, voiced speech is modelled as a sum of sinusoids whose frequencies and phases are controlled so as to track the evolving short-term spectra of the speech Sinusoidal systems suffer from some of the analysis errors typical of LPC-10 but, in general, provide cleaner speech than CELP at low rates However, Sinusoid systems incur a long coding delay

Mixed-Excitation coders are based on the LPC-10 system, but replace the binary voicing decision with a mixed impulse and white noise excitation [McCree and Barnwell 1992 1993] Separate voicing decisions are made for each sub-band of the speech signal This reduces the severity of voicing errors and removes the buzzy quality of standard LPC-10 systems. Subjective tests indicate that the quality of Mixed-Excitation systems at 2.4 kb/s approaches that of U.S. Federal Standard 1016 CELP for clean speech, and exceeds it for noisy speech.

Prototype Waveform Interpolation coders model the voiced excitation to an LP synthesis filter by transmitting a single prototype pitch cycle every 20-30 ms and reconstructing the signal by interpolation [Kleijn, 1991; Kleijn and Ganzow, 1991]. The interpolation can be done in the time or spectral domains with differential coding of the prototypes. Conventional CELP coding is used for unvoiced speech. An implementation of PWI has been reported to achieve an impressive quality compared with conventional schemes at 2.4-4 kb/s [Shoham, 1993a,b].

Glottal Excited Linear Prediction coders operate by extracting and parameterising the glottal excitation during voiced speech [Hedelin, 1984, 1986; Bergstrom and Hedelin, 1988, 1989; Alku and Laine, 1989a; Alku, 1990a,b, 1991]. The glottal excitation is generally represented by a time-domain waveform model, the parameters of which are determined by fitting the model to the glottal excitation estimated from the speech signal by inverse filtering. The parameters of the glottal waveform are slowly time-varying compared to those of a conventional LP residual model. Thus, the transmission rate required for the excitation is lower in GELP systems than in conventional coders. Also, in GELP systems the LP synthesis filter is equivalent to the actual vocal tract filter and so can be quantised very efficiently. Standard LPC-10 or CELP techniques are used for transmission of unvoiced speech. GELP systems have been shown to produce high quality speech at low rates. However, they are susceptible to phase distortion in the incoming speech signal which makes extraction of the time-domain glottal waveform parameters difficult.

At rates below 1 kb/s, speech coders operate on large segments of speech and so incur delays of hundreds of milliseconds [Liu, 1989, 1990, 1991; Kemp et al., 1991]. The systems range from barely intelligible to communications quality.

2.5 CONCLUSION

This chapter has described the basic theory underlying today's speech coding systems. The human speech production system has been explained in terms of its anatomical structures and the acoustic processes involved in the generation of speech. Various discrete-time models for the speech production system have been detailed. The lossless tube model for the vocal tract has been described, together with waveform excitation and lip radiation models. Additionally, the derivation of the Linear Prediction vocal tract model from the lossless tube model has been presented. Current speech coding standards, both waveform and vocoding, have been described in some detail. Finally, the state of the art in speech coding has been assessed. Current research is focused on low rate, medium delay systems which can achieve digital cellular quality. One such system, Glottal Excited Linear Prediction, has been chosen as the subject of this investigation. The next chapter describes the history and structure of GELP systems.

CHAPTER 3

FOCAL THEORY

3 1 INTRODUCTION

This chapter presents a survey of research that has been conducted in the area of glottal waveform processing The principles underlying glottal processing are considered, as well as techniques for glottal waveform estimation and modelling Methods for identifying the Glottal Closure Instant (GCI) from the speech signal are described In addition, systems employing glottal processing techniques for speech recognition, synthesis and coding are detailed

The chapter is organised as follows Section two covers the principles and assumptions underlying glottal processing Section three describes techniques that have been developed for estimating the glottal waveform from voiced speech The fourth section examines the various models which have been proposed for representing the voice source The associated problem of GCI detection is covered in the fifth section Section six describes how glottal processing techniques have been applied in the areas of speech recognition, synthesis and coding Section seven concludes the chapter

3.2 FUNDAMENTALS OF GLOTTAL WAVEFORM PROCESSING

Glottal waveform processing is based on the source-filter theory of speech production [Fant, 1970] This theory supposes that the human speech production system consists of three separable linear sub-systems - a glottal excitation, a vocal tract filter and a lip radiation function Although this representation of the speech production system has been used extensively and fruitfully in speech processing, the assumptions underlying it are incorrect Linearity requires planar airflow within the vocal tract. Direct measurements of flow within the tract suggest that this does not always occur [Teager and Teager, 1990] Separability requires that the glottal waveform is unaffected by the vocal tract configuration Measurements show that the glottal flow waveform is skewed, relative to the glottal opening area, due to the vocal tract load [Rothenberg, 1973, Rothenberg and Zahorian, 1977, Ananthapadmanabha and Fant, 1982] In addition, coupling between the vocal tract and the subglottal system increases formant damping during the glottal open phase, particularly for the first formant [Lindqvist, 1964, Fant, 1979, Krishnamurthy, 1992] This manifests itself as a formant ripple superimposed on the open phase of the glottal flow waveform derived by inverse filtering [Ananthapadmanabha and Fant, 1982, Fant, 1986, Krishnamurthy and Childers 1986] The assumptions of linearity and separability are however effective in simplifying the speech production model and facilitate the use of computationally efficient algorithms

Assuming that the source-filter theory is valid, the problem of glottal estimation reduces to separating the effects of the glottal excitation from those of the lip radiation function and the vocal tract filter Generally, the lip radiation function is considered to approximate a simple first order differentiation As such, its effects can be easily cancelled by an integration step or they can be incorporated into the glottal model Hence, the differentiated glottal volume velocity is often used in place of the glottal volume velocity Unfortunately, the influence of the glottal excitation is difficult to separate from the effects of the vocal tract filter The vocal tract filter is highly variable, for example formant frequencies change dramatically according to the vocal tract configuration Also, the filter is time-varying, due to the movement of the articulators and due to subglottal coupling During non-nasal speech, an all-pole Linear Prediction filter is effective in modelling the tract [Makhoul, 1975] However, during nasal speech, a pole-zero filter is more appropriate [Steiglitz and Dickinson, 1977, Atal and Schroeder, 1978, Fujisaki and Ljungqvist, 1987, Lobo and Ainsworth, 1992] In most systems an all-pole model is used, regardless of whether the segment is nasal or non-nasal

As well as the problem of source-tract deconvolution, careful consideration must be given to the recording channel [Holmes, 1975, Markel and Gray, 1976] Since the glottal waveform contains significant low frequency components, the recording must be of high quality Also, since a composite waveform is to be extracted, the signal must not be phase distorted In general, glottal extraction requires the use of FM or digital recording equipment together with phase linear microphones and anti-aliasing filters Although a number of techniques have been proposed to correct phase distortions in the recording process [Veeneman and BeMent, 1985, Hedelin, 1988], these factors remain obstacles to the widespread application of glottal processing techniques

3.3 GLOTTAL WAVEFORM ESTIMATION

Clinical inspection of the larynx has provided a great deal of information on the physiology of the voice source Methods, such as stroboscopy [Hertegård and Gauffin, 1995], high-speed cinematography [Flanagan, 1958], ultrasound [Hamlet and Reid, 1972] and photoglottography [Hanson et al , 1990], have all played an important part in extending our knowledge of phonation Also, techniques for making acoustic measurements, both within the vocal tract [Cranen and Boves, 1988] and at the lips [Rothenberg, 1970, 1973, Sondhi, 1975], have provided information on the airflow through the vocal cords Unfortunately, these methods, while very accurate, are invasive and are unsuitable for everyday speech processing applications What is required is an algorithm that provides accurate glottal waveform estimation from the speech signal alone Furthermore, the algorithm must be robust to noise and distortion

Algorithms for estimating the glottal waveform from the speech signal fall into two main categories - inverse filtering algorithms and joint source-tract estimation algorithms Inverse filtering algorithms attempt to retrieve the glottal excitation by estimating the vocal tract filter and applying its inverse to the speech signal Joint source-tract estimation algorithms attempt to determine the glottal excitation and the vocal tract filter by matching re-synthesised speech to the original

This section consists of two sub-sections The first describes inverse filtering algorithms and the second details methods for joint source-tract estimation

3 3.1 Inverse Fıltering

Inverse filtenng ınvolves usıng the ınverse of an estımated vocal tract filter $H'(z)$ and lıp radıatıon functıon $L'(z)$ to cancel the formant structure of the speech sıgnal $S(z)$ and so obtaın an estımate of the glottal excıtatıon $G'(z)$

$$G'(z)=\frac{S(z)}{H'(z)L'(z)} \tag{3 1}$$

An example of the ınverse filtenng process ıs shown ın Fıg 3 1

The earlıest ınverse filtenng systems consısted of an electncal filter network whıch was manually adjusted to cancel the vocal tract resonances [Mıller, 1957, Mathews et al , 1961, Holmes, 1962, Lındqvıst, 1965] The operator tuned the ınverse filter to produce mınımum formant npple dunng the closed phase of the glottal waveform estımate Later studıes used Dıgıtal Sıgnal Processıng technıques, ıncludıng automatıc formant trackıng, to expedıte the process [de Veth et al , 1989, Knshnamurthy, 1992] Whıle the manual methods produce good results, they are slow and are restncted to clear modal voıcıng dunng whıch the formants are clearly defined Obvıously, the technıque ıs very labour ıntensıve, whıch may be satısfactory for basıc research, but ıs unsuıtable for most speech processıng applıcatıons

The earlıest method for automatıc vocal tract filter estımatıon was Closed Phase Inverse Fıltenng (CPIF) [Berouti, 1976, Berouti et al , 1977, Wong et al , 1979, Hunt et al , 1978] Thıs technıque assumes that the speech sıgnal observed dunng the glottal closed phase ıs due solely to the freely decayıng vocal tract resonances Thus, LP analysıs performed over the closed phase should ıdentıfy the vocal tract filter alone, excludıng any glottal excıtatıon Thıs vocal tract filter estımate can then be used to recover the glottal excıtatıon by ınverse filtenng of the speech sıgnal

Unfortunately, closed phase analysıs may not provıde accurate vocal tract transfer functıon estımates for several reasons Fırstly, nasal couplıng may ıntroduce zeros ınto the vocal tract transfer functıon Thıs ıs dıfficult to account for usıng normal all-pole LP analysıs Secondly, the glottıs may not close completely Thus, excıtatıon wıll occur dunng the "closed" phase and, as a result, the vocal tract estımate wıll be ınaccurate [Hunt et al , 1978, Larar et al , 1985] Thırdly, even ıf the glottıs closes completely, there ıs evıdence to suggest that some excıtatıon of the vocal tract often occurs due to vertıcal motıon of the vocal folds [Holmes, 1976, Cranen and Boves, 1988] In the case of modal non-nasal voıced speech these effects are generally assumed to be neglıgıble A fourth problem wıth CPIF ıs that the closed phase may be too short to allow accurate LP analysıs Thıs ıs partıcularly evıdent for hıgh pıtched female voıces sınce denvıng LP filters from ıntervals shorter than 1 5 ms gıves erratıc results A number of remedıal methods have been demonstrated whereby successıve glottal cycles can be combıned to allow more accurate LP analysıs [Fans and Tımothy, 1974, Chan and Brookes, 1989, Lu et al , 1990] The fifth and final problem wıth CPIF ıs that ıt ıs frequently dıfficult to automatıcally ıdentıfy the closed phase Inaccurate closed phase ıdentıficatıon leads to the ınclusıon of glottal effects ın the vocal tract filter and causes poor glottal waveform estımatıon

Regardless of these problems, CPIF has produced good qualıty results ın a number of ınvestıgatıons In partıcular, Knshnamurthy and Chılders have reported that closed phase covanance analysıs provıdes very accurate formant trackıng [Knshnamurthy and Chılders, 1986] Also, Veeneman and BeMent have noted that CPIF can provıde relıable glottal waveform estımates for both normal and

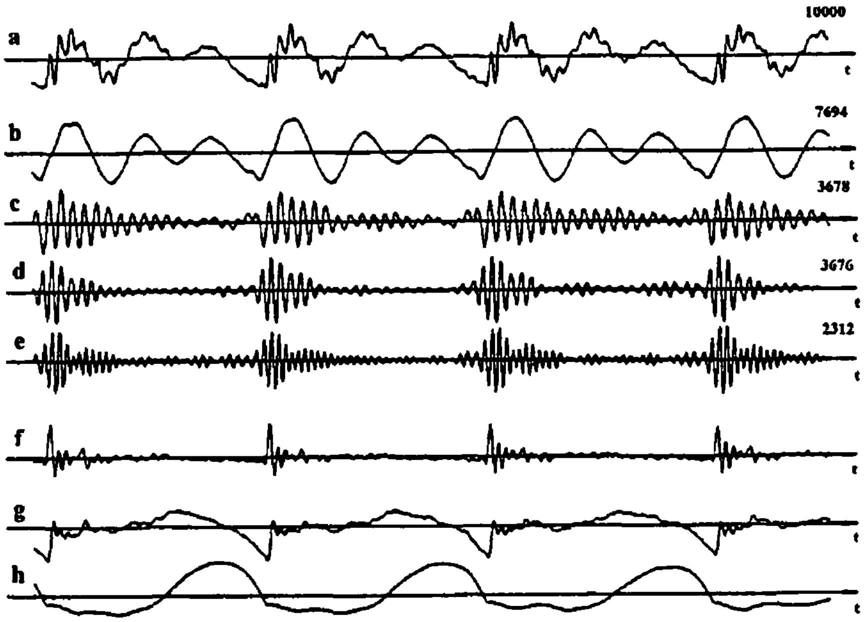


Fig 3 1 Inverse filtering analysis (a) speech signal, male vowel [e], (b) waveform of formant F1, (c) waveform of formant F2, (d) waveform of formant F3, (e) waveform of formant F4, (f) double differentiated glottal volume velocity, (g) differentiated glottal volume velocity, (h) glottal volume velocity [after Hess, 1983]

pathological speakers [Veeneman and BeMent, 1985] However, it must be noted that both investigations used high quality speech recordings and an electroglottograph for precise identification of the closed phase

Another approach to inverse filtering is to attribute certain spectral characteristics of the speech signal to the glottal excitation Once identified, these characteristics can be removed from the speech signal, leaving behind an estimate of the vocal tract filter The earliest example of this kind of technique was published by Miller and Mathews [Miller and Mathews, 1963] They attributed the zeros of the speech spectrum to the glottal excitation and attributed the poles to the vocal tract resonances Although they published some interesting results, the method is not generally applicable as, for example, nasal coupling introduces zeros into the vocal tract filter

In more successful work, Alku has proposed Pitch Synchronous Iterative Adaptive Inverse Filtering (IAIF) [Alku, 1992b] Developed from asynchronous and non-iterative versions [Alku and Laine, 1989a,b, Alku, 1990a,b, 1991, 1992a], the technique operates by attributing the gross spectral envelope of the speech signal to the glottal excitation A low order all-pole LP analysis is used to capture the spectral envelope of the speech signal The speech signal is inverse filtered by this initial glottal estimate The vocal tract filter is then obtained by applying high order all-pole LP analysis to the inverse filtered speech This vocal tract filter estimate is used to inverse filter the original speech signal to give the first glottal waveform estimate Low order LP analysis is performed on the glottal waveform estimate and the inverse filtering process is repeated to give a second, more accurate, glottal waveform The pitch synchronous version of the algorithm carries out this procedure twice - once pitch asynchronously, to determine the glottal pulse end-points and once more, pitch synchronously, to obtain a precise glottal waveform estimate

Alku has reported that the method works well for male and female speech, both synthetic and natural The only problems he notes are in processing the vowel [i], during which the algorithm fails to fully cancel the first formant. The method is fully automatic and, in contrast to CPIF, does not require accurate *a priori* identification of the closed phase The performance of the algorithm has not been independently assessed and the robustness of the procedure is unknown A similar inverse filtering procedure has also been proposed by Benitez, Galvez, Rubio and Diaz [Benitez et al , 1992]

In recent years, several other techniques for inverse filtering have been proposed using the conventional LP residual [Matausek and Batalov, 1980], the complex cepstrum [Yegnanarayana, 1981] and Higher Order Statistics [Chen and Chi, 1993] However, manual, closed phase and spectral allocation inverse filtering remain the most effective in terms of glottal waveform estimation and formant tracking

3.3 2 Joint Source-Tract Estimation

In recent years, due to the introduction of fast and inexpensive Digital Signal Processors, computationally complex iterative optimisation algorithms have been designed and applied in all areas of speech research This approach has also been taken in the field of voice source estimation A number of algorithms for joint estimation of the glottal waveform and vocal tract filter have been proposed In general, the algorithms proceed as follows The vocal tract filter is initialised, based on, say, conventional LP analysis Inverse filtering is performed and a glottal waveform model is fitted to the output. Iterative joint optimisation of the glottal waveform model $G'(z)$ and vocal tract filter $H'(z)$ then takes place This optimisation procedure is usually based on the now familiar analysis-by-synthesis technique and often makes use of a subjective error criterion Ultimately, the glottal and vocal tract parameters which minimise the error between the re-synthesised speech $S'(z)$ and the original, are stored or transmitted and processing continues to the next frame

$$S'(z) = G'(z)H'(z)L'(z) \qquad (3\ 2)$$

One of the earliest pieces of research on this topic was carried out by Takasugi [Takasugi, 1971] Since then similar systems have been proposed by other researchers Milenkovic used a polynomial glottal waveform to excite an all-pole LP vocal tract filter [Milenkovic, 1986, Thomson, 1992] Also, some studies have been made using a glottal excitation, combined with a pole-zero vocal tract filter, for the synthesising nasals [Fujisaki and Ljungqvist, 1987, Lobo and Ainsworth, 1992] To ease the computational burden of iterative optimisation, efficient numerical methods were developed by Isaksson and Millnert [Isaksson and Millnert, 1989]

Basing their work on articulatory modelling of the speech production system, Schroeter et al [Schroeter et al , 1987] devised a system whereby the results of an acoustic analysis of the speech signal were used to look up a linked code-book of vocal tract configurations and acoustic parameters These acoustic parameters were then used to perform inverse filtering before parameter re-optimisation

Another approach to simultaneous estimation of the glottal source and vocal tract parameters has been suggested by Krishnamurthy [Krishnamurthy, 1990, 1992] He used a sum-of-exponentials representation for a glottal excited LP speech production model The method shows promise but uses a large number of parameters and requires accurate closed and open phase identification by means of an EGG

An ARX formulation of the speech production model has been used for simultaneous source-tract estimation by Cheng and O'Shaughnessy [Cheng and O'Shaughnessy, 1993] A glottal model consisting of a number of polynomial functions is used to excite an all-pole filter The parameters of the glottal model and vocal tract filter are determined by a least-mean-square method which minimises the mean error between the re-synthesised and original speech The accuracy of the glottal waveform estimation is unclear but the system is said to produce natural sounding speech

While joint estimation methods can provide high quality results, their usefulness for glottal flow determination is limited Unlike inverse filtering algorithms which estimate the glottal excitation directly, source-tract methods are limited by the accuracy of their glottal model This limitation is less of a problem for speech coding systems which aim to produce good sounding speech regardless of the accuracy of the model However, for these applications the high computational complexity of the algorithms remains a problem

3.4 GLOTTAL MODELS

Models representing the voice source can be divided into two categories - dynamic models, which capture the movement of the vocal folds, and flow models, which parameterise the air flow from the glottis into the vocal tract Models from these categories are detailed in the next two sub-sections

3 4 1 Dynamic Models

The earliest dynamic model for the voice source was proposed by Flanagan and Landgraf [Flanagan and Landgraf, 1968] This model represents the vocal cords as an acoustic-mechanical oscillator wherein each cord is described by a single sprung mass The control parameters are the subglottal lung pressure, vocal cord tension, rest opening, vocal tract shape and nasal coupling The model was later elaborated to incorporate more physiological processes, leading to the development of the two mass model of Ishizaka and Flanagan [Ishizaka and Flanagan, 1972, Flanagan et al , 1975, Lucero, 1993]

Later, a glottal model based on the contact area of the vocal folds was developed by Titze [Titze, 1984, 1989] This work dealt primarily with the modelling of the three-dimensional glottis using kinematic parameters similar to those employed in articulatory vocal tract models In recent years, computational approaches have been used to study glottal dynamics, including finite element simulation of the flow through the glottis using the Navier-Stokes compressible viscous fluid flow model [Liljencrants, 1991, Iijam et al 1992, Ni and Alipour, 1993, Guo and Scherer, 1993]

Dynamic models are much more computationally complex than flow models Ideally, however, they are more accurate since they directly model the function of the human larynx This facilitates the inclusion of source-tract interaction effects in the overall speech production model

Unfortunately, dynamic models are difficult to use in speech processing because they employ parameters which are not easily measured Although some attempts have been made to calculate the glottal area function from the glottal flow [Rothenberg and Zahorin, 1977], the movement of the vocal cords can generally only be determined by invasive methods Thus, dynamic models are unsuitable for speech coding and recognition applications They have, however, been used with some success in speech synthesis systems [Ishizaka and Flanagan, 1972, Allen and Strong, 1985, Miller et al , 1988]

3 4 2 Glottal Flow Models

Flow models attempt to represent the waveform of the glottal airflow into the vocal tract Usually, the parameters of flow models are determined by time-domain fitting to the glottal waveform derived by inverse filtering or joint source-tract estimation Flow models are defined either in terms of the glottal volume velocity or the differentiated glottal volume velocity The latter formulation simplifies the overall speech production model since it incorporates the lip radiation effects The major types of model are polynomial, cosinusoidal and impulse excited filter

Today, the most commonly used glottal waveform model is the LF model [Fant et al , 1985] The differentiated volume velocity version of the model consists of a cosinusoidal open phase with exponentially growing amplitude, followed by an exponential return phase In analysis experiments, the model has been shown to capture the main details of the glottal excitation [Jansen et al , 1991] In addition, the LF model has been successful in speech research [Gobl, 1988, Karlsson, 1988] and speech synthesis applications [Childers et al , 1987, Carlson et al , 1990, Childers and Lee, 1991] One of the advantages of the LF model is its flexibility, that is, it allows the parameterisation of a large range of glottal wave shapes In particular, the inclusion of a controllable return phase has been determined as essential for good quality re-synthesis [Fujisaki and Ljungqvist, 1986] The model only requires four independent parameters which can be specified in forms suitable for analysis or synthesis One of the main disadvantages of the model is that automatic fitting of the LF waveform requires iterative optimisation [Riegelsberger and Krishnamurthy, 1993]

The first polynomial and cosinusoidal models were developed by Rosenberg [Rosenberg, 1971] He tested listener preferences to six volume velocity models in speech synthesis experiments Since that time, a plethora of polynomial and cosinusoidal models have been proposed, see Fig 3 2 [Takasugi, 1971 Fant, 1979b, 1982, Rothenberg, 1981, Anathapadmanabha, 1982, Hedelin, 1984, Fujisaki and Ljungqvist, 1986, Price, 1989, Klatt and Klatt, 1990, Cummings and Clements, 1992, Lobo and Ainsworth, 1992]

In a development of the polynomial model, Milenkovic has suggested using the sum of a number of component polynomial functions [Milenkovic, 1986] The idea has been further developed by other authors and it appears that the best structure for the model is either the sum of four to ten low order polynomial basis functions [Thomson 1992 Milenkovic 1993, Cheng and O'Shaughnessy, 1993], or a single high order polynomial [Childers and Hu, 1994] In synthesis experiments the method has proven successful [Childers and Hu, 1994] Although fitting the model is less computationally expensive than matching the LF model, the technique requires a greater number of parameters

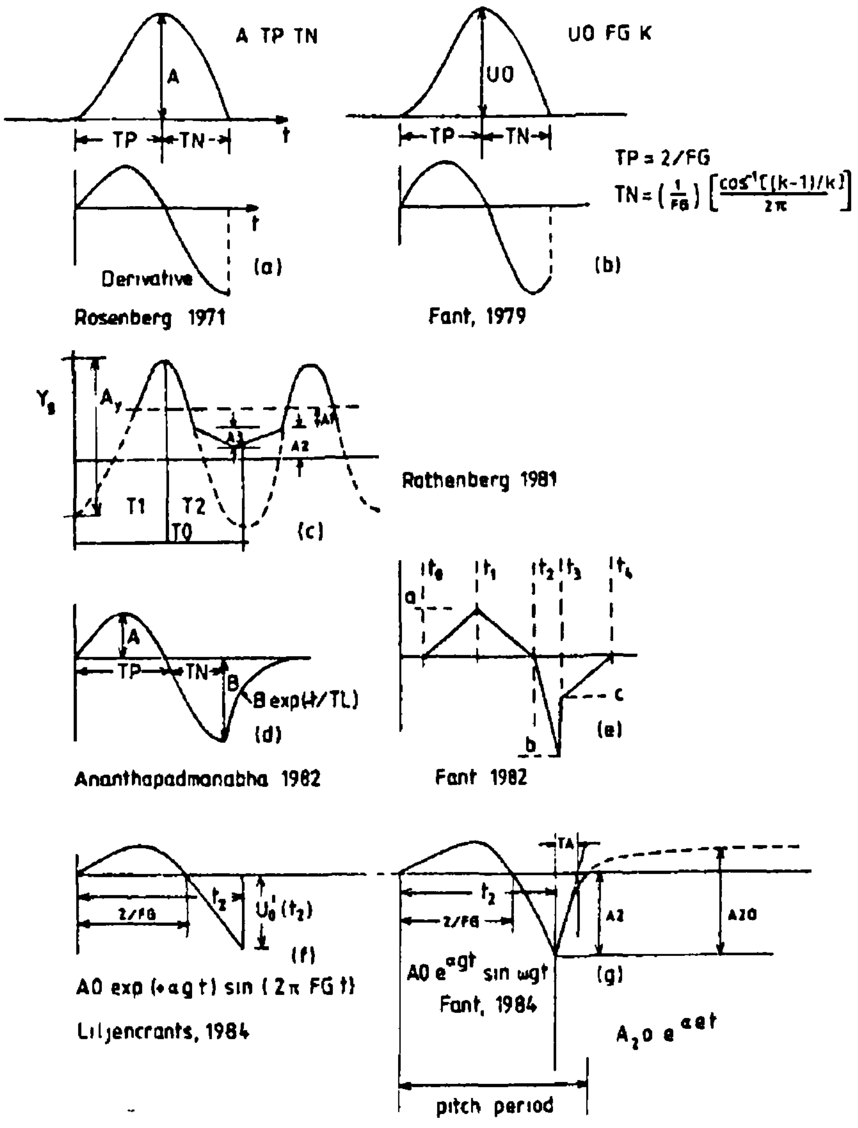


Fig 3 2 Some examples of time-domain glottal waveform models [after Ananthapadmanabha, 1984]

Another flow model has been proposed by Schoentgen [Schoentgen, 1988, 1989, 1990, 1992a,b] In a series of papers, he developed the idea of glottal waveform modelling via nonlinear Volterra shaping functions driven by a cosinusoidal signal The nonlinear shaping functions can be calculated over a small number of reference glottal cycles The long-term glottal excitation can then be generated by controlling the pitch and amplitude of the driving signal Schoentgen has reported that the method accurately tracks the output of the two-mass model and glottal waveforms estimated from natural male speech The method shows promise for coding applications due to its low update rate However, the current formulation requires a large number of parameters Furthermore, the reliability of the model in capturing the details of phonation across all voices and voicing types has yet to be established

Much less precise glottal models have been employed by a number of researchers Alku and Laine proposed the use of Lagrange interpolation between five reference points [Alku and Laine, 1989a,b]

Similarly, Benitz, Galvez, Rubio and Diaz used spline interpolation between reference points [Benitz et al , 1992] Leung et al have used a multipulse approximation to the glottal waveform [Leung et al , 1990]

An impulse excited filter model has been used in several investigations [Matau\u0161ek and Batalov, 1980, Deller, 1983, Alku, 1990b, 1991, Scordilis and Gowdy, 1990] This representation has the advantages of requiring few parameters, and being easy to implement The LP filter models the spectral magnitude contribution of the glottal excitation but loses a great deal of phase information Contrary to the commonly accepted notion that the human ear is phase deaf, a number of studies have commented on the fact that preserving the timing details of the glottal excitation is important for high quality synthesis [Flanagan, 1958, Wong and Markel, 1978, Childers et al , 1987] These filter models lose the glottal timing details and so generally provide lower quality re-synthesis than the polynomial and cosinusoidal models

Overall, glottal flow modelling by waveform matching has been the most successful means of incorporating voiced excitations in speech processing systems Currently, polynomial and cosinusoidal representations dominate In general, system designers are faced with a trade-off between the computationally expensive but low dimensional LF model and the polynomial models Alternatively, coarse glottal waveform models, such as spline interpolation or filter modelling, are computationally inexpensive and robust, but are imprecise

3.5 GLOTTAL CLOSURE DETECTION

Accurate reproduction of the pitch contour is essential for generating high quality speech Over the years, this requirement has led to the development of a large number of pitch detection algorithms [Hess, 1983] Aside from techniques which employ special apparatus, pitch detection algorithms fall into two categories - those that detect the periodicity of the speech signal and those that identify the Glottal Closure Instant (GCI) Although identification of the GCI is the more difficult approach, it has the advantage of preserving the pitch micro-melody which carries phonemic, linguistic and speaker information Also, accurate GCI identification is a necessity for automatic glottal waveform estimation by Closed Phase Inverse Filtering

This section, which is divided into three sub-sections, describes methods for detection of the GCI In the first sub-section, the most effective non-invasive technique for GCI identification electroglottography (EGG), is described Although limited to research and medical applications, the technique is important because it is the standard by which other pitch detection algorithms are assessed The second sub-section details algorithms for GCI identification from the speech signal by epoch detection The third sub-section covers algorithms for GCI identification from the speech signal by closed phase detection

3 5.1 Electroglottography

The electroglottograph (or electrolaryngograph) remains the most accurate non-intrusive method for determining the GCI Invented by Fabre [Fabre, 1957] in the late fifties, the EGG has had extensive use in speech research The device measures the electrical impedance of the glottis by feeding a weak, high frequency current between two electrodes placed on either side of the thyroid cartilage The impedance is high when the glottis is open due to the large electrical resistance of the air gap between the vocal folds Conversely, the impedance is low when the vocal folds are in contact In comparative testing the EGG has proven to give an accurate measurement of vocal fold activity [Fant et al, 1966, Fourcin and Abberton, 1971, Fourcin, 1974, 1986, Lecluse et al, 1975, Krishnamurthy and Childers, 1981, 1986, Childers et al, 1985, 1990, Hess and Indefrey, 1987 Orlikoff, 1991]

3.5.2 Epoch Detection

Epoch detection techniques attempt to find the GCI from the speech signal by identifying the abrupt change associated with closure In theory, few events other than glottal closure should excite all frequency bands coherently In practice, however, the approach is sensitive to noise and to excitations other than at the moment of closure For this reason, recent GCI epoch detection techniques have used secondary identification criteria, such as the limited period-by-period variation of the pitch and the linear predictability of the speech signal immediately following closure

GCI detection by identification of jumps in the speech energy was originally used by Smith [Smith, 1954] In his method, the signal is bandpass filtered into 20 frequency bands The signal from each bandpass filter is rectified and all the signals are summed to give an all-frequency energy estimate The instant of glottal closure is taken to be the moment when the summed waveform, i e the all-frequency energy, changes from decaying to rising The method was later employed in a channel vocoder constructed by Yaggi [Yaggi, 1962] Also, a variation on this method was used by Parthasarathy and Tufts [Parthasarathy and Tufts, 1987] This energy based approach is very sensitive to noise and secondary vocal tract excitations

Atal and Hanauer have suggested that the GCI can be identified by a peak in the prediction error after LP analysis [Atal and Hanauer, 1971] The idea is that decaying vocal tract oscillations are predictable and so can be cancelled by an LP filter, whereas the excitation at closure is unpredictable and so cannot be removed in this way Unfortunately, this method fails for certain sounds and for certain speakers Multiple peaks of either polarity can occur around the instant of closure due to the phase of the formant resonances and due to the presence of zeros in the speech spectrum In addition, the residual signal becomes very noisy when the prediction error is low, e g during voiced speech, or when frication is present Thus, the estimate is frequently inaccurate

Strube [Strube, 1974, 1980], basing his work on earlier methods devised by Sobakin [Sobakin 1972], proposed the use of the autocovariance matrix for GCI detection The method involves calculating the autocorrelation matrix of a window slid over the speech signal Strube suggested that the matrix determinant displays its short-term maximum when the start of the window coincides with the GCI He argued that, since the largest excitation occurs at the GCI, then the instant would be characterised by high energy and high prediction error, both of which lead to a high autocorrelation matrix determinant Strube reported that the method could only be considered in the case of vigorous vocal cord vibration with sharp glottal closure

Following early work by Young on radar signals [Young, 1965], Anathapadmanabha and Yegnanarayana suggested that the moment of glottal closure could be found by identifying the point of maximum discontinuity in the derivative of the speech waveform [Anathapadmanabha and Yegnanarayana, 1975] Their hypothesis is that the glottal closure excitation causes discontinuities in the speech signal, whereas freely decaying oscillations exhibit a continuous waveform This approach works best when applied to a spectrally flat signal [Larsson, 1977] The idea has been implemented in three different ways In the first implementation, the output from a bandpass filter was used [Anathapadmanabha and Yegnanarayana, 1975] In the second, a narrow bandpass filter centred around the formants was applied [de Mori et al , 1977] Both of these approaches provide results of limited resolution and require the use of clean data since only a narrow frequency band is analysed To circumvent this problem, the third implementation used a moderate bandpass filter applied to the LP residual [Anathapadmanabha and Yegnanarayana, 1979] This implementation relies on the accuracy of LP analysis which is unsatisfactory in certain cases, as was explained previously

Very good results have been reported for an epoch detection algorithm proposed by Cheng and O'Shaughnessy [Cheng and O'Shaughnessy, 1989] Maximum Likelihood Epoch Detection (MLED) is based on the principle that the speech signal following the GCI is equivalent to the impulse response of an all-pole system Conventional LP analysis is applied to the speech signal to obtain the parameters of the all-pole system A wavelet modelling the speech signal at, and immediately following, closure is obtained by exciting the all-pole system with a Dirac delta pulse The GCI is identified as the maximum of the cross-correlation between the wavelet and the speech signal Post-processing is applied to the cross-correlation signal to facilitate the voicing decision and to aid in determining the GCIs Cheng and O'Shaughnessy report that the method works well for all vowels, nasals, voiced fricatives and voiced plosives Furthermore, they state that the technique is resistant to white noise and to certain amplitude and phase distortions

A similar algorithm has since been published by Harris and Nelson [Harris and Nelson, 1993] This algorithm finds the GCIs by cross-correlating the speech signal with a time-varying adaptive filter matched to previous glottal pulses The cross-correlation is scored by a pseudo-metric which is invariant under affine transformations of the incoming signal Harris and Nelson claim that the accuracy of the algorithm is comparable to that of hand-marking However, they do not propose an automatic initialisation procedure for the adaptive filter The algorithm requires hand-marking of the first glottal pulse and so is unsuitable for many applications

A different approach to the pitch and GCI detection problem has been suggested by Dologlou and Carayannis [Dologlou and Carayannis, 1989] Their algorithm extracts the fundamental frequency of voiced speech by iteratively removing the high frequency components of the signal using a zero-phase filter with monotonically decreasing frequency response After each iteration, the results of autocorrelation and second order LP analysis are compared If only one sinusoid remains the filtering process is terminated and the remaining sinusoid is taken as the fundamental The minimum of each cycle of the fundamental is marked as the GCI In comparisons with EGG output, Dologlou and Carayannis report good results for the method, both in terms of its precision and robustness However, some problems associated with the halting criterion have been reported [Hult, 1991, Dologou and

Carayannis, 1991]. Also, the precision with which the minimum of the fundamental identifies the GCI is open to question.

3.5.3 Closed Phase Identification

Algorithms for closed phase identification are based on the assumption that little or no excitation of the vocal tract takes place during the closed phase. Thus, when the glottis is closed, the speech signal is due solely to decaying vocal tract resonances. Most closed phase identification techniques use the Linear Predictability of this region to distinguish it from the open phase. Once the closed phase has been identified, the GCI is taken as occurring immediately prior to it.

El Mallawany proposed a closed phase detection method whereby low order and high order LP analysis is carried out on a short-time window, slid one sample at a time over the speech signal [El Mallawany, 1977]. He suggested that the window was positioned over the closed phase when the largest decrease in prediction error moving from low order to high order LP analysis was observed. The procedure is highly computationally complex and sensitive to noise.

Using a similar approach, Wong, Markel and Gray proposed that the close phase could be found by applying covariance analysis to a short-time window slid over the speech signal [Wong et al, 1979]. They suggested that the close phase is identified as the window position for which the normalised prediction error is minimum. Veeneman and BeMent found that Wong's method was sufficient for normal speech but that it gives ambiguous results for high pitched or breathy speech [Veeneman and BeMent, 1985]. In addition, the method is computationally complex.

Recently, a unifying framework for the Strube and Wong methods has been developed by Ma, Kamp and Willens [Ma et al., 1994]. The methods were compared under a Singular Value Decomposition (SVD) approach and a better formulation of the methods was proposed. The new formulation involves applying a sliding window to the speech signal and calculating the arithmetic mean of the squared singular values obtained from the Frobenius norm of the window. The mean is a measure of the predictability of the speech signal within the analysis window and its local maxima coincide with the GCI. The SVD method is much less computationally expensive than the Strube and Wong techniques and has been reported to be less sensitive to noise.

Funada has designed a new algorithm for GCI identification based on a re-interpretation of the AR model [Funada, 1989]. The conventional AR model assumes a white input. Funada suggests the use of an AR model with an unknown non-white input signal (*u*-input) whose parameters are estimated by a Kalman filter. The recovered *u*-input signal captures the dynamics of the glottal excitation and can be used to determine the GCI. In tests on synthesised and natural male speech, Funada reports good results for the method. In particular, he finds it to be more successful than Wong's approach. However, he comments that the results are poor for the vowel [i] due to its low first formant.

In a recent paper, Moulines and Di Francesco [Moulines and Di Francesco, 1990] proposed and tested two new methods for closed phase identification. The first method relies on the assumption that, due to the effects of subglottal coupling, the vocal tract resonances during the closed phase are very different to those during the open phase. It is assumed that the speech signal observed during a single pitch period is best modelled by a succession of two (unknown) Gaussian AutoRegressive processes. For each possible transition instant the parameters of these models are identified. Simultaneously, a single model is identified over the same period. The likelihood ratio between the alternatives of a single process and two processes with an abrupt change occurring at the transition instant, is computed and used as a GCI indicator.

The second method proposed by Moulines and Di Francesco, is an adaptation of earlier methods for phonetic segmentation of speech [Basseville and Benveniste, 1986, Andre-Obrecht, 1988]. The technique locates glottal events by detecting jumps in the divergence between a short-term Probability Density Function (PDF) and a long-term PDF. During steady states the divergence function is convex because the PDFs are similar. In contrast, during transient regions such as at the GCI, the divergence function falls rapidly.

In experiments on normal speech both of the methods proposed by Moulines and Di Francesco were assessed as being reliable for all of the vowels. However, the algorithms performed poorly during voiced fricatives, nasals and voiced plosives.

Based on Moulines and Di Francesco's work, Murgia, Mann and Feng have suggested a similar GCI identification technique [Murgia et al., 1994]. The method uses a long-term and a short-term window applied to the speech signal. The windowed signals are analysed using two LP models and the residual probability densities are calculated. Jumps in the cumulative sum of the log-likelihood ratio is used to test the hypothesis that the short-term model is significantly different to the long-term model. This is equivalent to the hypothesis that the GCI occurs within the short-term window. Mugia, Mann and Feng report that the technique works well for vowels, voiced fricatives and voiced plosives, but has some difficulty during transients.

3.6 GLOTTAL WAVEFORM APPLICATIONS

The preceding techniques for glottal processing have been employed in a number of applications areas. The next three sub-sections survey the use of glottal waveform processing in speech recognition, synthesis and coding systems.

3.6.1 Speech Recognition

The glottal waveshape carries information on the speaker's emotional state and identity. This adds variability to the speech signal. To account for this, Blomberg has proposed a pre-processing technique whereby source spectrum adaption is carried out prior to the recognition process [Blomberg, 1991, 1993]. The adaption has proven successful, improving the accuracy of isolated word recognition from 88% to 96%. Techniques such as this are of particular benefit in recognising speech in high stress environments [Stanton et al., 1989].

Cummings and Clements have published a method for recognising emotional states based on the parameters of the estimated glottal waveform [Cummings and Clements, 1990, 1992]. The emotional state of the speaker is determined by comparing the extracted glottal parameters with those of eleven prototype styles. Methods have also been proposed for altering the voicing style, for example to change stressed speech to modal [Cummings and Clements, 1993, Mizuno and Abe, 1994].

3 6 2 Speech Synthesis

Glottal excitation models have been used to improve the naturalness of synthesis systems for some time The most common approach is to use a glottal waveform excitation with a LP or formant vocal tract model Early work was carried out by Rosenberg and Holmes, who concluded that representing the glottal pulse shape is important for synthesising natural sounding vowels [Rosenberg, 1971, Holmes, 1973] In listening tests, glottal waveform models have repeatedly been shown to provide more natural sounding speech than impulse excitations [Pinto et al , 1989, Childers and Wu, 1990, Carlson et al , 1990] In addition, the more precise glottal models have proven capable of synthesising different voicing styles, such as modal, vocal fry, falsetto and breathy [Childers and Lee, 1991, Lalwani and Childers, 1991, Childers and Ahn, 1995] Accurate reproduction of the source waveform has also proven useful in the synthesis of female voices [Klatt and Klatt, 1990, Karlsson, 1990, 1991, 1992] Furthermore, there is evidence to suggest that naturalness can be further improved by using phoneme-specific glottal parameters [Fries, 1994]

Various techniques for representing source-tract coupling have been investigated In general, glottal flow models capture waveform skewing but do not represent source-tract interaction during the open phase Several methods of modelling open phase coupling have been proposed, including the use of modified glottal volume velocity models [Guérin et al , 1976], different vocal tract filters during the open and closed phases [Brookes and Naylor, 1988, Krishnamurthy, 1992, Childers and Wong, 1994] and an electrical analog synthesiser [Allen and Strong, 1985] The results of these experiments are inconclusive as to the importance of source-tract coupling for speech synthesis Certainly the interaction effect exists, but, since the human ear is relatively insensitive to formant bandwidth changes, it does not appear to be important for re-synthesis purposes

3.6.3 Speech Coding

Glottal waveform based coding schemes have the potential to achieve higher quality speech at a lower bit rate than conventional systems It is already well established that glottal waveform excitation provides high quality speech synthesis Furthermore, the parameters of the glottal excitation are fewer and are slower time-varying than those of conventional LP residual models The main problem with glottal based coding is that reliable and robust techniques for automatic extraction of the glottal waveform from the speech signal have proven difficult to develop

Early work on glottal coding was carried out by Hedelin, who examined the use of glottal excitation models in a LP coding system [Hedelin, 1984] The system inverse filters the speech signal and fits a polynomial waveform model to the estimated glottal signal Following this, ARX estimation is carried out to determine the optimum LP synthesis filter, given the glottal excitation A simple LPC-10 type voicing decision is used to cope with unvoiced frames Hedelin found that the system produced much higher quality speech than LPC-10 at a comparable bit rate

Also using an inverse filtering approach, Alku and Laine reported the use of their Adaptive Inverse Filtering technique in a speech coding system The incoming speech signal is inverse filtered to obtain a glottal waveform estimate A flow model is fitted to the extracted glottal excitation and speech is re-synthesised by applying it to the estimated vocal tract filter They studied the use of three glottal waveform models - a polynomial [Alku and Laine, 1989b], a Lagrange interpolation scheme using five reference points [Alku, 1990a] and a two pole filter plus white noise [Alku, 1990b, 1991] They concluded that the polynomial model was most sensitive to distortions of the incoming speech signal but that it provided good quality speech at roughly 4 kb/s In contrast, the Lagrange and filter schemes were more robust but operated at rates of 4-8 kb/s and 5 kb/s, respectively A similar iterative inverse filtering based coding scheme, employing glottal modelling by spline functions, was also developed by Benitez et al [Benitez et al , 1992]

An analysis-by-synthesis approach was first used in glottal coding by Hedelin [Hedelin, 1986] In this, the initial inverse filtering, glottal model fitting and ARX estimation steps are executed as before Next, the glottal and filter parameters are iteratively optimised to minimise a perceptually weighted error criterion between the re-synthesised speech and the original Hedelin claimed that the system gave good speech quality at a rate of 3 kb/s However, due to the iterative optimisation procedure the technique is very computationally complex

In order to improve the robustness of the procedure, Bergstrom and Hedelm later proposed the use of a mixed excitation consisting of glottal waveform, impulse and noise sources operating in tandem [Bergstrom and Hedelin, 1988] The glottal and vocal tract parameters are estimated and optimised as before Following this, a standard multipulse technique is employed to determine the optimum time instants and amplitudes for the impulses Finally, the energy of the noise source is calculated from the prediction error of the model From preliminary experiments, Bergstrom and Hedelm determined that the speech quality was superior to that of LPC-10, at a transmission rate of approximately 5 kb/s Furthermore, due to the mixed excitation, the system is much more robust than either of Hedelin's earlier glottal coding schemes

The principles of multipulse systems have also been applied to glottal codmg by Leung et al [Leung et al , 1990] During voiced speech, a glottal pulse function and a standard impulse source are used to excite an LP filter As before, the parameters of the glottal pulse and the impulse source are chosen to minimise the perceptually weighted error in an analysis-by-synthesis scheme In this case, the glottal pulses are thinned to reduce the computational burden The system is robust and provides a SNR gain of 2-3 dB over standard multipulse coding for voiced speech

Bergstrom and Hedelm also investigated the performance of a codebook-driven glottal coding scheme [Bergstrom and Hedelin, 1989] A codebook of two double differentiated glottal pulses is used in tandem with a stochastic codebook and a long-term predictor The combined excitation is passed to a conventional LP filter for synthesis The codebook indices and gains are determined via a perceptually weighted analysis-by-synthesis scheme Bergstrom and Hedelm found that the coder produced higher quality speech than conventional CELP and was robust to both noisy and phase distorted speech The coder achieved an overall transmission rate of 7-9 kb/s

In a recent development, Cheng and O'Shaughnessy have applied the principles of the glottal estimation to very low bit rate speech coding [Cheng and O'Shaughnessy, 1993] The method uses an all-pole vocal tract model with a glottal excitation provided by two polynomial basis functions operating in tandem with a white noise source The parameters of the model are determined by a least mean square error method A large reduction in coding rate is achieved through short-term temporal compression of the speech and vector quantisation. In addition, finite-state vector quantisation is introduced to further decrease the coding rate. The system provides natural-sounding speech at a bit rate of 450-600 kb/s with a delay of about 200 ms. Due to the provision of a mixed excitation, the system is robust to noise and voice classification errors.

3.7 CONCLUSION

Glottal modelling is based on an accurate representation of the human speech production system. As such, it is a powerful tool for capturing the dynamics of voiced speech, especially the acoustic effects associated with speaker identity and voicing style.

Various techniques for extracting the glottal waveform from the speech signal have been proposed. Most fall into one of two categories - inverse filtering or source-tract estimation. Of the two approaches, inverse filtering is by far the most computationally efficient. Currently, the most successful inverse filtering algorithms are Closed Phase Inverse Filtering and Iterative Adaptive Inverse Filtering. CPIF is the older method and has been studied in the greatest detail. IAIF has the advantage of not requiring precise *a priori* GCI identification. No studies comparing the performance of the two algorithms have been published as yet.

Two types of model have been used to represent the glottal excitation. Dynamic models capture the movement of the vocal cords and flow models parameterise the airflow from the glottis into the vocal tract. Flow models are generally the easiest to use since dynamic models require the use of anatomically based control parameters which are hard to measure. Currently, the most common flow model is the LF model which has been shown to capture the essential characteristics of the glottal excitation using just four independent parameters. Matching the LF model to estimated glottal waveforms can be computationally expensive compared to the fitting of polynomial models.

Determining the GCI is important for extracting the micro-melody of voiced speech. Precise identification of the GCI is also a pre-requisite for CPIF. Aside from using special apparatus, such as the EGG, techniques fall into two classes - those that detect the epoch associated with closure and those that detect the closed phase following the GCI. The most modern methods use both of these criteria to improve the reliability of their GCI estimates. Although no comparative experiments have been carried out, the most effective methods currently available appear to be MLED, SVD and Murgia's method.

All of these glottal processing techniques have been applied in a number of ways to the fundamental speech processing problems of recognition, synthesis and coding. Cancelling the glottal excitation from voiced speech removes variability and has improved the accuracy of recognition systems. In addition, using a glottal excitation has considerably improved the naturalness of speech synthesis systems. The use of glottal models has also shown promise in coding applications. The slowly time-varying nature and smooth trajectory of the glottal parameters make them ideal for low rate coding. However, as has been explained, extraction of the glottal parameters from voiced speech is difficult, particularly under conditions of noise and phase distortion. This has led to the use of very coarse glottal waveform approximations, together with impulsive or stochastic innovations. The rationale is that, since precise extraction is difficult, then only rough glottal models can be used. In general, this approach is counter-productive as the advantages of glottal coding cannot be gained without using precise glottal models As yet, no detailed quantitative study has been made of the speech quality, transmission rate and robustness achievable by precise glottal coding systems in comparison to conventional coders A study of this nature is important since it would define the limits of glottal coding techniques and provide a direction for further research in the area

CHAPTER 4

REVERBERATION MODELLING

4.1 INTRODUCTION

Glottal waveform extraction algorithms show poor performance in processing reverberant speech [Holmes, 1975, Markel and Wong, 1976]. Echoes from previous pitch periods can corrupt the current speech signal. Thus, for example, the echo of a previous GCI may be misinterpreted as a new glottal closure. Alternatively, the echo may be incorrectly judged to be an abnormally strong glottal opening, etc. In order that glottal extraction algorithms may be used in conventional speech coding applications, it is important that their reverberation sensitivity be determined experimentally. Obviously, test reverberant speech data must be employed in these experiments. This chapter describes two investigations carried out to ensure that the procedure used to generate the test data accurately models the true reverberation process.

Since a human speaker cannot repeat the same phonation pattern on different occasions, it is desirable that a noiseless, anechoic recording be made and have reverberation added to it. In this way, the true glottal parameters may be estimated by applying glottal extraction algorithms to the noiseless speech. After this, the errors induced by reverberation may be easily identified by comparing these results to those obtained by processing the same speech segment with added reverberation.

Consider the reverberation process [Kuttruff, 1991]. When a sound source operates it displaces a volume of gas either by its movement, e.g. a loudspeaker, or by the emission of air, e.g. human speech. This volume flow produces a pressure wave which moves away from the source. When the source is operating in a room, some of the sound energy striking the walls is reflected back into the room. This is picked up by a microphone as echoes of the direct signal. Obviously, these echoes may themselves be again reflected and so on. In this manner, a reverberant sound field is created in a normal room.

The amount of reverberation captured in a speech recording depends on the nature of the room in which the recording is made and on the lip to microphone distance. Rooms with hard walls reflect more sound energy and so corrupt the speech signal to a greater extent than rooms with soft furnishings. The sound energy directly received at a microphone falls off with increasing source-receiver distance. In contrast, the reverberant sound energy remains relatively constant throughout the room. Thus, the energy of reverberation increases, relative to that of the direct field, as the microphone is moved away from the source. Generally, echoes returning to the source itself are much weaker than the direct signal from the source. Therefore, the pressure generated for a given flow from the source is normally assumed to be constant in all enclosures, regardless of the reflected signal. In other words, it is assumed that the operation of the source is independent of the enclosure into which it radiates. This assumption must be correct if the conventional approach of adding reverberation to anechoic recordings is to be used. That is, the signal recorded under anechoic conditions must be the same as the signal which would be generated at the source if it were actually in a reverberant room. In the next section, the validity of this assumption is examined theoretically and tested experimentally.

Making the above assumption, reverberation approximates to a linear time invariant process [Kuttruff, 1991]. The pressure signal at the receiver $p_R(t)$ can be calculated as a convolution of the pressure signal at the source $p_S(t)$ and the impulse response between the source and the receiver $h(t)$

$$p_R(t) = \int_{-\infty}^{+\infty} p_S(\tau)h(t-\tau)d\tau \tag{4.1}$$

The impulse response depends only on the characteristics of the room and on the positions of the source and receiver. Thus, if the room impulse response is known then reverberant speech can be produced from anechoic recordings.

Two options exist for the determination of typical room impulse responses - measurement or simulation. Obviously, measurement produces the more accurate results. However, it is difficult to construct an apparatus which can make reliable impulse response measurements over the entire speech bandwidth (20-4000 Hz). Thus, the decision was taken to obtain the room impulse responses by simulation. A standard technique, the Image Method [Allen and Berkley, 1979], was selected for this purpose. The third section of this chapter describes experiments carried out to establish the accuracy of the Image Method. In this work, narrowband room responses were measured and compared to those produced by the Image Method.

In summary, this chapter addresses two key issues in the generation of reverberant speech from anechoic recordings. Firstly, the assumption that source-reverberant field interaction is negligible for speech in normal rooms is investigated. Secondly, the accuracy of the Image Method in generating artificial room impulse responses is studied.

The chapter is divided into four sections. The next section describes the experiments undertaken to ensure that the pressure signal radiated from the lips is independent of the enclosure. The section develops theory for predicting the variation in the radiation impedance due to reverberation at a piston in an infinite baffle. This theory is validated by comparisons with in-room measurements and is applied to the problem of predicting the variation in the lip radiation impedance due to reverberation. Section three explains the Image Method and compares impulse responses generated by it to those measured in normal rooms. This comparison is made in terms of the decay rate and the spectral variation of the responses. Section four concludes the chapter.

4.2 VARIATION OF THE RADIATION IMPEDANCE

This section examines the variation in the lip radiation impedance due to reverberation. At low frequencies, less than 4 kHz, the radiating area of the mouth approximates to a piston in an infinite baffle [Flanagan, 1972; Miki et al., 1987]. Theory is developed below for predicting the variation in the radiation impedance due to reverberation at a piston in an infinite baffle. This theory is tested by measuring the radiation impedance variation at a loudspeaker in a normally reverberant room. Based on these results, simulations are conducted to determine the variations which would be encountered at the lips.

The section is broken into four sub-sections. The relevant acoustic theory is developed in sub-section one. The experimental method and results are presented in sub-sections two and three, respectively. Lastly, the findings of the investigation are then discussed in sub-section four.

4.2.1 Theory

The pressure produced in the near-field of an acoustic source can be characterised by the mechanical radiation impedance function [Kinsler et al., 1982]. This is defined as the ratio of force applied by the source to the particle velocity of the source

$$Z_R(\omega) = \pi\ a^2 \frac{P(\omega)}{U(\omega)} \qquad (4.2)$$

where $P(\omega)$ is the pressure amplitude at the source, $U(\omega)$ is the particle velocity amplitude of the source and a is the radius of the source. The radiation impedance of a circular piston set in an infinite baffle placed in the free-field is given by [Morse and Ingard, 1968]

$$Z_R(\omega) = \pi\ a^2 \rho_o c\left(R(2ka) + jX(2ka)\right) \qquad (4.3)$$

where

$$R(x) = 1 - \frac{2J_1(x)}{x} = \frac{x^2}{2^2\ 1!\ 2!} - \frac{x^4}{2^4\ 2!\ 3!} + $$

$$X(x) = \frac{4}{\pi}\left[\frac{x}{3} - \frac{x^3}{3^2\ 5} + \frac{x^5}{3^2\ 5^2\ 7} - \quad \right]$$

This expression specifies, for a baffled source in the free-field, the pressure generated per unit particle velocity of the source. Plots of the functions $R(x)$ and $X(x)$ are shown in Fig. 4.1.

The power radiated away from such a source is determined by the real part of the radiation impedance [Kinsler et al., 1982]

$$\Pi(\omega) = \frac{1}{2} U_0^2(\omega)\, Re\left(Z_R(\omega)\right) \qquad (4.4)$$

where $U_o(\omega)$ is the magnitude of the amplitude of the particle velocity of the source.

In order to make statistical predictions about the behaviour of the reverberant sound field, it is necessary to assume that the field is diffuse [Kuttruff, 1991], that is, the average energy density is the same throughout the volume of the enclosure and all directions of propagation are equally probable. This model over-simplifies the actual behaviour of sound in a room, particularly at low frequencies. It neglects the presence of normal modes, the distribution of absorptive materials and the shape of the room. Schroeder has calculated that the model is reasonable, provided that there are at least three overlapping normal modes at the frequency under consideration [Schroeder, 1962]. Thus, the model is assumed to be valid above the Schroeder frequency given by

$$f_S = 2000\left(\frac{T_{60}}{V}\right)^{1/2} \qquad (4.5)$$

where T_{60} is the reverberation time of the enclosure and V is the volume of the enclosure. The reverberation time of a room is the length of time from when a source in the steady-state is switched off

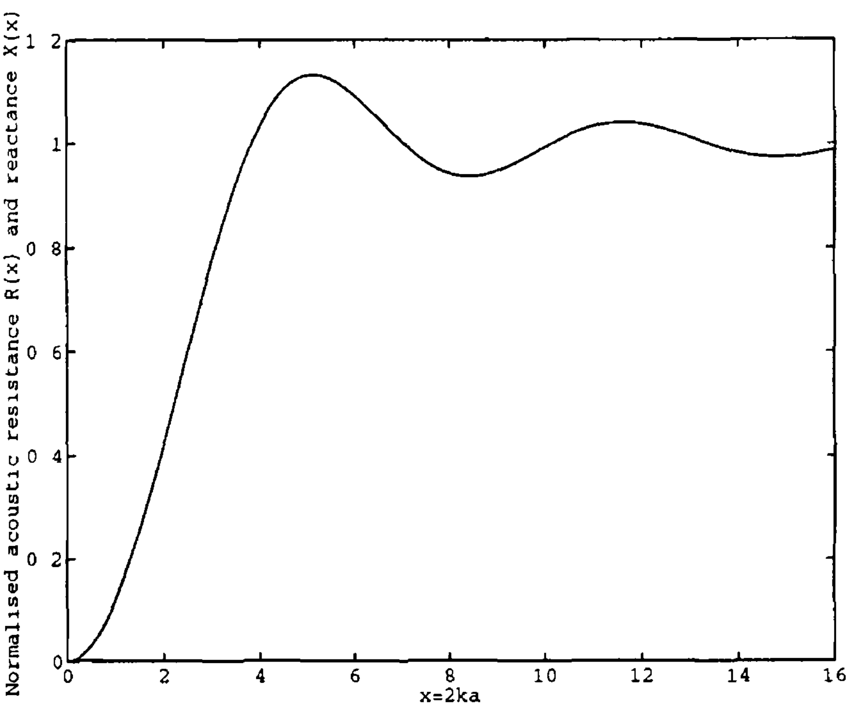


Fig 4 1 Radiation impedance functions for a piston in an infinite baffle solid line - real part R(x), *dotted line - imaginary part* X(x)

until the sound pressure level in the room drops by 60 dB When the sound field in a reverberant room reaches the steady state, the sound power lost from the room equals the power generated by the source Assuming that the reverberant sound field is diffuse, it can be shown that the spatially averaged, time averaged squared sound pressure amplitude of the reverberant field is given by [Kinsler et al , 1982]

$$\overline{P_R^2(\omega)} = \frac{4\Pi(\omega)\rho_o c}{A} \qquad (4\ 6)$$

where A is the total sound absorption of the room, ρ_0 is the air density and c is the speed of sound The total sound absorption can be obtained from [Kinsler et al , 1982]

$$A = \frac{0\ 161\ V}{T_{60}} \qquad (4\ 7)$$

From Eqs (4 2), (4 3), (4 4) and (4 6) it can be shown that the ratio of the mean square pressure level of the reverberant field to the squared magnitude of the near-field pressure is given by

$$\frac{\overline{P_R^2(\omega)}}{|P_D(\omega)|^2} = \frac{2\pi\ a^2 R(2ka)}{A\left[R(2ka)^2 + X(2ka)^2\right]} \qquad (4\ 8)$$

Herein, this quantity is referred to as the reverberant pressure ratio For low frequencies or small pistons ($ka<<1$) it reduces to

$$\frac{\overline{P_R^2(\omega)}}{|P_D(\omega)|^2} \approx \frac{9\pi^3 a^2}{64A} \qquad (4\ 9)$$

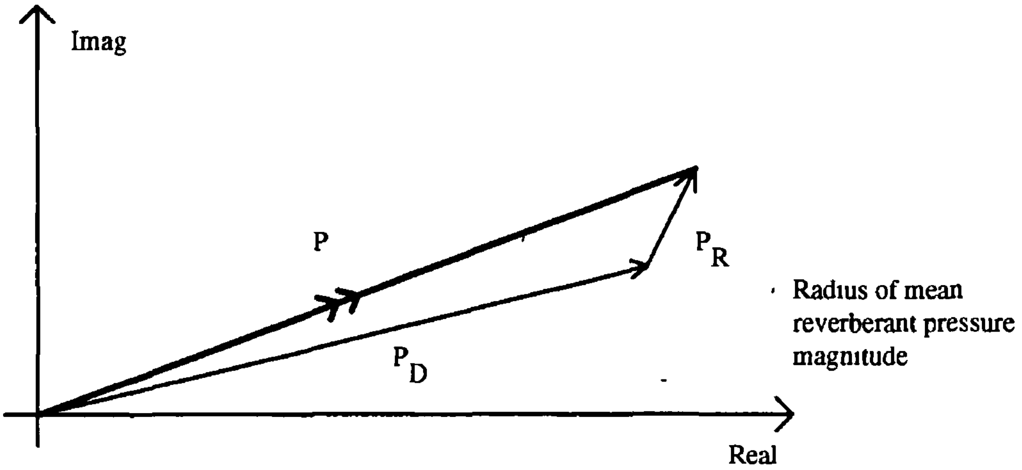


Fig 4 2 Resultant pressure calculated as the complex sum of the direct and reverberant components

It can be seen that the reverberant pressure ratio is directly proportional to the area of the piston Thus, as piston area decreases, the reverberant pressure decreases with respect to the pressure at the source This suggests that source-reverberant field interaction becomes increasingly small for sources of reducing surface area

In the diffuse reverberant field, the pressure at any point is due to a number of pressure components which combine in a random fashion Under the assumption that these components each have a Gaussian distribution, Schroeder has demonstrated that the resultant reverberant pressure amplitude has a distribution given by [Schroeder, 1954]

$$W(z) = \exp(z - \exp(z)) \tag{4 10}$$

where

$$z = ln\left(P_R^2(\omega)/\overline{P_R^2(\omega)}\right)$$

When the direct and reverberant sound fields interact, the pressure experienced is the complex sum of the direct and reverberant pressures, see Fig 4 2 Assuming unit direct pressure at the source, the mean square pressure level of the reverberant field is given by Eq (4 9) The distribution of the reverberant pressure magnitude is governed by its average level as per Eq (4 10) and the phase of the reverberant pressure has a uniform distribution

Using these relations, a Monte Carlo method [Schroeder and Kuttruff, 1962] can be used to simulate the variation in the pressure at the source Reverberant pressures are generated according to the above statistical process and the resultant pressure is calculated as the sum of the direct and reverberant components The variation between the resultant pressure and the free-field pressure, that is, the direct pressure, can then be calculated

Simulations such as these allow the prediction of the variation in the radiation impedance which will occur at a piston-like source placed in a room with a known reverberation time Thus, the radiation impedance variation which occurs at the lips during speech in normal rooms can be determined The next sub-section describes experiments conducted to compare the predictions of this model to actual radiation impedance measurements carried out using a loudspeaker.

4.2.2 Method

In order to confirm the above theory, measurements were made of the variation in the radiation impedance occurring at a loudspeaker in a reverberant enclosure. There are two main approaches to measuring acoustic impedance functions. In the first, a source of constant volume velocity is used and the pressure signal is measured using a normal microphone. This technique is simple to use. However, sources of constant volume velocity are difficult to manufacture, particularly over the frequency range and Signal to Noise Ratio required in this experiment. The second approach is to use a normal sound source but to measure its volume velocity. This method is more cost effective. The method of Salava [Salava, 1988], which conforms to the second approach, was chosen for use in these experiments. The accuracy of the method has been confirmed experimentally [Anthony and Elliott, 1991].

Salava's method involves the measurement of pressure with a normal microphone and the measurement of flow with an inverted loudspeaker. The inverted or passive loudspeaker is acoustically coupled to an identical driver unit, see Fig. 4.3. The driver unit is excited by a pseudo-random sequence and the inverted cone vibrates in sympathy. The movement of the passive cone generates an e.m.f. in its speaker coil and the acoustic signal is radiated from its back. At low frequencies the passive cone behaves as a rigid piston of constant area. Thus, the induced e.m.f. is directly proportional to the velocity of the cone and so to the volume velocity of the air displaced by it. The pressure at the source is measured by placing a microphone close to the back of the passive cone.

The radiation impedance $Z_R(x_S,\omega)$ is calculated using the cross-spectral technique [Bendat and Piersol, 1971] as the transfer function between the volume velocity, measured by the passive cone, and the pressure, measured by the microphone

$$Z_R(x_s,\omega) = \frac{\sum_{i=1}^{N} X_i^*(x_s,\omega) Y_i(x_s,\omega)}{\sum_{i=1}^{N} X_i^*(x_s,\omega) X_i(x_s,\omega)} \tag{4.11}$$

where $X(x_S,\omega)$ and $Y(x_S,\omega)$ are the Fourier Transforms of the cone $x(n)$ and microphone $y(n)$ sequences respectively and N is the number of recordings made for the source location x_S.

The experiments were performed using a PC and a Loughborough Sound Images DSP data acquisition card with on-board ADC/DAC. The driver speaker was excited by a pseudo-random maximal length sequence (length 32767) [Kuttruff, 1991] at a sampling frequency of 16 kHz. The excitation signal was anti-aliased using a passive 5kHz lowpass filter and amplified by a JVC AX-11 amplifier. The speaker, a Radionics 8 ohm 6.5 in. bass/mid-range unit, was installed in a 30 cm by 20 cm by 13 cm wooden speaker cabinet which was lined with sound absorbing foam. The acoustic coupling between the speakers was stiffened by reducing the air volume between the cones. This was achieved using a perforated wooden plate with metal bolts attached to it. This increased the frequency range over which the cones moved in sympathy. The pressure signal was measured using a Brüel and

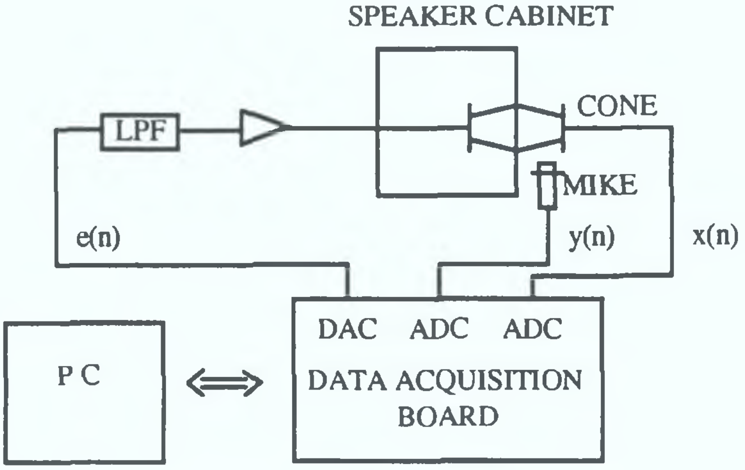


Fig. 4.3. Room impedance measurement apparatus.

Kjær microphone (model 4006) with diffuse head and, like the passive cone e.m.f., was amplified using an Alice Soundtek pre-amplifier.

For each measurement, 5 records of the pressure and flow signals were taken (N=5). Each of these records was obtained by repeating the excitation signal 11 times, discarding the results of the first cycle and averaging the remaining 10 cycles in the time-domain. The coherence function between the cone and microphone signals and the excitation signal was estimated within an individual recording to ensure linearity [Bendat and Piersol, 1971]. The 95% confidence limits for the measurements were found to be ±0.2 dB and ±0.02 rads in the frequency range 50-2000 Hz.

In order to determine any change in the impedance function due to the enclosure, it was decided to perform the measurements in two very different rooms. A hemi-anechoic studio was used to make an almost free-field measurement of the radiation impedance. For comparison purposes, a small highly-reverberant room was used for the other radiation impedance measurement. The studio measured 3.0 m by 3.0 m by 2.7 m. The walls were lined with acoustic wadding and heavy curtains, the floor was carpeted and acoustic tiles were fixed to the ceiling. The reverberant room was 3.4 m by 2.6 m by 2.7 m with smooth plastered walls, acoustic ceiling tiles, a concrete floor and no windows. Neither room contained any furniture.

To provide information on the reverberant process in the two rooms, the reverberation times were determined using Schroeder's integrated impulse response technique [Schroeder, 1965]. The impulse response measurements were made using the m-sequence cross-correlation method, again proposed by Schroeder [Schroeder, 1979]. As in the impedance experiments, the measurements were made using a PC and LSI development board. An m-sequence pseudo-random signal was emitted by a Fostex 6301B active loudspeaker and recorded using the B&K microphone. The impulse response was measured at six receiver locations for each of four source locations. These responses were filtered into third-octave bands [Beranek, 1992] and the integrated impulse responses were calculated. The reverberation time was estimated for each source-receiver position by manually fitting a straight line to the early (-5 to -35 dB) decay of the sound pressure level. The resulting average third-octave reverberation times of the two

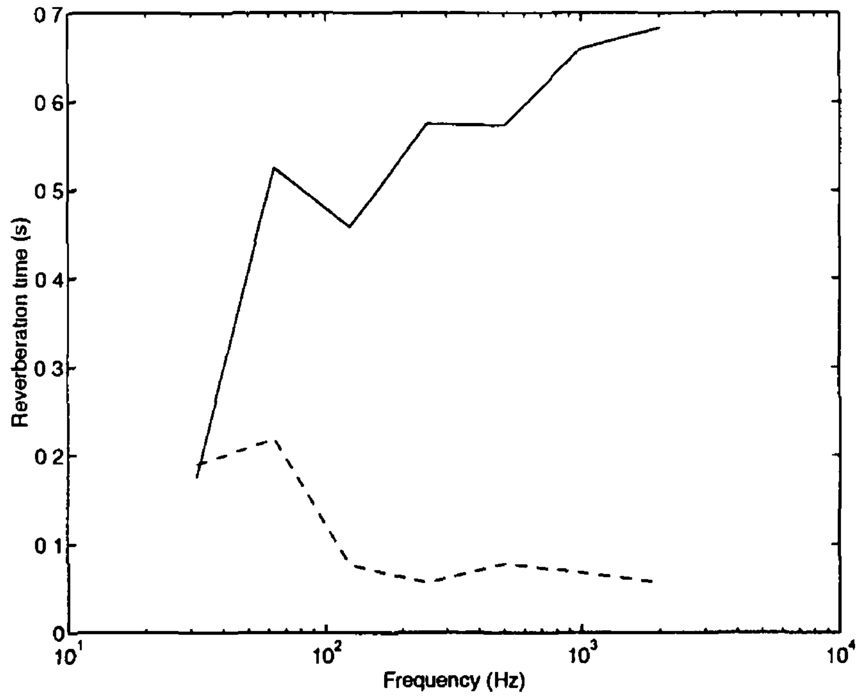


Fig 4 4 Measured reverberation times in third octave bands solid line - reverberant room dashed line - studio

rooms are shown in Fig 4 4 From these results, the Schroeder frequency for the reverberant room was calculated as approximately 300 Hz Further information on the error analysis procedure used in the impulse response measurements is provided in [Bleakley and Scaife, 1995] (Appendix B)

To compare the radiation impedance measurements with the newly developed theory, Monte Carlo simulations were carried out The frequency band 1-1 6 kHz was modelled based on a reverberation time of 0 68 s and a loudspeaker radius of 8 95 cm The reverberant pressure ratio for the room was calculated from the reverberation time using Eq (4 8) From the resulting average reverberant pressure and following the probability density function given in Eq 4 10, 10000 samples of the combined direct and reverberant pressure fields were calculated The distribution of the impedance variation so generated was then compared to that measured for the loudspeaker by Salava's method

In order to determine the variation in radiation impedance occurring for speech, the Monte Carlo simulations were repeated using typical lip radiation areas A man articulating a rounded vowel, such as [u], produces a mouth opening of appropriately 0 9 cm^2 [Flanagan, 1972] For an open vowel, such as [a], the mouth area increases to 5 cm^2 These areas correspond to circular pistons with radii of 0 5 cm and 1 3 cm, respectively The Monte Carlo simulations were repeated using the same room parameters and these new piston radii In this way, the radiation impedance variation occurring at the lips during speech was determined

4.2.3 Results

The measured radiation impedance of the loudspeaker in the studio and the room are shown in Figs 4 5 (a) and (b), respectively Clearly, the randomisation caused by reverberation is at a low level

Fig 4 6 shows the variation in magnitude and phase of the radiation impedance between the room and studio The variation increases with frequency as does the reverberation time, cf Figs 4 4 and 4 6 Increased reflection of sound energy from the walls at higher frequencies leads to greater energy in the reverberant field and so to greater variation in the radiation impedance

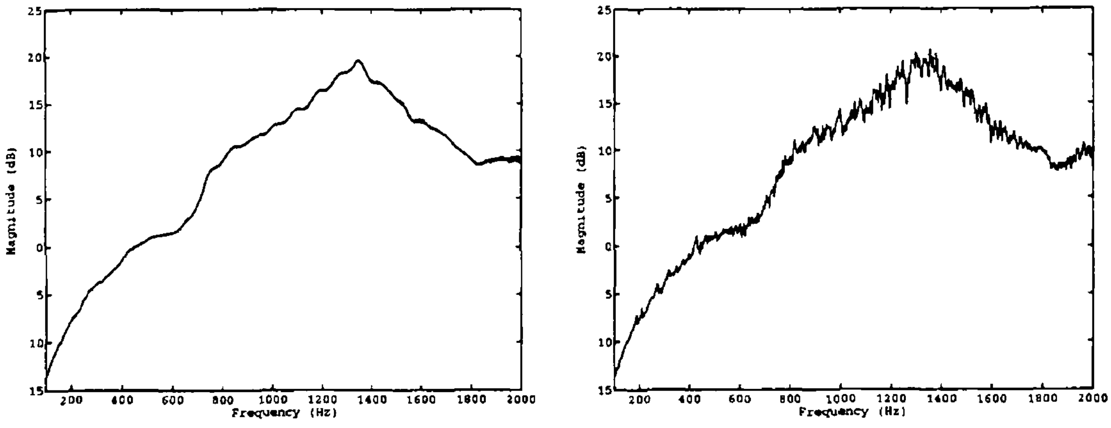
The distribution of the measured magnitude and phase variation occurring between 1 kHz and 1 6 kHz is shown in Fig 4 7 From these graphs it can be seen that, in this middle frequency band, 90 % of the magnitude variation occurs in a band less than 4 dB wide Similarly, 90 % of the phase variation occurs in a band less then 0 4 rads wide

The calculated reverberant pressure ratio (Eq 4 8) for the loudspeaker in this room is shown in Fig 4 8 The mean ratio in the frequency range under consideration is -21 6 dB The resulting distribution of the variation of the simulated radiation impedance is shown in Fig 4 9 Over 95 per cent of the variation occurs in a band less than 4 dB and 0 4 rads wide

Comparing Figs 4 7 and 4 9, the similarity between the measurements and the Monte Carlo simulations can be seen However, the simulations predict slightly less variation than is actually encountered There are two possible explanations for this Firstly, the studio is not truly anechoic As a result, the variation between the in-studio and in-room measurements is larger than that which would be observed between free-field and in-room measurements Secondly, the reverberant field in the room is not diffuse, that is, the energy density is not the same at all points in the room Thus, the actual acoustic measurements may differ from the predictions due to the position of the source Nevertheless, the discrepancy is small and, overall, the results support the accuracy of the derived formulae and the Monte Carlo simulation technique

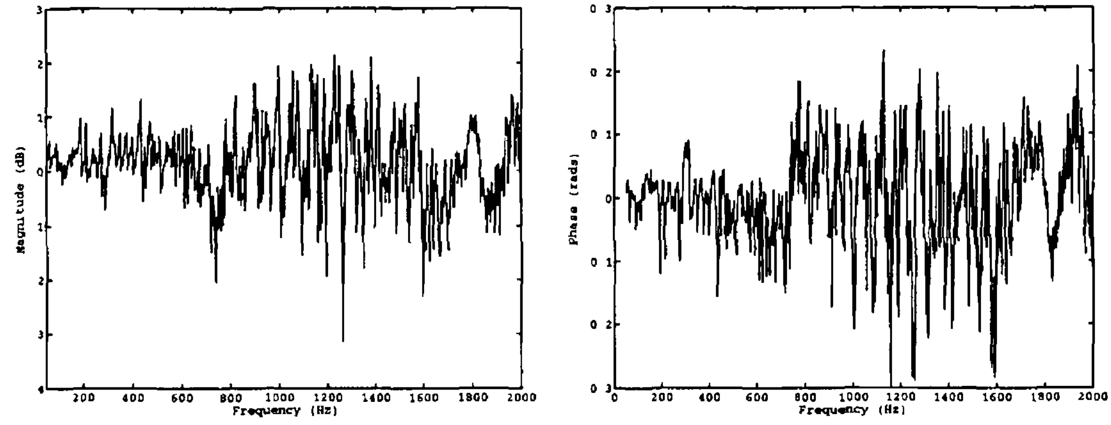
Applying Eq (4 8) to the lip areas, the reverberant pressure ratios for the vowels [u] and [a] are -47 dB and -39 dB respectively The variation in the radiation impedance for the two vowels, as determined by Monte Carlo simulations, is shown Fig 4 10 In both cases, the magnitude and phase variations are negligible

On further experimentation it has been found that, even for open vowels, a radiation impedance variation of ±0 5 dB would require the room to have a reverberation time of 3 s Reverberation times of this length are not normally encountered in small enclosures



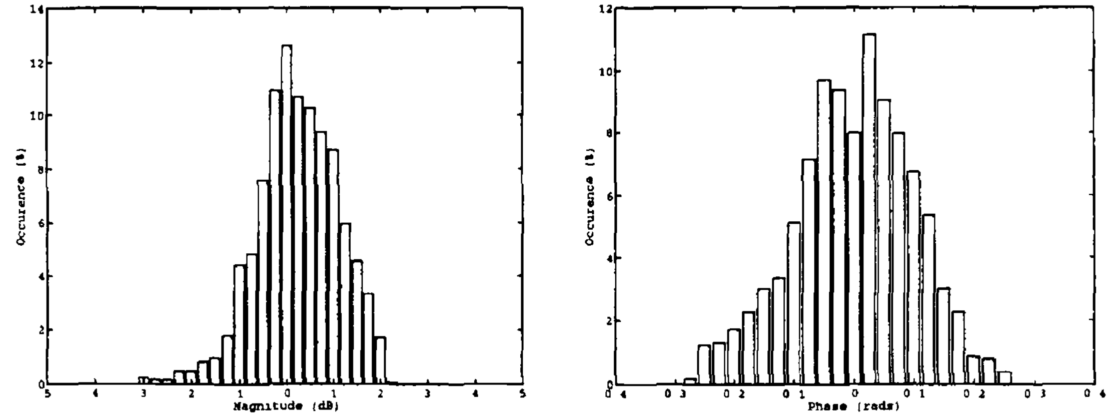
(a) (b)

Fig 4 5 Measured radiation impedance of speaker assembly (a) in studio, (b) in room.



(a) (b)

Fig 4 6 Measured variation in the radiation impedance of the loudspeaker occurring in the room with respect to the studio (a) magnitude, (b) phase



(a) (b)

Fig 4 7 Distribution of the measured variation in the radiation impedance of the loudspeaker in the room with respect to the studio (a) magnitude (b) phase

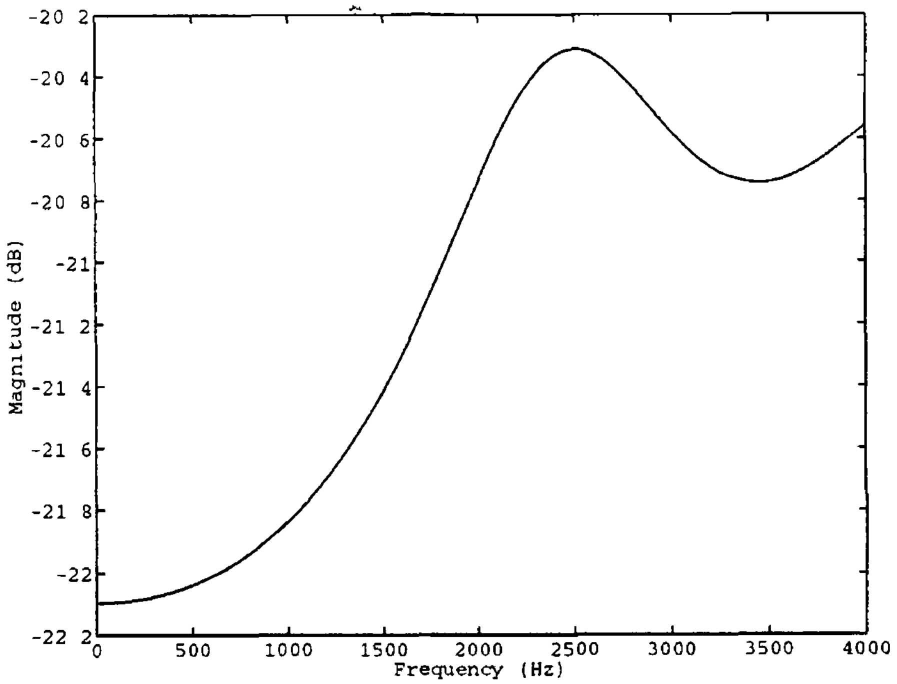
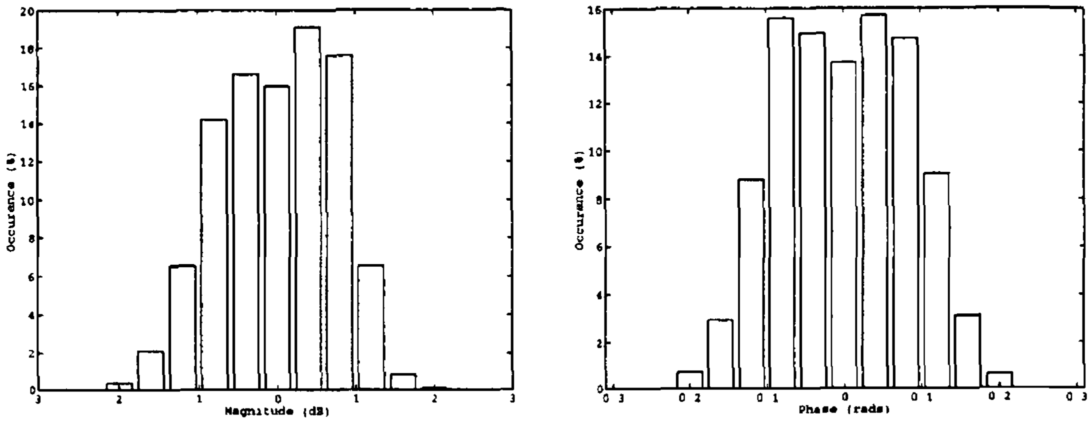
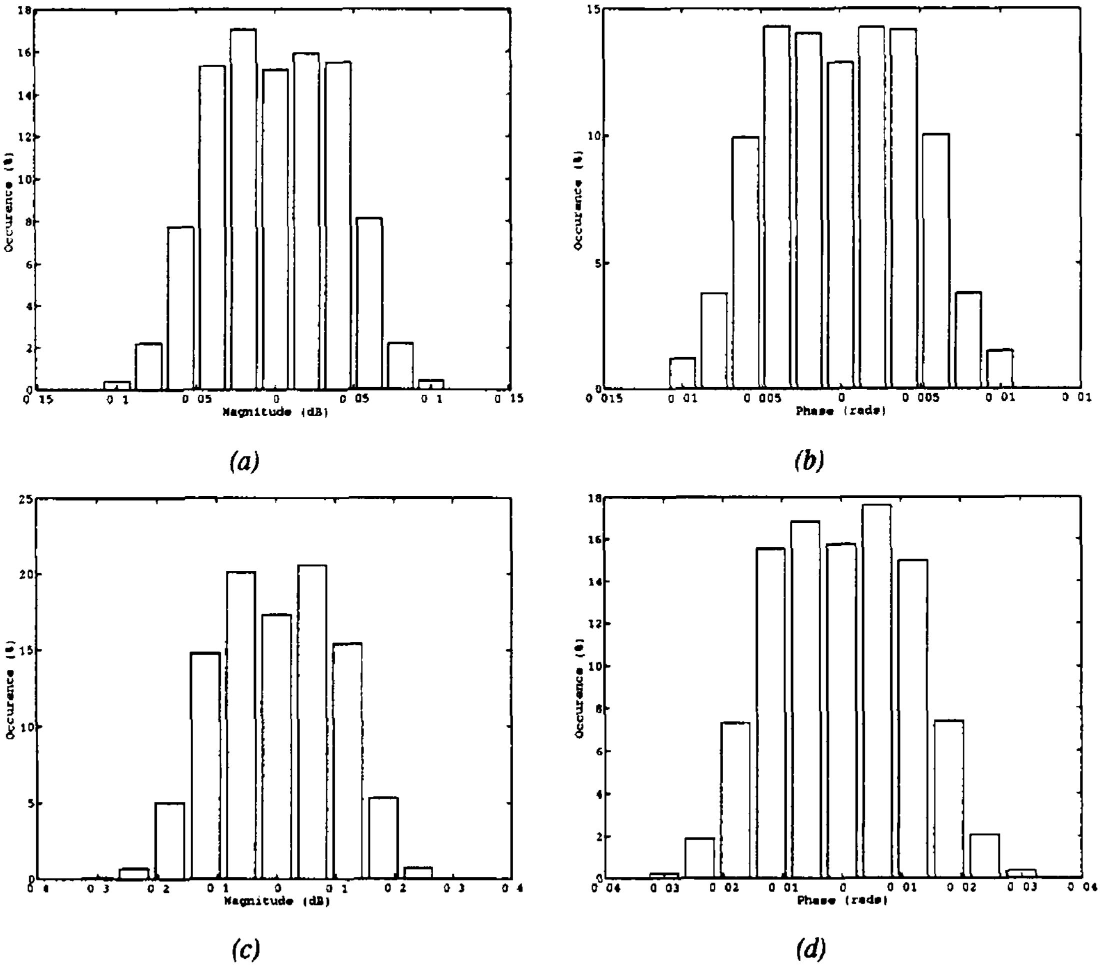


Fig 4 8 Reverberant pressure ratio calculated for the loudspeaker



(a) *(b)*

Fig 4 9 Distribution of the simulated variation in the radiation impedance of the loudspeaker in the room with respect to the studio (a) magnitude, (b) phase



(a) (b)

(c) (d)

Fig. 4.10 Distribution of the simulated variation in the radiation impedance in the room with respect to the free-field for piston radii of (a) 0.5 cm, magnitude, (b) 0.5 cm, phase, (c) 1.3 cm, magnitude, (d) 1.3 cm, phase

4.2.4 Discussion

The theory and results presented above indicate that the steady state variation in the radiation impedance at the lips due to reverberation is, in general, less than ±0.5 dB and ±0.05 rads. Of course, steady state conditions are never reached by a person talking in a room. The amplitude of the direct pressure from the source rises and falls according to what is being said. At onsets there is no time for reverberation to build up. Therefore, there will be no variation in the radiation impedance due to reverberation. During prolonged segments of speech, however approximate steady state conditions will be reached. At these times, the preceding findings are applicable and the radiation impedance variation is negligible. In contrast, at offsets the direct pressure falls rapidly, while the reverberant pressure decreases slowly. The above results indicate that, for the lips, the steady state ratio of the reverberant to the source pressure level is less than -39 dB. This means that the source and reverberant pressure levels will be almost equal if the energy in the speech signal falls by 40 dB in a short time. A fall of this kind is quite possible at offsets. Therefore, some interaction may occur at these times. As a result, glottal extraction algorithms may fail at voicing offsets due to reverberation.

For practical purposes, however, the variation in the radiation impedance of the lips due to reverberation is negligible It can be concluded that, for radiation areas equal to or smaller than the mouth, the interaction between the source and the reverberant field can be ignored Therefore, modelling reverberation by filtering speech data recorded under anechoic conditions is a valid procedure Additionally, the results indicate that glottal waveform extraction is always possible under reverberant conditions, provided that the recording microphone can be placed sufficiently close the lips

4 3 REVERBERATION SIMULATION

This section describes experiments carried out to ensure the accuracy of the Image Method for generating artificial room impulse responses The first sub-section describes the Image Method and analyses some of its inherent assumptions Sub-section two explains the experimental method used to measure actual room impulse responses The third sub-section presents the results of the impulse response measurements and compares them with those generated by the Image Method Following this, the section is concluded with a discussion of the results

4.3 1 Theory

There are two commonly used approaches to generating artificial room impulse responses - ray tracing and the Image Method [Kuttruff, 1991] The Image Method was chosen for use in this investigation for two reasons Firstly, the ray-tracing approach may miss valid source-receiver paths since the scheme only samples the emission space Secondly, the Image Method is of lower computational complexity for the small rectangular enclosures under consideration

The Image Method [Allen and Berkley, 1979] is based on the idea that each wall is a "mirror" for sound Thus, a mirror image of a room can be imagined as existing on the other side of each wall Furthermore, each wall in the mirror image is itself a mirror and so the process repeats itself infinitely For a rectangular room this leads to a grid of virtual or imaginary rooms as shown in Fig 4 11 Each virtual room contains a virtual source The impulse response between the source and receiver can be simulated by imagining that all the sources emit a pulse at $t=0$ The energy of each pulse is attenuated by the walls that the signal "passes through" on its way to the receiver This attenuation is modelled by the reflection coefficient of the wall That is, the pulse arriving at the receiver is multiplied by the reflection coefficients of all the walls that it passes through The impulse response is the record of the total energy received at each sampling instant Finally, a highpass filter with cut-off at 50 Hz is used to remove the DC offset of the response Typical room impulse responses generated using the Image Method are shown in Fig 4 13

The basic Image Method of Allen and Berkley does have a number of limitations, some of which have been addressed in later work by other authors

(a) The pressure produced at the source is assumed to be independent of the enclosure into which it radiates This assumption was found to be valid for speech, as explained in the previous section

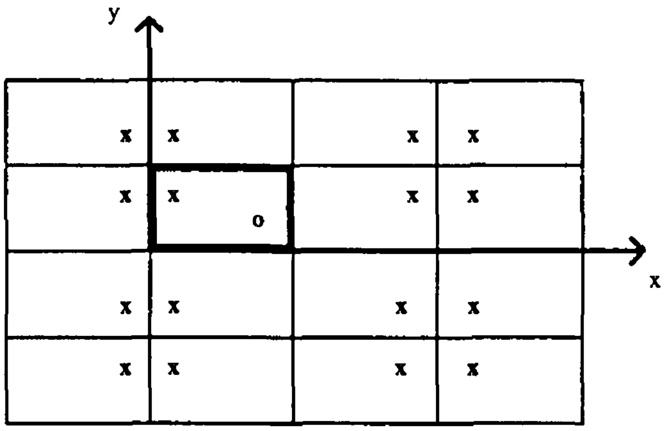


Fig 4 11 A two dimensional slice through the image space showing how the images are arranged spatially The solid box represents the original room, the crosses denote the position of sources and the circle denotes the receiver position [after Allen and Berkley, 1979]

(b) Only angle independent, frequency independent, specular reflections are modelled Allen and Berkley do not believe that this introduces serious problems

(c) The wall reflection coefficients must be greater than 0 7 This is reasonable for normal rooms

(d) Only simple, omni-directional sources and receivers are included in the basic model For the purposes of this investigation, the mouth roughly approximates a point source, particularly at low frequencies In addition, at low frequencies most microphones are not directional It should be noted that these effects could be included with some computational cost [Czyzewski and Nabelek, 1991, Kompis and Dillier, 1993, Culling et al , 1994]

(e) Pulses are shifted to the nearest sample, regardless of the actual arrival time In this investigation, this was corrected by using lowpass filter impulse responses centred on the exact arrival time [Peterson, 1986, Culling et al , 1994]

These limitations are investigated in the next sub-section by comparing measured impulse responses with those generated by the Image Method

4 3.2 Method

Real room impulse responses were obtained by measuring the transmission impedance of a speaker in a normally reverberant room The impedance measurements were made using Salava's method (see Section 4 2 2) The room impulse responses were calculated from the transmission impedance functions as follows

The sound field generated by an acoustic source is often characterised by use of the transmission impedance This is defined as the ratio of the pressure observed at some receiving point to the volume velocity of the source

$$Z_T(x_R, x_S, \omega) = \frac{P(x_R, \omega)}{V(x_S, \omega)} \tag{4 12}$$

where $P(x_R,\omega)$ is the pressure amplitude at the receiving point x_R and $V(x_S,\omega)$ is the volume velocity of the source positioned at x_S.

Consider a point source operating in the free-field. The transmission impedance only depends on the source-receiver distance d [Kinsler et al., 1982]

$$Z_T^{FF}(d,\omega)=\frac{P(d,\omega)}{V(0,\omega)} \tag{4.13}$$

Now consider a simple source operating in a reverberant room. Assuming that source-reverberant field interaction is negligible, as shown in the previous section, and that reverberation is a linear time invariant process, then the pressure signal recorded at the receiver is the convolution of the pressure signal radiated by the source and the room impulse response (Eq. 4.1). Therefore

$$Z_T^{REV}(x_R,x_S,\omega)=H(x_R,x_S,\omega)Z_T^{FF}(0,\omega) \tag{4.14}$$

where $H(x_R,x_S,\omega)$ is the room transfer function for that source-receiver configuration.

Analysing this expression, it can be seen that the room transfer function incorporates a time delay and a scaling factor. The delay is equal to the time required for sound to travel directly from the source to the receiver. The scaling factor is equal to the reduction in amplitude of the direct signal between the source and the receiver. Altering the formula to remove the delay and normalise the transfer function and assuming an omni-directional source, we obtain

$$Z_T^{REV}(x_R,x_S,\omega)=H'(x_R,x_S,\omega)Z_T^{FF}(x_R-x_S,\omega) \tag{4.15}$$

Thus, the room impulse response can be obtained from the free-field and in-room transmission impedances

$$h(n)=\mathrm{IFT}\left(\frac{Z_T^{REV}(x_R,x_S,\omega)}{Z_T^{FF}(x_R-x_S,\omega)}\right) \tag{4.16}$$

where IFT(x) is the Inverse Fourier Transform.

Based on this result, real room impulse responses were calculated from the transmission impedances measured by Salava's method. In the previous section, the acoustic radiation impedance was measured by recording the pressure at the speaker. In this experiment, however, the transmission impedance was measured by placing the microphone some distance from the source. Aside from this, the impedance measurement procedure was carried out in exactly the same manner as before.

The free-field transmission impedance of the speaker was estimated by measuring the spatially averaged transmission impedance in the studio. This was done by measuring the transmission impedance at certain fixed on-axis source-receiver distances (7, 12, 32, 62 and 100 cm) for four different source locations. The results were averaged over the four source locations. This spatially averaged transmission impedance tends to the free-field value. At high frequencies, the impedance variation is due to the sum of many overlapping modes and is a random function of source and receiver position. These effects cancel by averaging over a number of source and receiver locations [Davy, 1981]. At low frequencies, the high damping of the room wall coverings minimises the impedance variations caused by standing waves. The spatially averaged transmission impedances obtained in this way are smooth functions of frequency to within ±0.5 dB. This strongly supports the assumption that reverberant effects were minimised by the averaging process and that the spatially averaged studio measurements approximate the free-field values.

The transmission impedance of the speaker was measured in the reverberant room at the same source-receiver distances The room impulse responses were calculated using Eq 4 16 A phase linear highpass filter, with cut-off at 100 Hz, and a phase linear lowpass filter, with cut-off at 2 kHz, were applied to remove frequencies at which the measurements were unreliable

Artificial room impulse responses corresponding to the measured responses were generated using the Image Method The parameters of the simulations consisted of the room dimensions, the source and receiver positions and the wall reflection coefficients The average wall reflection coefficient β was calculated from the total sound absorption of the room obtained from the reverberation time by Eq 4 7

$$\beta^2 = 1 - A/S \qquad (4\ 17)$$

where S is the surface area of the room Using a reverberation time of 0 6 s (see Fig 4 4) leads to an average wall reflection coefficient of 0 935 The values 0 93, 0 92 and 0 92 were chosen for the walls, ceiling and floor, respectively The resulting impulse responses were then bandpass filtered in the same manner as the measured responses

4.3.3 Results

The measured room impulse responses are shown in Fig 4 12 and the corresponding simulated room impulses are shown in Fig 4 13 Clearly, the envelopes of the measured and simulated responses are very similar

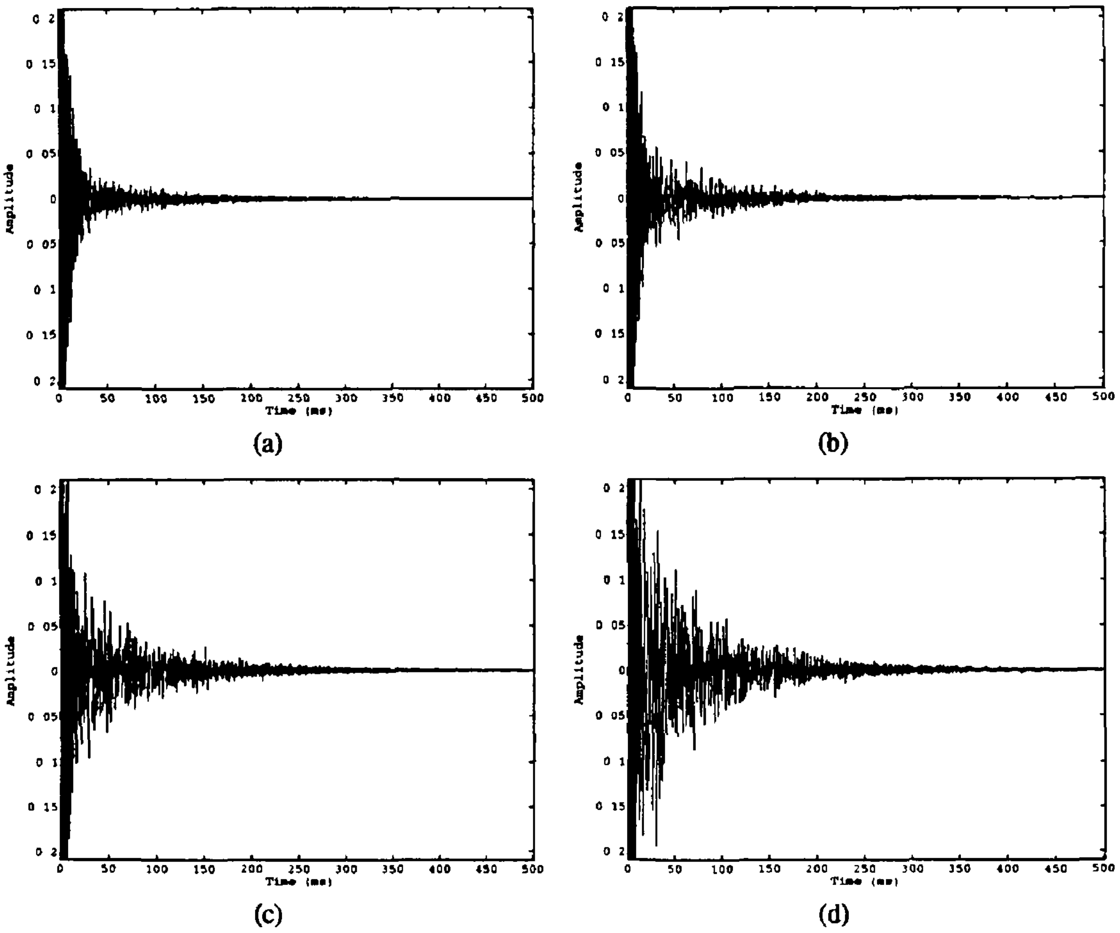
In order to compare the responses in more detail, energy decay curves were calculated The energy decay curve is the average sound pressure level decay occurring at the receiver after a white noise source in the steady state is switched off Schroeder has shown that the decay curve for a particular source-receiver configuration can be calculated as the time reversed integration of the squared impulse response [Schroeder, 1965]

$$E(t) = \int_t^{\infty} h^2(\tau)d\tau \qquad (4\ 18)$$

In the case of the decay curves calculated for impulse response measurements, Chu's method was used to compensate for the effects of residual background noise [Chu, 1978] The calculated energy decay curves are shown in Fig 4 14 Again, a good match was found between measurement and simulation

Another important property of room impulse responses is the spectral randomisation occurring due to reverberation As source-receiver distance increases, the direct-to-reverberant energy ratio decreases and the amount of spectral variation increases The standard deviation of the spectral response was calculated for the measured and simulated responses in the range 500-1500 Hz The results show close similarity and are shown in Fig 4 15

The standard deviation of the spectral response was investigated by Jetzt [Jetzt, 1979] Using Monte Carlo simulation experiments, he determined the standard deviation of the spectral response which would occur for certain source-receiver distances These distances were normalised according to the reverberation distance of the enclosure The reverberation distance is defined as the source-receiver



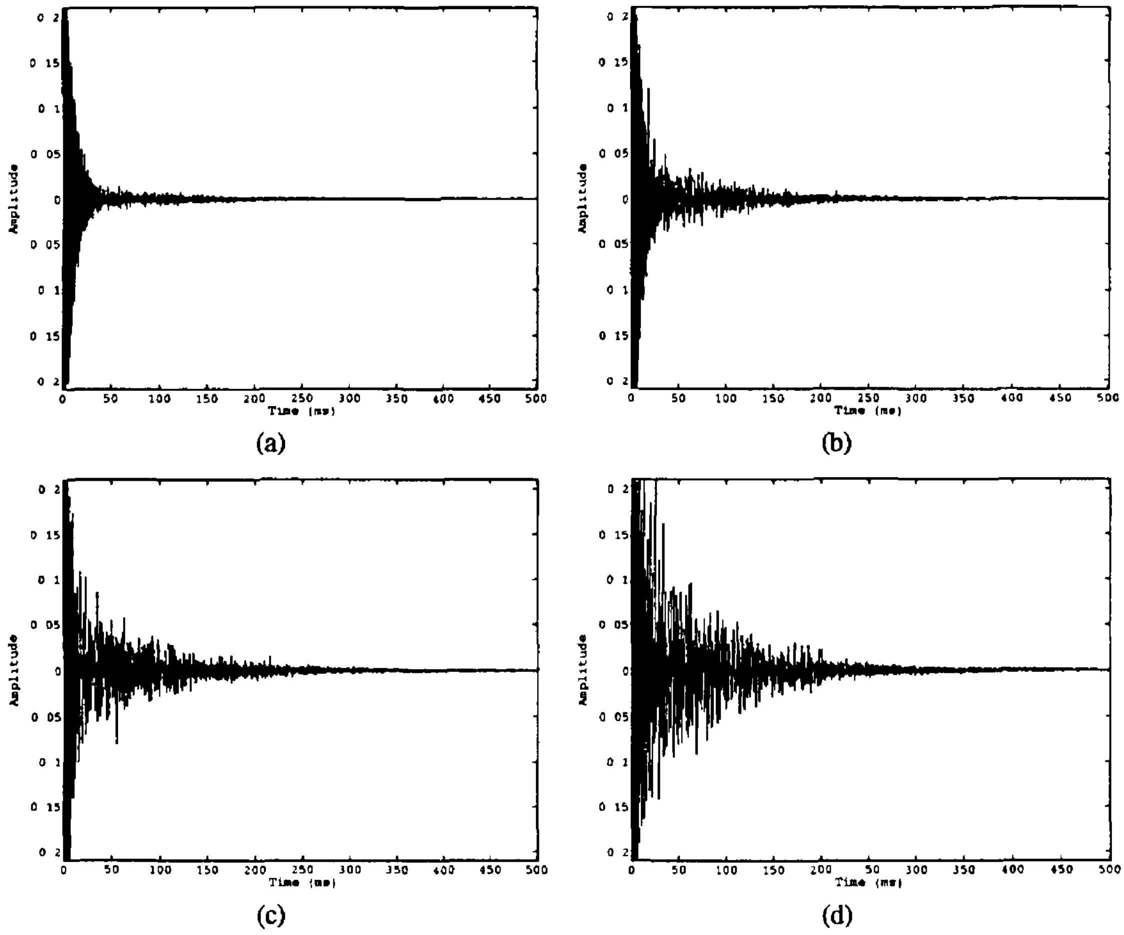
(a)

(b)

(c)

(d)

Fig 4 12 Measured room impulse responses for source-receiver distances of (a) 13 cm, (b) 32 cm, (c) 50 cm, (d) 100 cm.



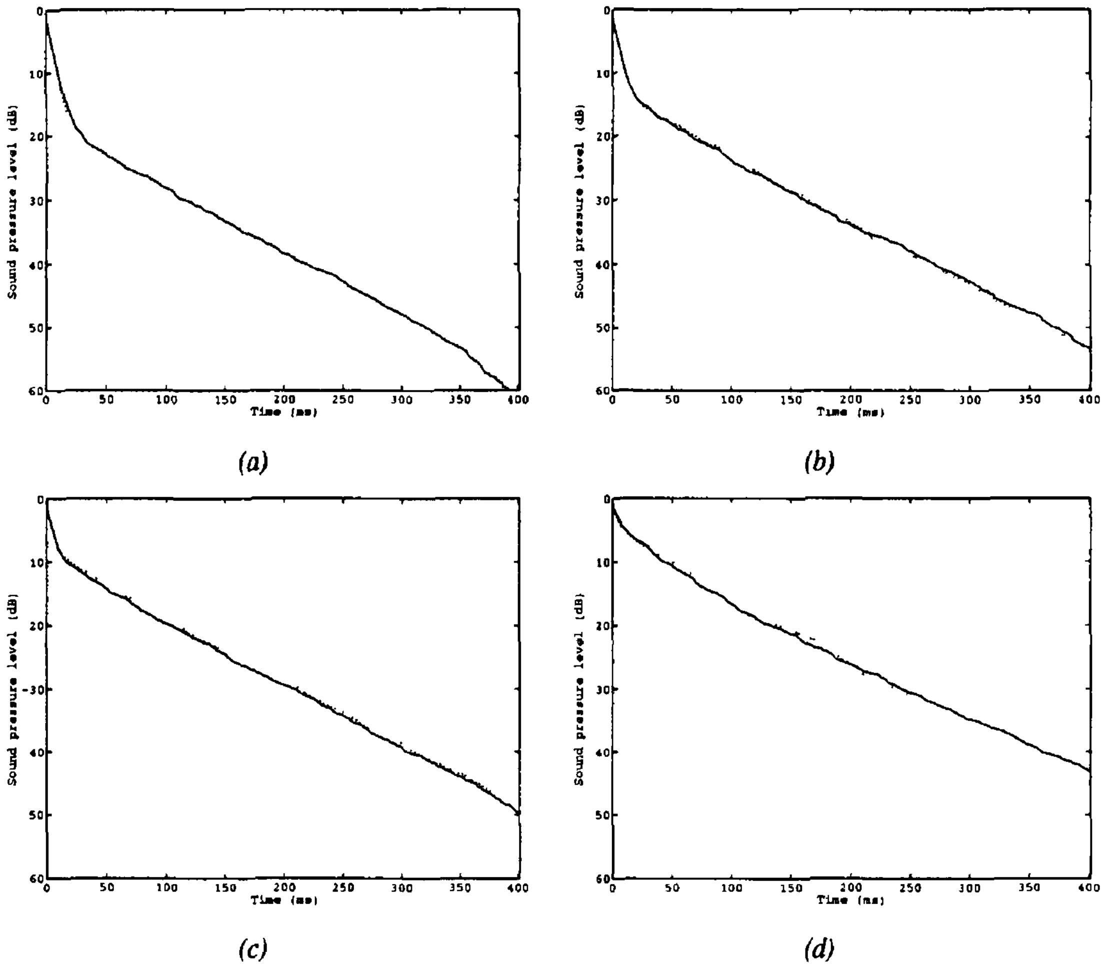
(a)

(b)

(c)

(d)

Fig 4 13 Simulated room impulse responses for source-receiver distances of (a) 13 cm, (b) 32 cm, (c) 50 cm, (d) 100 cm.



(a)

(b)

(c)

(d)

Fig 4 14 Sound pressure level decay for source-receiver distances of (a) 13 cm, (b) 32 cm, (c) 50 cm, (d) 100 cm solid line - measured, dotted line - simulated

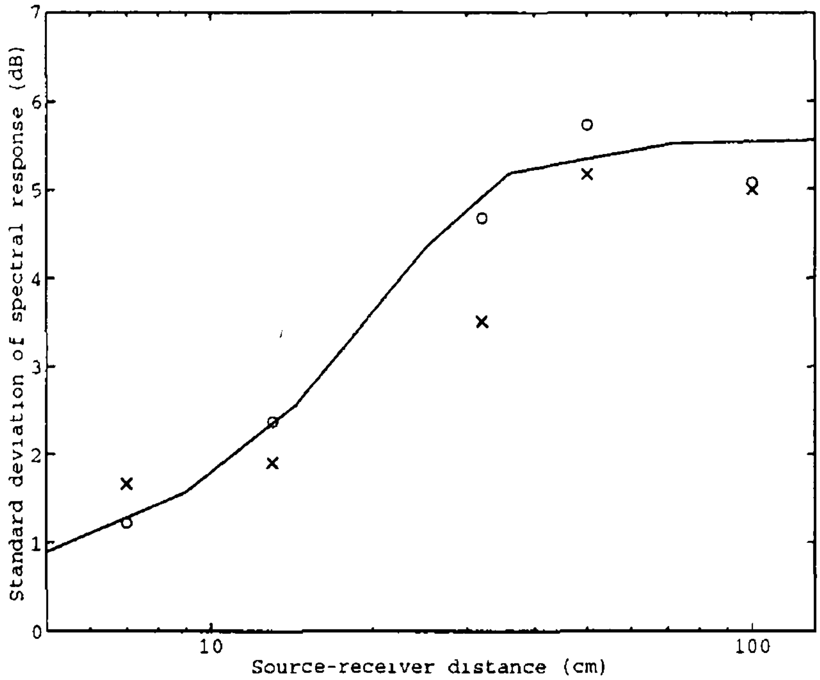


Fig 4 15 Standard deviation of spectral response solid line - theory, crosses - measured room responses, circles - simulated room responses

distance at which the direct and reverberant fields have equal energy For an omni-directional source, the reverberation distance is equal to

$$r_{60} = 0\ 1 \left(\frac{V}{\pi\ T_{60}} \right)^{\frac{1}{2}} \qquad (4\ 19)$$

Jetzt supported his findings with actual measurements of the variations occurring in real rooms

The reverberation time of 0 6 s leads to a reverberation distance of 36 cm Based on this normalisation, the predictions of Jetzt are included in Fig 4 15 The close correspondence between theory, measurement and simulation is clearly illustrated

4 3 4 Discussion

The results presented in this section support the accuracy of the Image Method in generating artificial room impulse responses The measured and simulated impulse responses show a high degree of similarly in terms of their decay rate and spectral response These two properties define the nature of the reverberant response Therefore, the satisfactory performance of the Image Method has been established

Recently developed algorithms for synthesising room impulse responses avoid some the limitations of the Image Method [Lewers, 1993, Nakagawa et al , 1993, Heinz, 1993] However, these new systems are significantly more computationally complex and remain, to some extent, imprecise For the purposes of this investigation, the Image Method has been shown to be sufficiently accurate

4.4 CONCLUSION

Since reverberation has a significant effect on the glottal extraction algorithms under investigation, the accuracy of the proposed method for adding reverberation to anechoic recordings was verified by experiment and simulation Two main issues were addressed

Firstly, the assumption that source-reverberant field interaction is negligible at the lips was investigated New theory was developed for predicting the variation in the radiation impedance of a piston in an infinite baffle source operating in a reverberant enclosure The theory was confirmed by comparing the results of Monte Carlo simulations and measurements made in real rooms using a loudspeaker Predictions were then made for the variation which would occur at the lips These results indicated that the radiation impedance variation occurring at the mouth in a normally reverberant enclosure is negligible Therefore, source-reverberation field interaction can be ignored for the purposes of adding reverberation to anechoic speech

A further implication of this result is that close to the lips, the pressure signal in a room is almost equal to that which would occur if the speaker were in the free-field Therefore, glottal extraction algorithms will operate satisfactorily in any normally reverberant enclosure, provided that the microphone can be placed sufficiently close to the lips The question of how close the microphone must be is investigated in later chapters Further experiments conducted by the author on the effects of reverberation on glottal waveform extraction are described in [Bleakley and Scaife, 1994] (Appendix A)

Secondly, the assumption that the Image Method produces reasonable room acoustic impulse responses was tested Room impulse responses generated by the Image Method were compared with actual room responses measured by Salava's impedance method The simulated impulse responses were found to capture correctly the main features of the reverberant process Hence, the technique of generating reverberant speech data by filtering noiseless, anechoic speech with impulse responses produced by the Image Method was established as being reasonably accurate

Regardless of its broad underlying assumptions, the general accuracy of the Image Method has been verified Moreover, the relationship between the reflection coefficient parameter and the sound decay of the resulting impulse response has been confirmed These findings support the use of the Image Method in room impulse response simulations Such simulations facilitate the synthesis of in-room sound fields for applications such as auditory experiments [Wattel et al , 1981, Culling et al , 1994] and hearing aid research [Kompis and Dillier, 1993]

CHAPTER 5

GLOTTAL CLOSURE DETECTION

5.1 INTRODUCTION

Automatic detection of the Glottal Closure Instant (GCI) is an important problem in speech science Correct detection of the GCI is necessary in order to determine period-by-period variations in the pitch of the speech signal This micro-melody carries phonemic, linguistic and speaker information Accurate reproduction of the micro-melody can improve the quality of speech coding and synthesis systems Furthermore, identification of the micro-melody can improve the accuracy of speech recognition strategies

As detailed in Chapter 3, GCI detection systems fall into two main categories Algorithms in the first category detect the presence of the closed phase, which occurs immediately after the GCI, by the high linear predictability of the region These methods tend to fail for certain vowels due to the presence of large residual pulses around the GCI They are also computationally complex Algorithms in the second category detect the GCI by the discontinuities or epochs associated with closure The main drawback with these methods is their sensitivity to ambient and excitation noise

One of the most promising methods for GCI detection is Maximum Likelihood Epoch Detection (MLED), as proposed by Cheng and O'Shaughnessy [Cheng and O'Shaughnessy, 1989] This method uses both the presence of a discontinuity and the linear predictability of the subsequent waveform to identify the GCI This makes the system robust to both noise and different types of voicing Unfortunately, as will be shown later, the system fails for certain vowels

This chapter proposes a re-formulation of the MLED technique called Pre-emphasised Maximum Likelihood Epoch Detection (PMLED) The new approach improves the accuracy of GCI identification and works for all voices and voicing types Furthermore, new pitch tracking and post-processing algorithms, which facilitate the use of the PMLED technique in a speech coding system, are detailed The performance of the new system for natural male and female speech in noise and reverberation is assessed

In the next section, Cheng and O'Shaughnessy's MLED technique is described Section three details why the method fails for certain voiced sounds Section four describes the new PMLED algorithm In section five the performance of the system is studied and section six concludes the chapter

5.2 MAXIMUM LIKELIHOOD EPOCH DETECTION

Maximum likelihood theory for epoch detection was initially developed for use in radar applications [Helstrom, 1960, Young, 1965] Cheng and O Shaughnessy adapted this theory for use in GCI detection They assumed that the speech signal within a pitch period is caused by a single pulse at the Glottal Closure Instant Assuming that speech production can be modelled as an all-pole linear system, then the wavelet due to the GCI can then be expressed as

$$s'(n) = \begin{cases} \sum_{i=1}^{p} a_i s'(n-i) & 0 < n \le N \\ G & n = 0 \\ 0 & \text{otherwise} \end{cases}$$

(5 1)

where G is an arbitrary constant and p is the order of the all-pole system They supposed that the difference between the observed signal $s(n+n_0)$ and the wavelet is a Gaussian process X with N independent observations, each with unit variance σ

$$X = S - S'$$

where

$$\begin{aligned} X &= [x(0) \quad x(1) \quad x(2) \qquad x(N-1)] \\ S &= [s(n_o) \quad s(n_o+1) \quad s(n_o+2) \qquad s(n_o+N-1)] \\ S' &= [s'(0) \quad s'(1) \quad s'(2) \qquad s'(N-1)] \end{aligned}$$

(5 2)

Maximum likelihood estimation states that the epoch occurs when the parameter values, $\Omega=\{a_1, a_2, a_p, n_0, \sigma\}$ maximise the conditional probability density or likelihood function

$$p(X/\Omega) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{ -\sum_{n=0}^{N-1} [s(n+n_o) - s'(n)]^2 \Big/ 2\sigma^2 \right\}$$

(5 3)

Cheng and O'Shaughnessy demonstrated that maximising the likelihood function as a function of n_0 is equivalent to maximising the cross-correlation of the observed signal and the wavelet Thus, the optimum closure point can be found as the maximum of

$$f'(n_o) = \sum_{n=0}^{N-1} s(n+n_o) s'(n)$$

(5 4)

where $f'(n_o)$ is referred to as the MLED signal Furthermore, they found that the AutoRegressive coefficients of the wavelet which produce the maximum likelihood function are the speech Linear Prediction coefficients produced by the autocorrelation method

$$\sum_{i=1}^{p} a_i \Phi(i-k) = \Phi(k)$$

where

$$\Phi(k) = \sum_{n=k}^{N_f-1} s(n) s(n-k)$$

(5 5)

and N_f is the frame size

Fig 5 1 shows typical speech and MLED signals for the vowel [a] The MLED signal can be seen to display a strong peak about 0-8 samples after the GCI Based on this empirical evidence, Cheng and O'Shaughnessy proposed that the best mark for the GCI was the 50 percent amplitude point on the rising edge of the strongest positive peak in the MLED signal

As can be seen in Fig 5 1, a number of weaker pulses occur close to the strong epoch pulse These represent sub-optimal epoch candidates In order to improve the strength ratio between the correct epoch pulse and the sub-pulses, Cheng and O'Shaughnessy proposed the use of a selection signal This signal is designed to have a symmetric and real spectrum which has its maximum amplitude at the

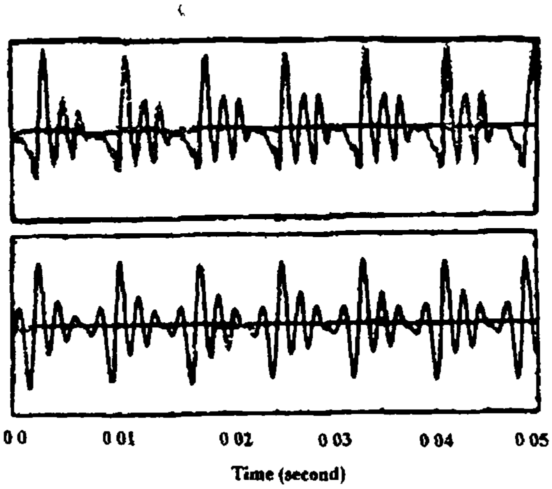


Fig 5.1 Typical signal in Maximum Likelihood Epoch Determination (MLED) for vowel [a] top - speech signal, bottom - MLED signal [after Cheng and O'Shaughnessy, 1989]

origin, gradually falling off with increasing frequency. They recommended use of the Hilbert Envelope of MLED signal

$$g(n_o)=\left[f'^2(n_o)+f_H'^2(n_o)\right]^{1/2} \tag{5.6}$$

where $f_H'(n_o)$ is the Hilbert Transform of the MLED signal $f'(n_o)$. The Hilbert Transform can be described as a filter [Ansari, 1987] with transfer function

$$H(\omega)=\begin{cases} -j & 0<\omega<\pi \\ 0 & \omega=0,\omega \\ j & -\pi<\omega<0 \end{cases} \tag{5.7}$$

and discrete-time impulse response

$$h(n)=\begin{cases} \dfrac{2\sin^2(\pi\ n/2)}{\pi\ n} & n\neq 0 \\ 0 & n=0 \end{cases} \tag{5.8}$$

This selection signal is made more pulse-like by average value subtraction

$$g'(n_o)=\begin{cases} g(n_o)-\overline{g(n_o)} & g(n_o)\geq\overline{g(n_o)} \\ 0 & g(n_o)<\overline{g(n_o)} \end{cases}$$

where

$$\overline{g(n_o)}=\frac{1}{N_f}\sum_{n_o=0}^{N_f-1}g(n_o) \tag{5.9}$$

The GCI Determination Signal (GCIDS) is calculated by multiplying the MLED signal by the selection signal

$$\theta(n_o)=f'(n_o)\ g'(n_o) \tag{5.10}$$

A block diagram of the overall system can be seen in Fig 5.2. The system implemented by Cheng and O'Shaughnessy operated at a sampling frequency of 10 kHz. An anti-aliasing filter with a 4.3 kHz

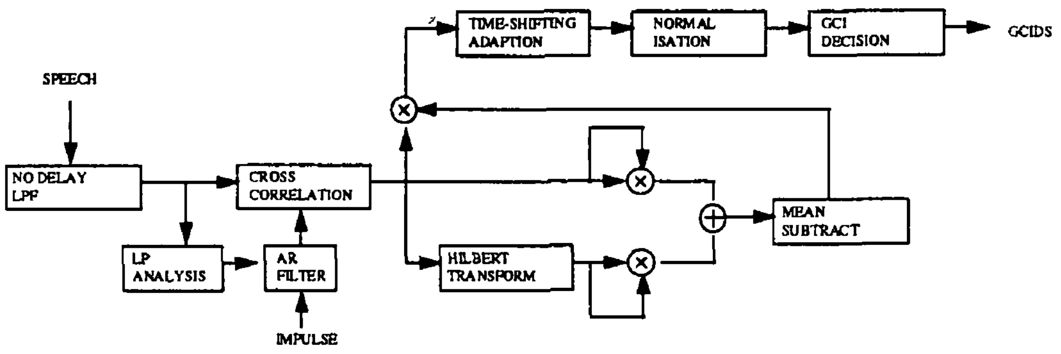


Fig 5 2 Block diagram of the GCI determination system [after Cheng and O'Shaughnessy, 1989]

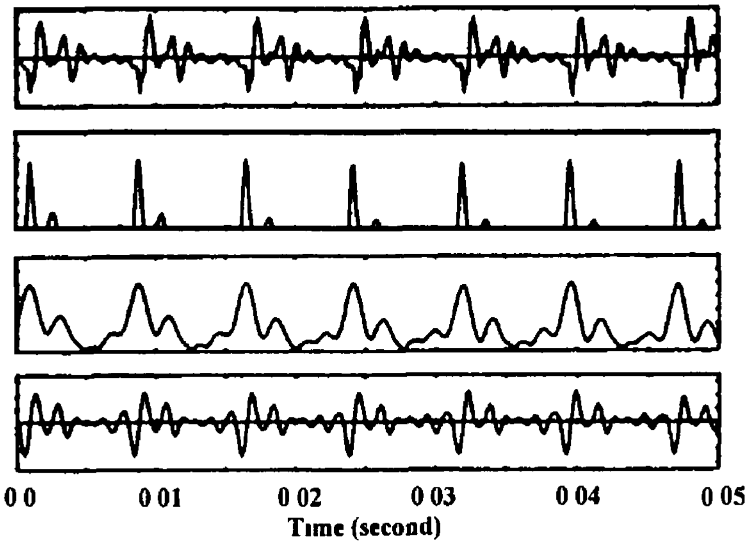


Fig 5 3 The results of GCI determination for the male vowel [a] from top to bottom - speech, the GCI determination signal (GCIDS), the selection signal, and the MLED signal [after Cheng and O'Shaughnessy, 1989]

cut-off frequency was applied before sampling and a no-delay, lowpass filter (NDLPF) after sampling The NDLPF was designed to decrease high frequency noise and had a cut-off frequency of 2 5 kHz with a gradual spectral roll-off An analysis frame length of 256 samples ($N_f=256$), with an overlap between successive frames of 56 samples, was chosen Twelve coefficients and a rectangular window were used in the Linear Prediction analysis The length of the wavelet was chosen to be 40 samples ($N=40$) Also, time-shifting adaptation was introduced to compensate for the time difference between the 50 percent amplitude point and the maximum amplitude point

Fig 5 3 shows the results obtained by Cheng and O'Shaughnessy in applying the systems to male speech The segment reproduced is for the vowel [a]

Cheng and O'Shaughnessy report good results for vowels, nasals, voiced fricatives and voiced plosives The system is also shown to be robust to white noise as well as to certain phase and amplitude distortions All of the tests conducted by Cheng and O'Shaughnessy used male speech

5.3 WHY MAXIMUM LIKELIHOOD EPOCH DETECTION FAILS

The MLED technique developed by Cheng and O'Shaughnessy has been found to fail for certain vowel sounds This section explains the causes of this failure

Firstly, consider a segment of male speech for which the method works well Fig 5 4 (a) shows the components of the MLED technique as calculated over a single analysis window Fig 5 4 (b) shows, superimposed on the magnitude spectrum of the speech signal, the spectral response of the 10th order all-pole LP filter estimated over the analysis window Fig 5 4 (c) shows the impulse response of the filter (i e the wavelet) which was used in calculating the MLED signal Note that 10th order LP analysis was used in this experiment since the speech signal was recorded at a sampling frequency of 8 kHz, see Appendix C

Examining the speech spectrum, Fig 5 4 (b), it can be seen that it rolls off with increasing frequency The roll-off is due to a combination of the -6 dB/octave de-emphasis characteristic of natural speech and the lowpass filtering effects of the NDLPF In order to model this spectrum, the LP filter allocates two poles to each formant and two poles to the roll-off These roll-off poles constitute a high energy low frequency resonance The resonance can be clearly seen to dominate the impulse response of the filter, Fig 5 4 (c)

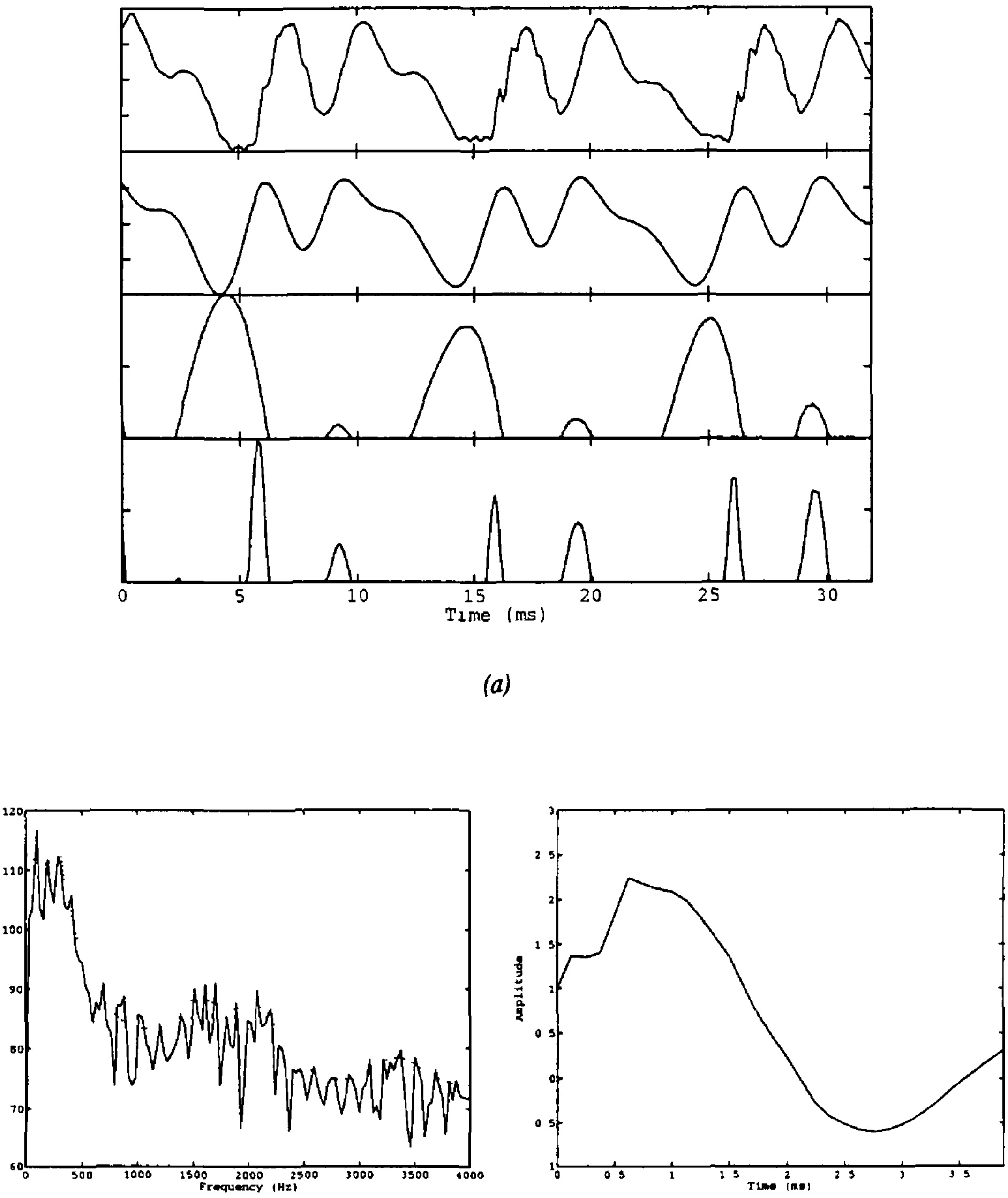
The MLED signal is calculated by cross-correlating the wavelet with the speech signal Due to the form of the wavelet, that is a broad positive peak followed by a broad negative peak, the MLED signal, Fig 5 4 (a) second panel, displays its maxima just before strong positive-to-negative transitions in the speech signal and its minima just before strong negative-to-positive transitions in the speech signal As a result, the MLED signal has its local minima just before the GCIs and its local maxima just after the GCIs

Now consider the selection signal The impulse response of the Hilbert transform filter is shown in Fig 5 5 Obviously, the output from this filter will show its maximum at the instant of strongest negative-to-positive transition in the input signal Therefore, the selection signal, calculated as the Hilbert envelope of the MLED signal, shows broad maxima around points of strong negative-to-positive transition in the MLED signal, see Fig 5 4 (a) third panel

The MLED and selection signals are multiplied to give the final GCIDS In the example, the broad maxima of the selection signal emphasise the negative-to-positive transitions in the MLED signal Thus, the GCIDS displays sharp peaks close to the GCI, see Fig 5 4 (a) fourth panel

Now consider a segment of speech for which the method fails Fig 5 6 shows the components of the MLED technique calculated over a single analysis window for the female vowel [ɪ] from "year" As before, the MLED signal shows its local maxima just before the instants of strongest positive-to-negative transition in the speech signal However, in this case, the point of strongest positive-to-negative transition does not occur immediately after the GCI Consequently, the final GCIDS does not correctly identify the GCIs

Since the speech signal always has a similar spectral roll-off, the all-pole LP model always contains a low frequency resonance As a result of this, the wavelet invariably consists of a broad positive peak followed by a broad negative peak Consequently, the overall MLED procedure identifies the instant of strongest transition with a pitch period As has been illustrated, the instant of strongest



(a)

(b)

(c)

Fig 5 4 The results of GCI determination for the male vowel [i] (a) from top to bottom - speech, MLED signal, selection signal and GCIDS (b) solid - magnitude spectrum of speech, dotted - spectrum of all-pole filter estimated over speech, (c) wavelet

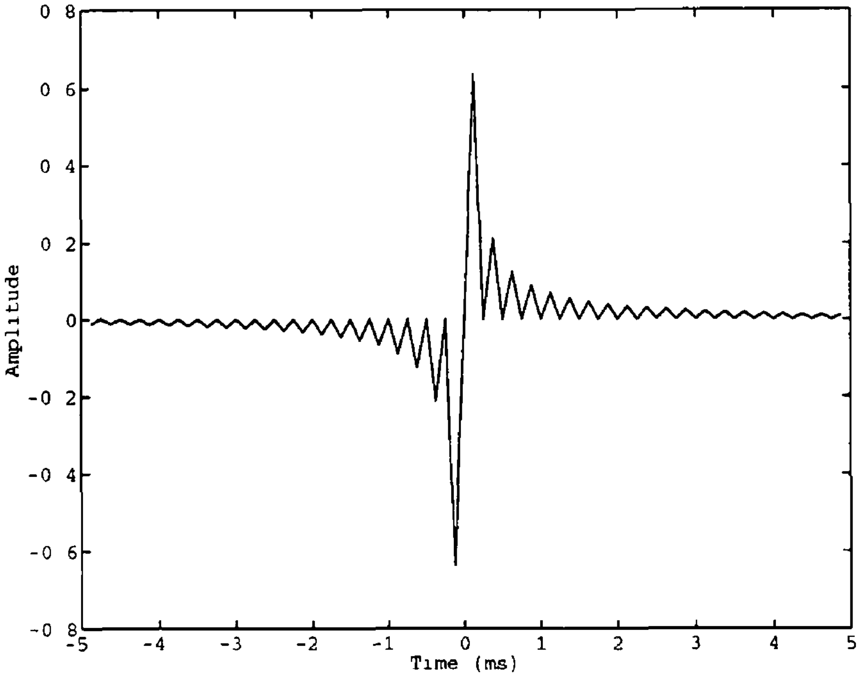
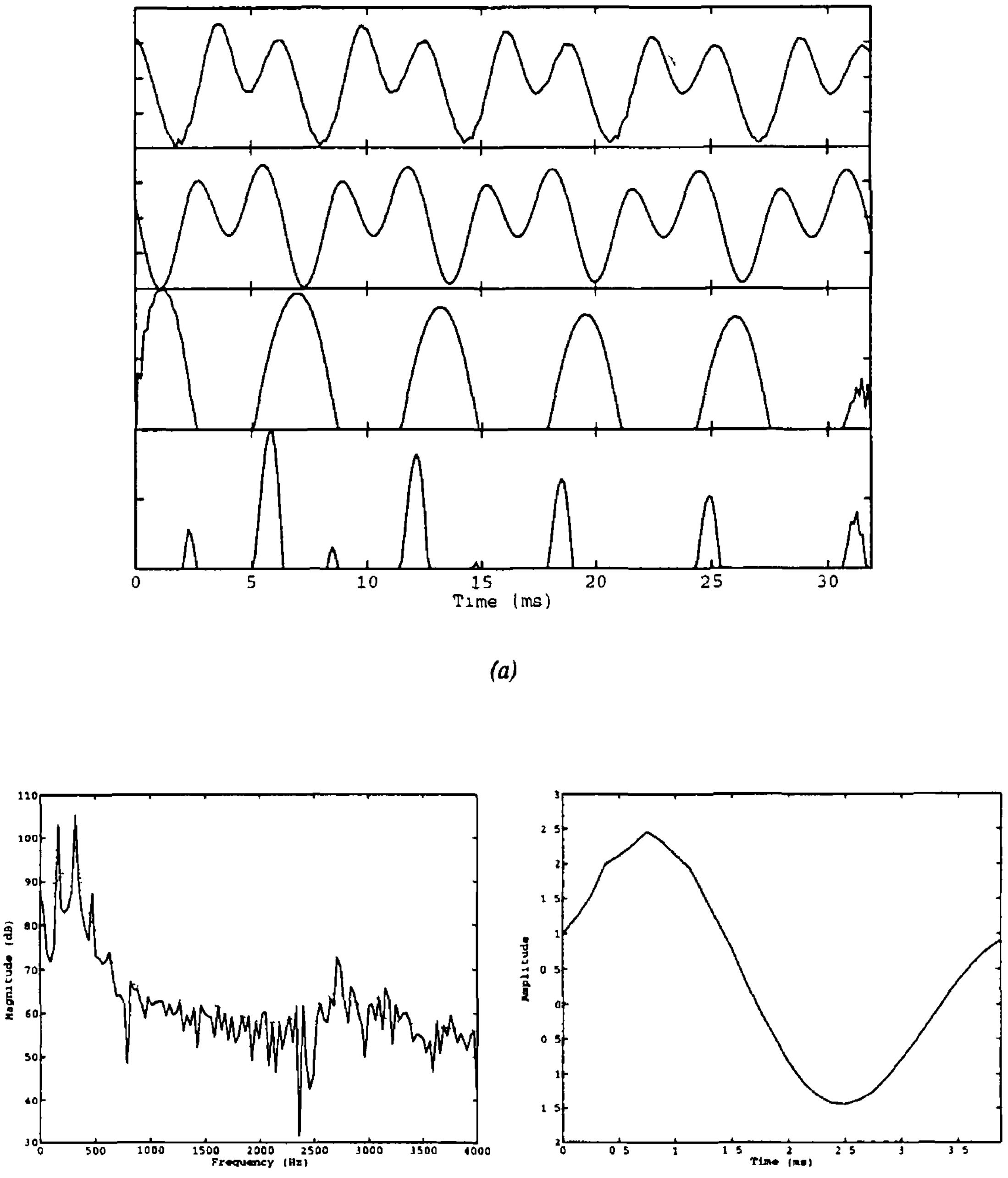


Fig 5 5 Impulse response of the Hilbert transform filter



(b) *(c)*

Fig 5 6 The results of GCI determination for the female vowel [ɪ] (a) from top to bottom - speech, MLED signal, selection signal and GCIDS, (b) solid - magnitude spectrum of speech, dotted - spectrum of all-pole filter estimated over speech, (c) wavelet transition does not necessarily coincide with the GCI. The re-formulation proposed in the next section corrects this problem by applying pre-processing to ensure that low frequency poles are excluded from the LP filter. This allows the new GCI detection algorithm to correctly identify the GCIs of the incoming speech signal.

5.4 PRE-EMPHASISED MAXIMUM LIKELIHOOD EPOCH DETECTION

This section describes the new reformulated MLED technique. The new method removes the spectral effects of the glottal waveform by applying a +6 dB/octave pre-emphasis filter to the speech signal. A standard one pole filter can be used for this purpose

$$H(z) = 1 - \frac{15}{16} z^{-1} \qquad (5.11)$$

In order to retain the timing information in the speech signal, the filter must be pass forwards and backwards across the speech signal. This ensures that the overall filtering operation has zero phase.

A schematic diagram of the overall system is shown in Fig. 5.7. The MLED signal is calculated as before, except for the lowpass filtering operation (NPLPF) which is removed. Since the spectral roll-off of the speech signal has been cancelled, only an 8th order all-pole filter is needed to model the formant resonances. Furthermore, in calculating the wavelet to be used in the cross-correlation operation (Eq. 5.1), the constant G is chosen to be negative. This is because the glottal closure pulse, whose position is to be determined, is itself negative. In order to improve the contrast of the resulting MLED signal, a mean subtraction is applied and negative samples are set to zero. This new MLED signal is termed the Pre-emphasised Maximum likelihood Epoch Detection (PMLED) signal. Fig. 5.8 shows the speech signal, pre-emphasised speech, wavelet and the PMLED for a typical male vowel.

A pitch detection algorithm is used to select the true GCI from the candidate pulses provided by the PMLED technique. A lowpass filter with sharp cut off at 1 kHz is passed, forwards and backwards, across the speech and PMLED signals. This removes high frequency noise and improves pitch determination. Next, autocorrelation is performed on the filtered speech and PMLED signals. The delays associated with the maxima of these autocorrelation functions are candidate pitch periods. The two candidates and the pitch period from the previous window are then compared. If two or more of the candidates match to within 10 % then that pitch is chosen as the period length for this window. If no candidates match, the values of the two autocorrelation functions at the three candidate periods are multiplied. The candidate pitch period with the largest product is chosen as the pitch.

Once the pitch has been determined, the delay between the start of the window and the first GCI must be found. To do this, a matched pulse train of broad peaks at the pitch period is cross-correlated with PMLED signal. The resulting cross-correlation function displays a peak at the optimum lag between the matched pulse train and the PMLED signal. The PMLED signal is multiplied by the appropriately lagged matched pulse train to give the Pre-emphasised GCIDS (PGCIDS). The instant of the maximum in the PGCIDS within each peak of the matched pulse train is then marked as the instant of glottal closure (see Fig 5.8).

For high pitched speech (F_0 > 2 kHz) significant pitch drift may occur within a single analysis window. Therefore, when high pitched speech is encountered, the pitch and lag determination

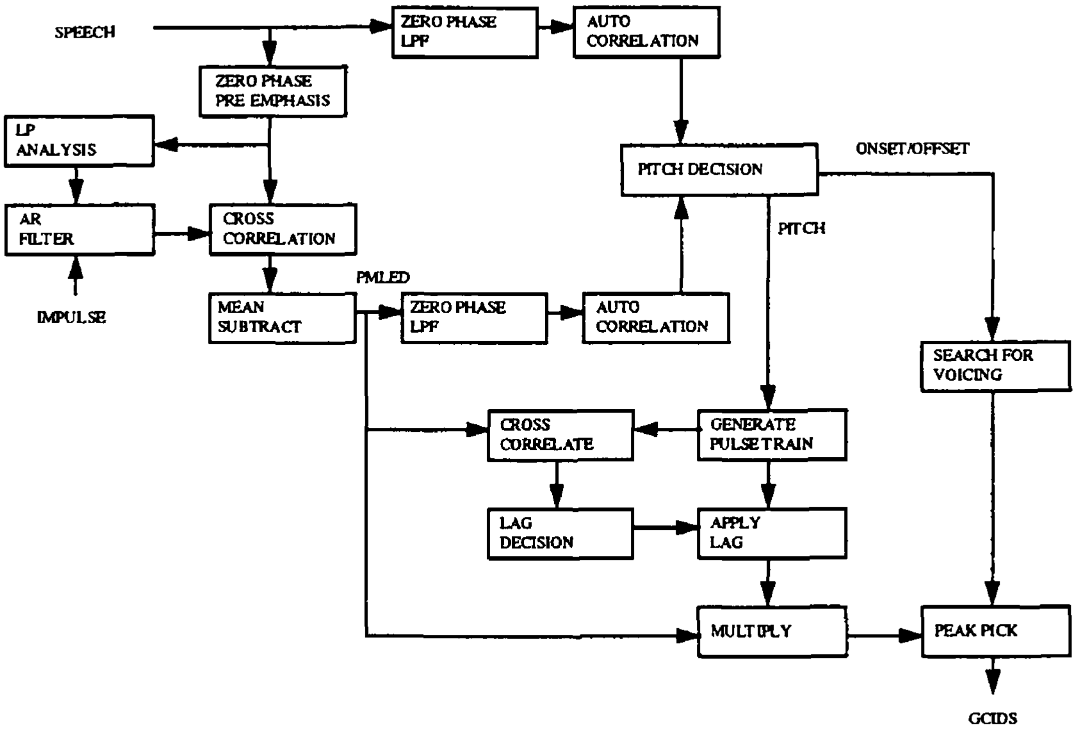
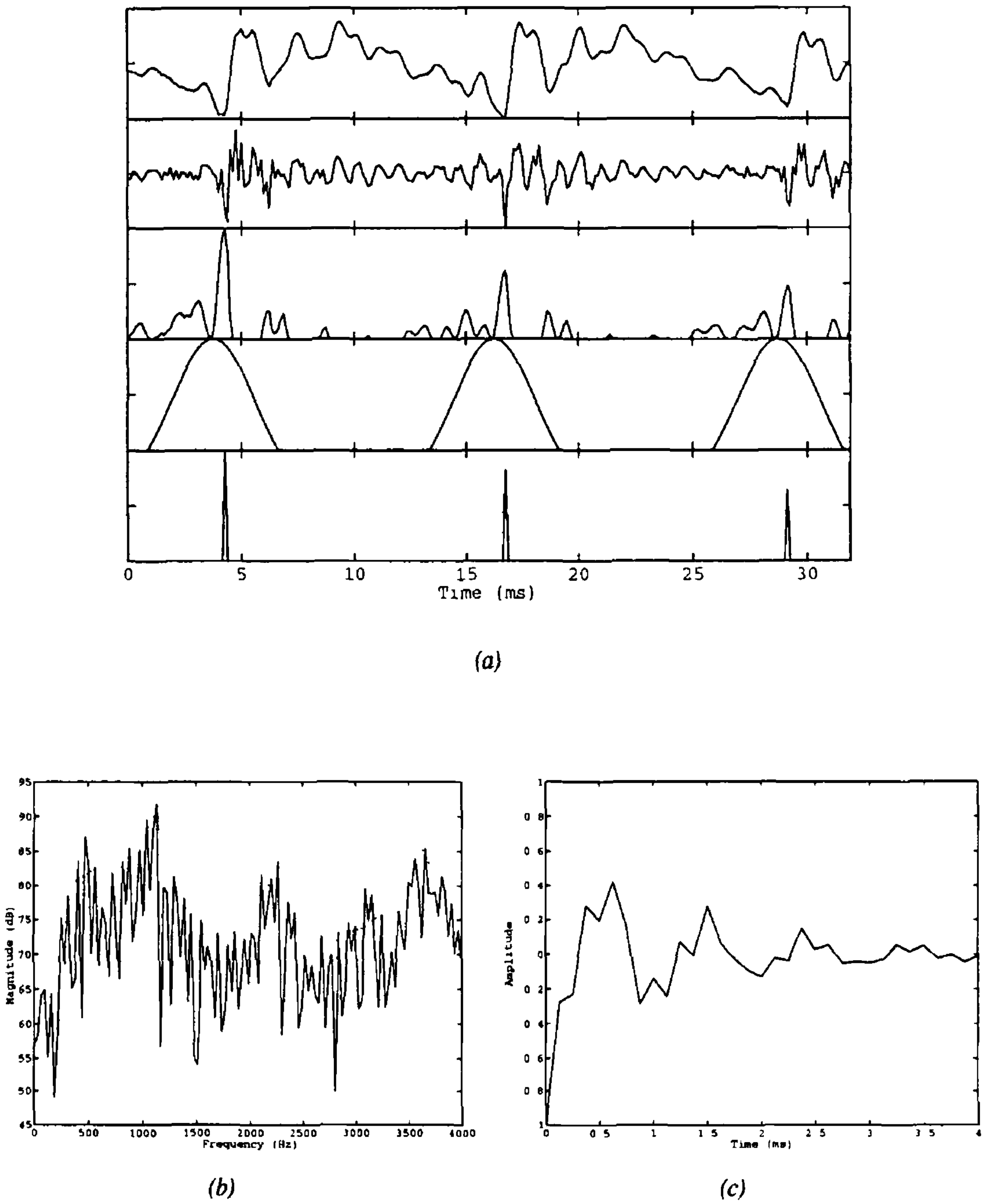


Fig 5 7 Block diagram of the PMLED system.



(a)

(b)

(c)

Fig 5 8 Typical results in PMLED for male vowel [ə] (a) from top to bottom - speech signal, pre-emphasised speech signal, PMLED signal, matched pulse train and PGCIDS, (b) solid - spectrum of the pre-emphasised speech signal, dotted - spectrum of the all-pole filter estimated over the pre-emphasised speech signal, (c) wavelet

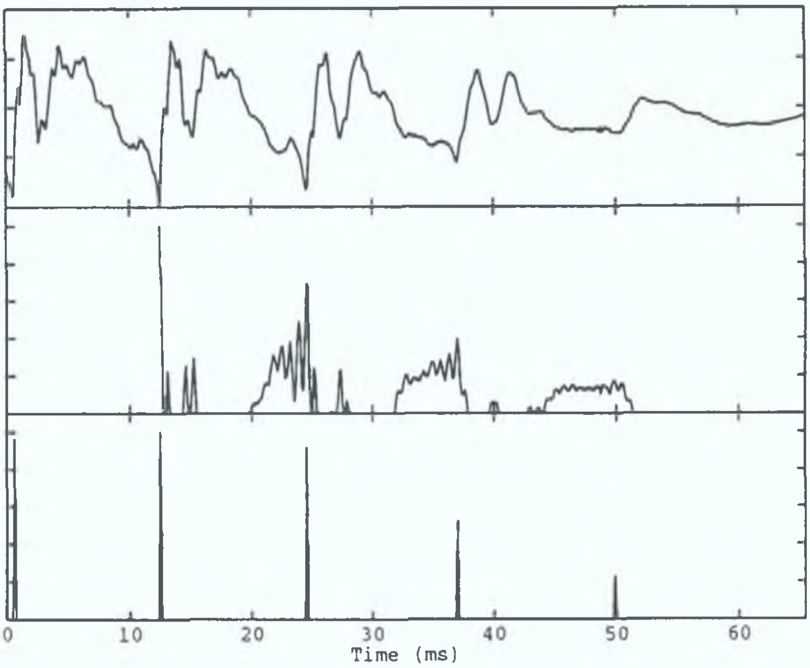


Fig. 5.9. Typical results in PMLED for male voiced constant offset [r]: from top to bottom - speech signal, PMLED signal and PGCIDS.

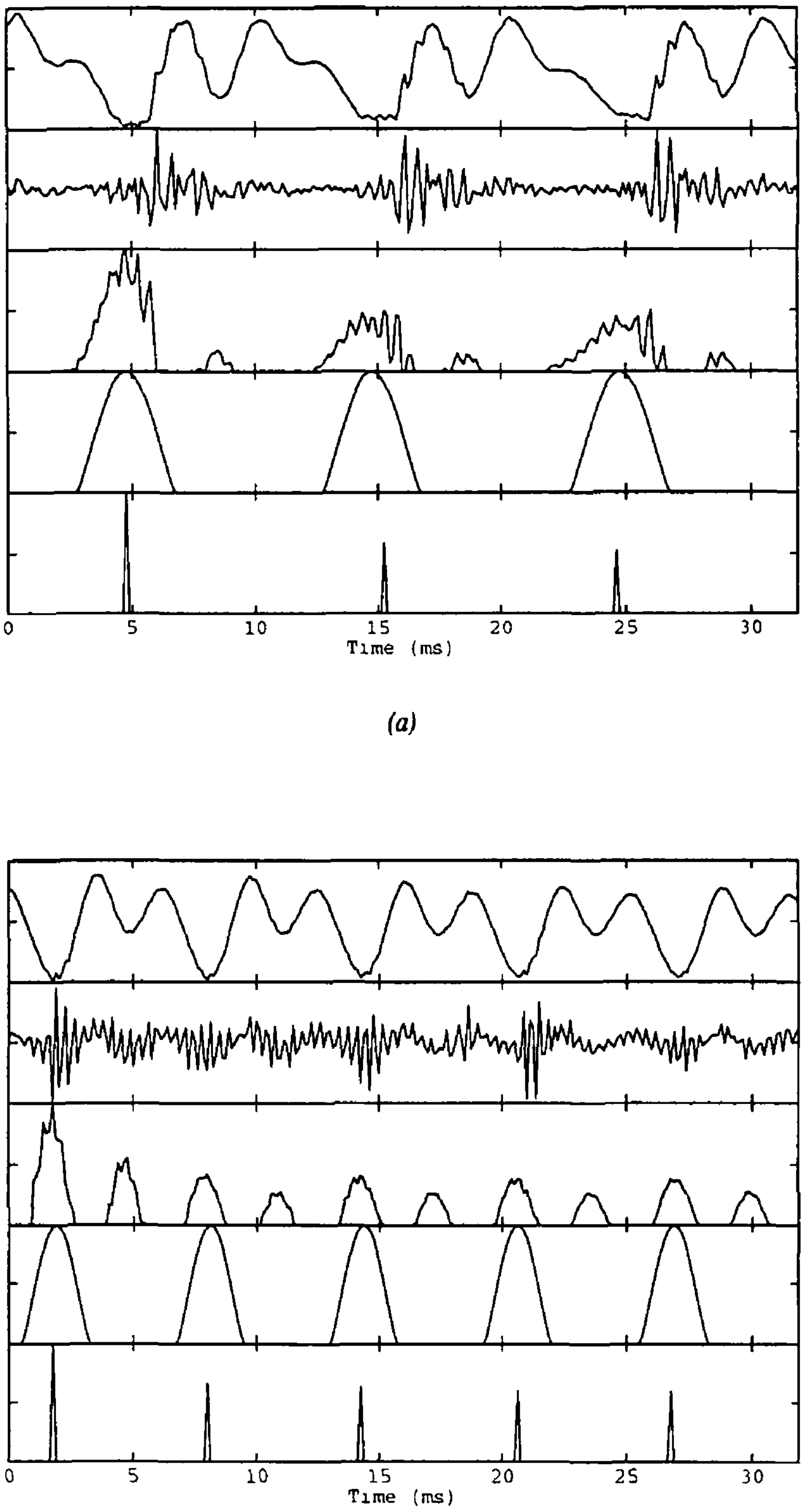
algorithms are applied to half of the analysis window at a time. This allows for more accurate pitch tracking and GCI determination.

Voicing onsets and offsets present a special problem for GCI determination. At these times, voicing is weak and the pitch period may change vary rapidly. Thus, the pitch estimation algorithm may produce ambiguous results. In the case of offsets, the system searches backwards for the nearest occurrence of strong voicing. PMLED is applied from the last closure in the strongly voiced segment to the beginning of the silence region. A search is conducted across the PMLED signal for peaks at, or below, the pitch of the strongly voiced region. For voicing onsets the technique is similar, except that the system searches forwards for the nearest strong voicing and backwards for the onset closures. Fig. 5.9 shows this technique applied to a typical voicing offset.

5.5 PERFORMANCE STUDY AND RESULTS

The PMLED method was tested for different voicing types under noise and reverberation. Input speech was anti-aliased using a zero-phase filter with a 3.8 kHz cut-off and sampled at 8 kHz, for more details see Appendix C. The analysis frame length was chosen to be 256 samples, $N_f=256$, with a 56 sample overlap between frames. Eighth order Linear Prediction ($p=8$) and a wavelet 32 samples in length ($N=32$) were used in all of the experiments.

Fig. 5.10 shows the results of the PMLED technique applied to male and female vowels. The vowel segments are the same as those tested on the previous MLED system, Figs. 5.4 and 5.6.



(a)

(b)

***Fig 5 10** The results of PMLED, (a) male vowel [i], (b) female vowel [i], from top to bottom - speech signal, pre-emphasised speech signal, PMLED signal, matched pulse train and GCIDS (Note, the vowel segments are the same as those appearing in Figs 5 4 and 5 6)*

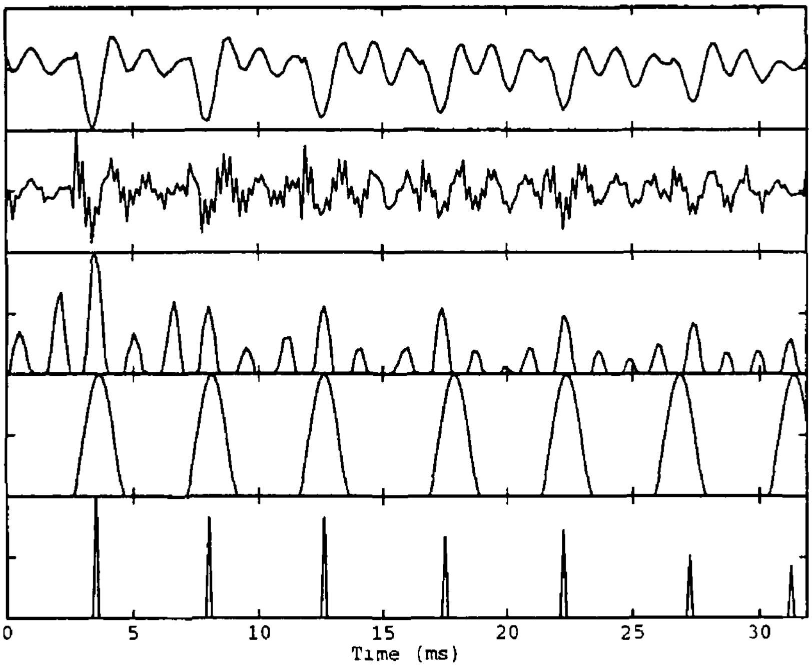


Fig 5 11 The results of PMLED for male vowel [u] from top to bottom - speech signal, pre-emphasised speech signal, PMLED signal, matched pulse train and GCIDS

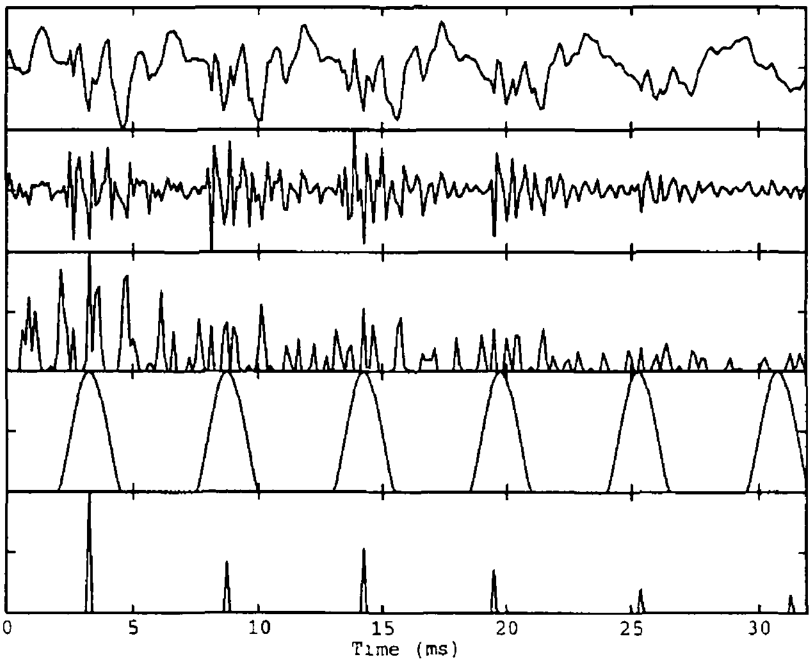


Fig 5 12 The results of PMLED for male voiced fricative [v] from top to bottom - speech signal, pre-emphasised speech signal PMLED signal, matched pulse train and GCIDS

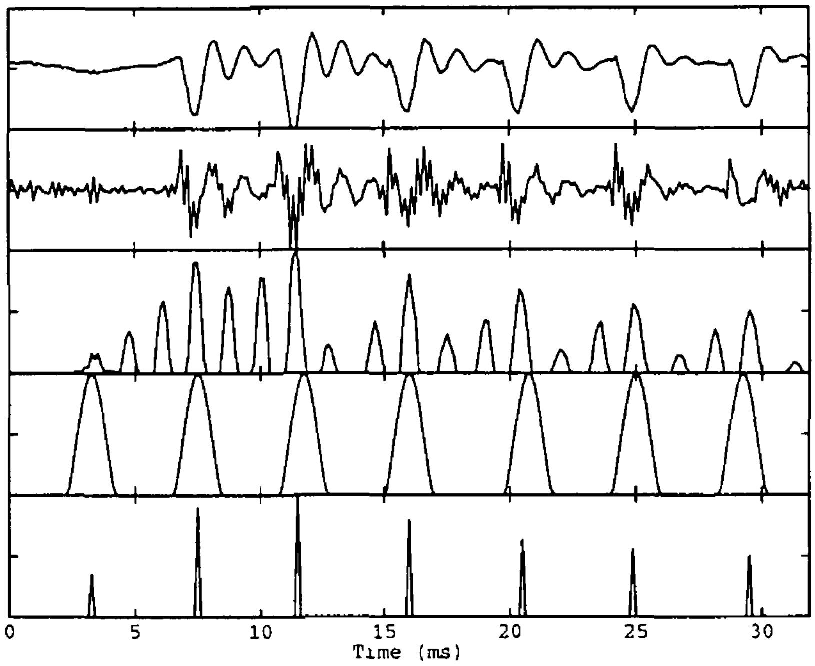


Fig 5 13 The results of PMLED for male voiced plosive [b] from top to bottom - speech signal, pre-emphasised speech signal, PMLED signal, matched pulse train and GCIDS

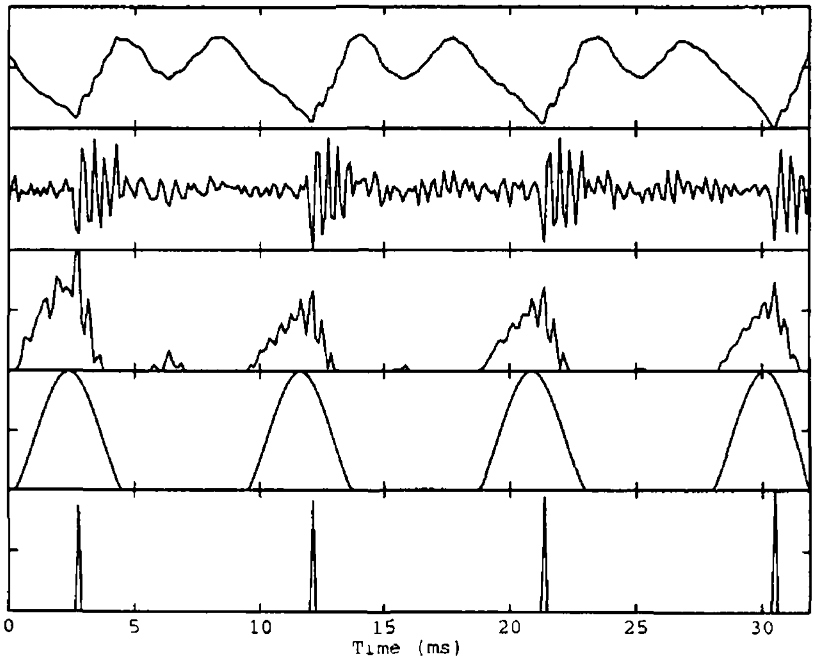


Fig 5 14 The results of PMLED for male nasal consonant [n] from top to bottom - speech signal, pre-emphasised speech signal, PMLED signal, matched pulse train and GCIDS

On examining Fig. 5.10 (a), it can be seen that the speech signal displays a flat negative peak around the GCIs. This makes it difficult to manually determine the exact instant of closure. Similarly, the PMLED technique has difficulty, showing a number of equal peaks representing candidate closure points. The matched pulse train selects the most promising peaks based on the estimated pitch period and pitch lag. Comparing Figs. 5.10 (a) and 5.4, it is clear that PMLED has chosen GCIs slightly earlier than those selected by MLED. In this case, it is unclear which algorithm is the more accurate. However, the general character of both GCIDS is correct.

In Fig. 5.10 (b) it is obvious that the new system has correctly determined the instants of glottal closure. This is in stark contrast to the performance of the MLED technique, see Fig. 5.6. On examining Fig. 5.10 (b) in detail, it can be seen that, in this case, accurate pitch determination was essential in removing secondary peaks from the PMLED signal.

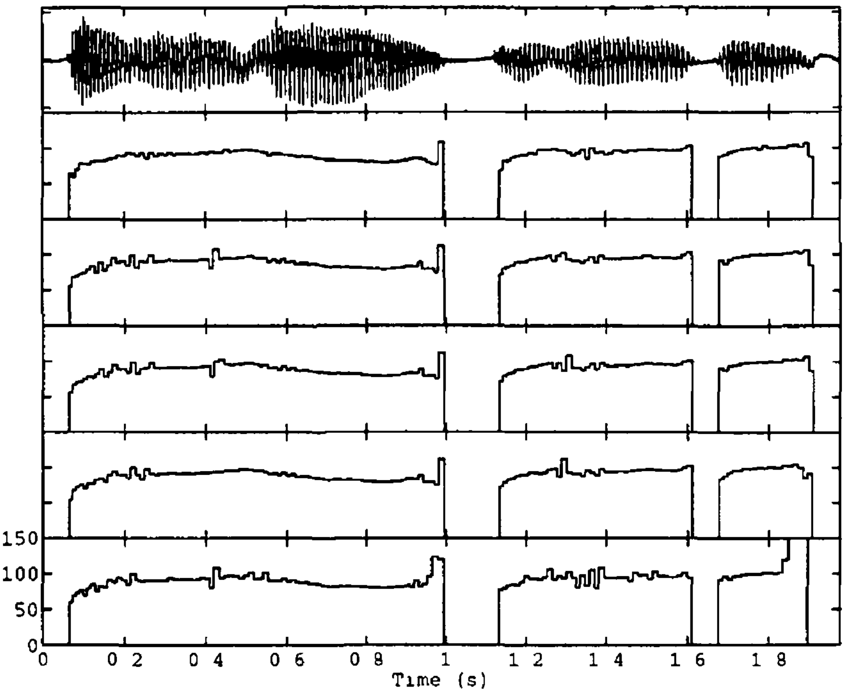
Performing accurate glottal closure identification on the vowel [u] has proven difficult [Strube, 1974; Ananthapadmanabha and Yegnanarayana, 1979]. Most epoch detection methods try to detect the high frequency energy associated with a waveform discontinuity. The vowel [u], however, contains little energy at these frequencies, most being concentrated at low frequency. MLED was one of the first algorithms to succeed for this vowel. Fig. 5.11 shows the performance of PMLED for a male [u]. The method correctly identifies all the closure points, even though the amplitude of the vowel is steadily decreasing.

Voiced fricatives are another class of sound for which the MLED method works well. In this type of speech, voiced and unvoiced excitation occur simultaneously. This makes it difficult to determine the excitation point associated with glottal closure. Fig. 5.12 shows PMLED applied to the consonant [v]. Again, the closure point is difficult to determine manually. Comparing the speech signal and the GCIDS, it can be seen that the marked GCIs occur before the maximum negative peak in the speech signal. However, comparing the pre-emphasised speech and the GCIDS, it is clear that the chosen GCIs are actually at the points of maximum energy innovation to the vocal tract. Thus, the PMLED system has correctly identified the glottal closure points. Again, the importance of accurate pitch determination is obvious.

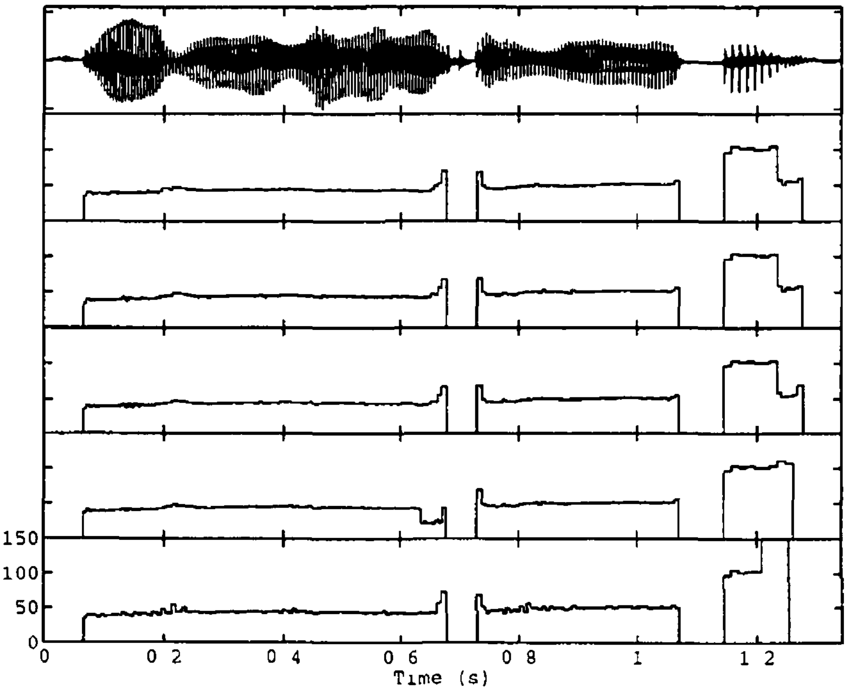
Voiced plosives present difficulties for conventional pitch detection algorithms. These sounds are characterised by a sudden burst of energy at the start of the consonant. In addition, there are often associated rapid changes in pitch. Fig. 5.13 shows the results obtained from applying PMLED to the voiced plosive [b]. Strong peaks in the PMLED signal clearly indicate the instants of glottal closure.

Another category of sound which can cause problems for glottal closure detection algorithms is nasal consonants. This is due to the presence of zeros in the speech spectrum caused by anti-resonances in the nasal tract. The results of PMLED applied to the nasal [n] are shown in Fig. 5.14. Obviously, the performance of PMLED is unaffected by nasal coupling.

To be of use in speech coding applications, a GCI detection algorithm must be accurate over prolonged segments of voiced speech and robust to noise. Fig. 5.15 (a) and (b) show the pitch contours produced by PMLED when applied to the sentence "We were away a year ago" as spoken by a male and a female subject, respectively. The top panel shows the recorded speech signal. The second panel displays the manually determined pitch contour. The third panel shows the pitch contour produced by

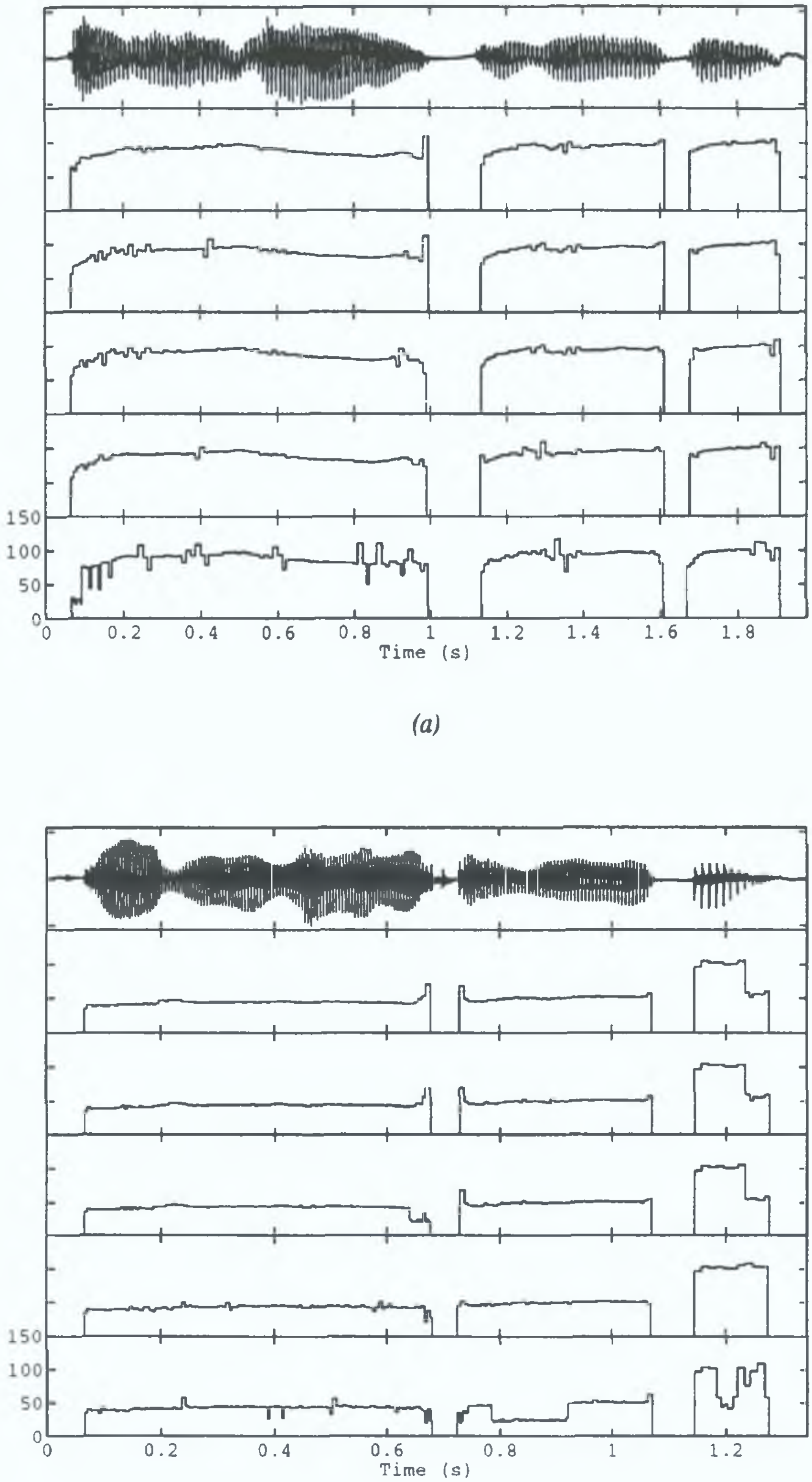


(a)



(b)

Fig 5 15 Pitch contours for the sentence "We were away a year ago" with added noise (a) male subject, (b) female subject from top to bottom - speech, manually estimated pitch contour, PMLED estimated pitch contour using clean speech and speech with SNR of 35 dB 25 dB and 15 dB



(a)

(b)

Fig. 5.16. Pitch contours for the sentence "We were away a year ago" with added reverberation: (a) male subject; (b) female subject: from top to bottom - speech, manually estimated pitch contour, PMLED estimated pitch contour using clean speech and speech with simulated source-receiver distance of 10cm, 30cm and 50cm.

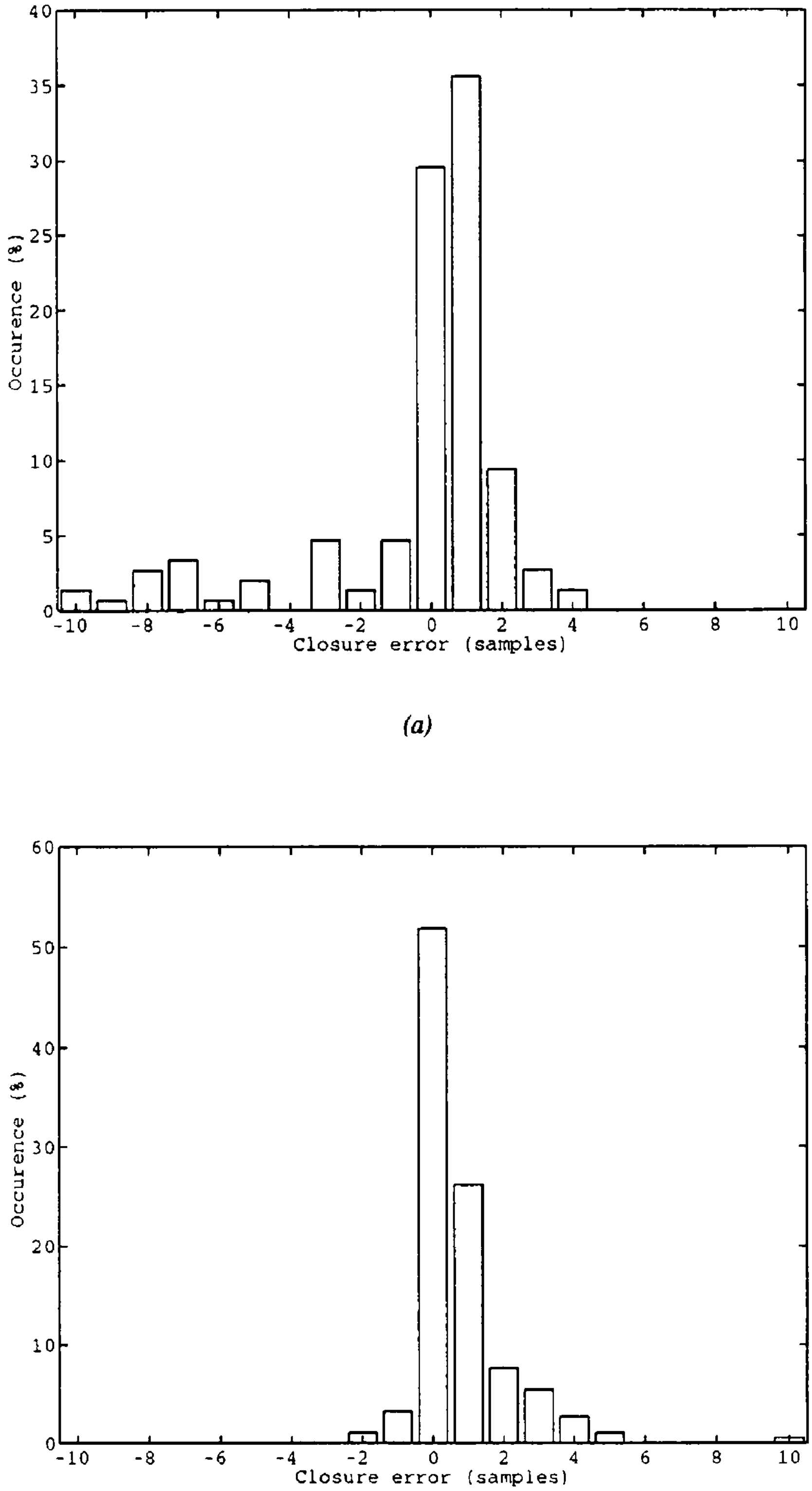
PMLED applied to the clean speech signal The fourth, fifth and sixth panels show the pitch contours calculated by PMLED over the speech signal with added noise

The accuracy of the new system can be clearly seen by comparing the manual pitch contour and that produced by PMLED on the clean speech (second and third panels) The pitch contours match to a high degree of accuracy The sudden changes in pitch at voicing onsets and offsets are particularly well identified This is difficult to achieve with conventional pitch prediction algorithms In the case of the male segment, some dither can be seen in the early part of the PMLED generated pitch contour (Fig 5 15 (a) third panel) It must be noted that this region corresponds to the [ı] vowel of the first word "we ' As can be seen in Fig 5 10 (a), the closure points in this region are difficult to determine Thus, some deviation from the manually marked GCIs must be expected

The performance of glottal closure detection algorithms tends to deteriorate in the presence of white noise Noise has a twofold effect Firstly, discontinuities in the noise signal can be misinterpreted as epochs Secondly, noise reduces the predictability of the overall signal and so can disguise the AutoRegressive decay which follows a glottal closure Fig 5 15 panels three, four and five show the pitch contour produced by PMLED applied to speech segments with Signal to Noise Ratios (SNRs) of 35dB, 25dB, and 15 dB, respectively In general, the PMLED algorithm remains reasonably robust with increasing noise The dither of the identified pitch increases slowly with the noise In addition, pitch halving and doubling occurs at some of the voicing offsets It must be remembered that the quoted SNRs are average values calculated across the entire segments As a result, the actual SNRs at the offsets are, in fact, much lower than the nominal values given above Using a more complex expert system for the pitch decision might alleviate this problem

Reverberation creates a special problem for GCI detection systems Generally, sound reflections have properties very similar to the original signal Thus, a reflected glottal pulse can be easily misinterpreted as a new pulse coming directly from the sound source In particular this presents great difficulties at voicing offsets At these points the energy of the direct signal is low, while the reflected energy is high Fig 5 16 panels three, four and five show the pitch contour produced by PMLED applied to speech segments with simulated reverberation equivalent to source-receiver distances of 10cm, 30cm and 50cm, respectively The performance of the PMLED algorithm deteriorates rapidly with increasing source-receiver distance At a distance of 50cm the technique is unsatisfactory for speech coding purposes However, this is well beyond the distances normally used for telephone speech The problem could be alleviated somewhat by the use of a directional microphone This would reduce the level of the reverberant sound energy relative to the direct

To further investigate the accuracy of the new system, the temporal differences between the GCIs identified by PMLED applied to the clean speech and those marked manually were calculated Fig 5 17 shows histograms of the error in closure instant detection using the PMLED system The tests indicate that the PMLED identified all of the closures marked manually, for both the male and female segments Furthermore, it was found that over 75% of the PMLED estimated GCIs were within ±2 and ±1 samples of the manually identified GCIs for the male and female segments, respectively The better accuracy in the case of the female segment can be attributed to the higher fundamental frequency of the speech In general, this means that the closure points are associated with a sharper negative peak in the speech



(a)

(b)

Fig 5 17 Percentage occurrence of GCI identification error for PMLED applied to the noiseless recordings (a) male subject, (b) female subject

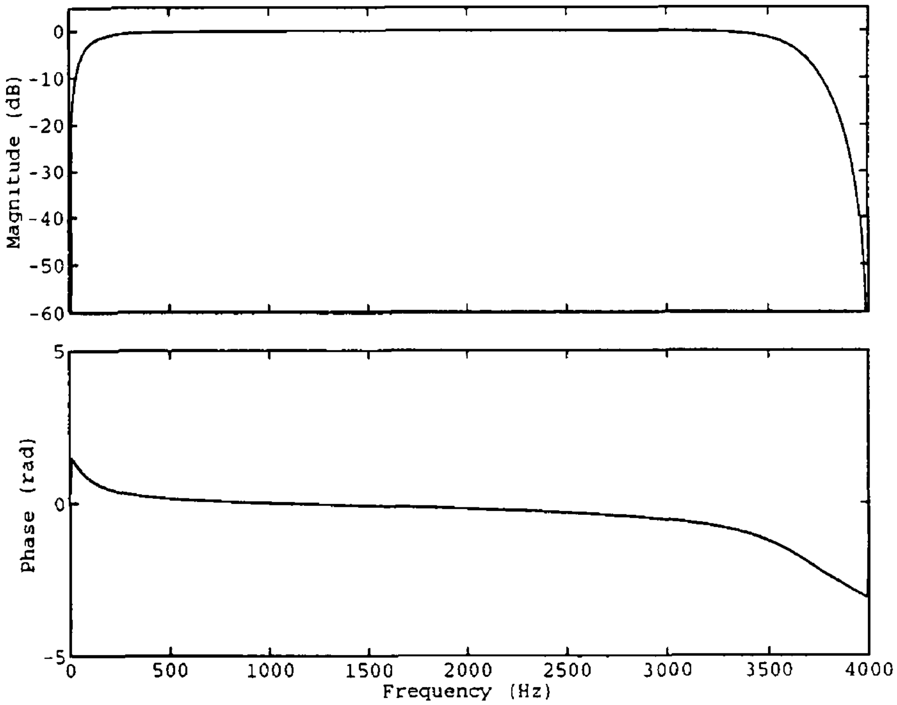
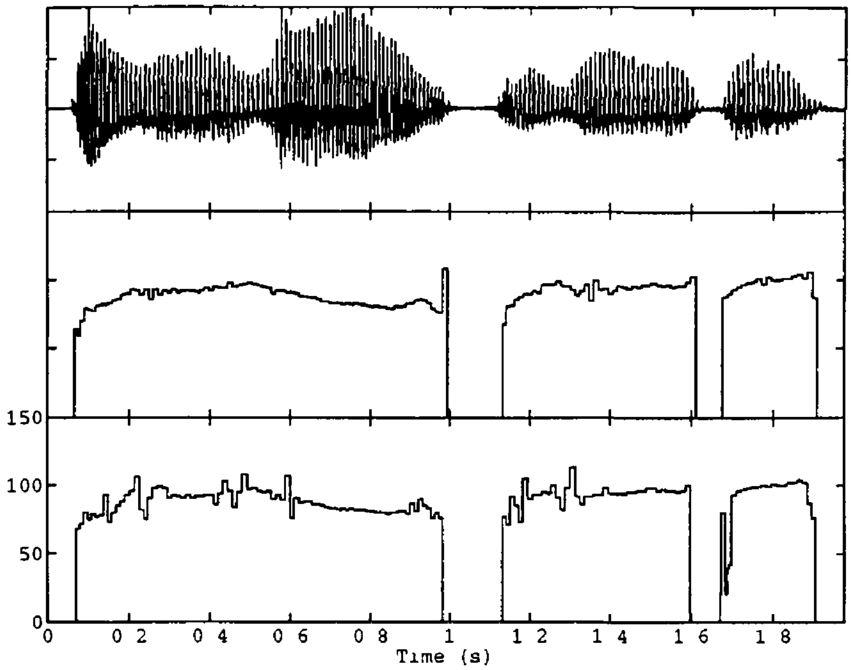
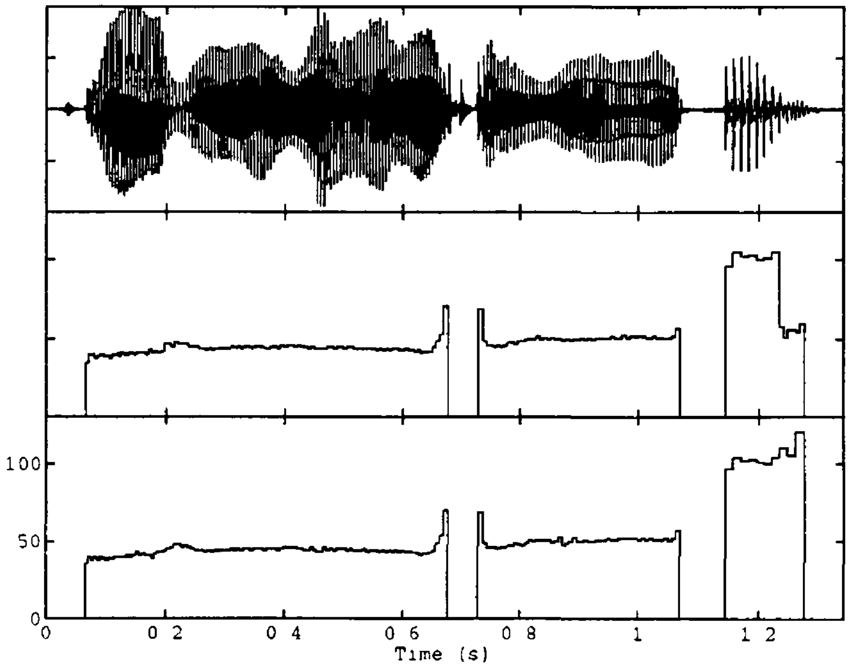


Fig 5 18 Bandpass filter top - magnitude spectrum, bottom - phase spectrum.

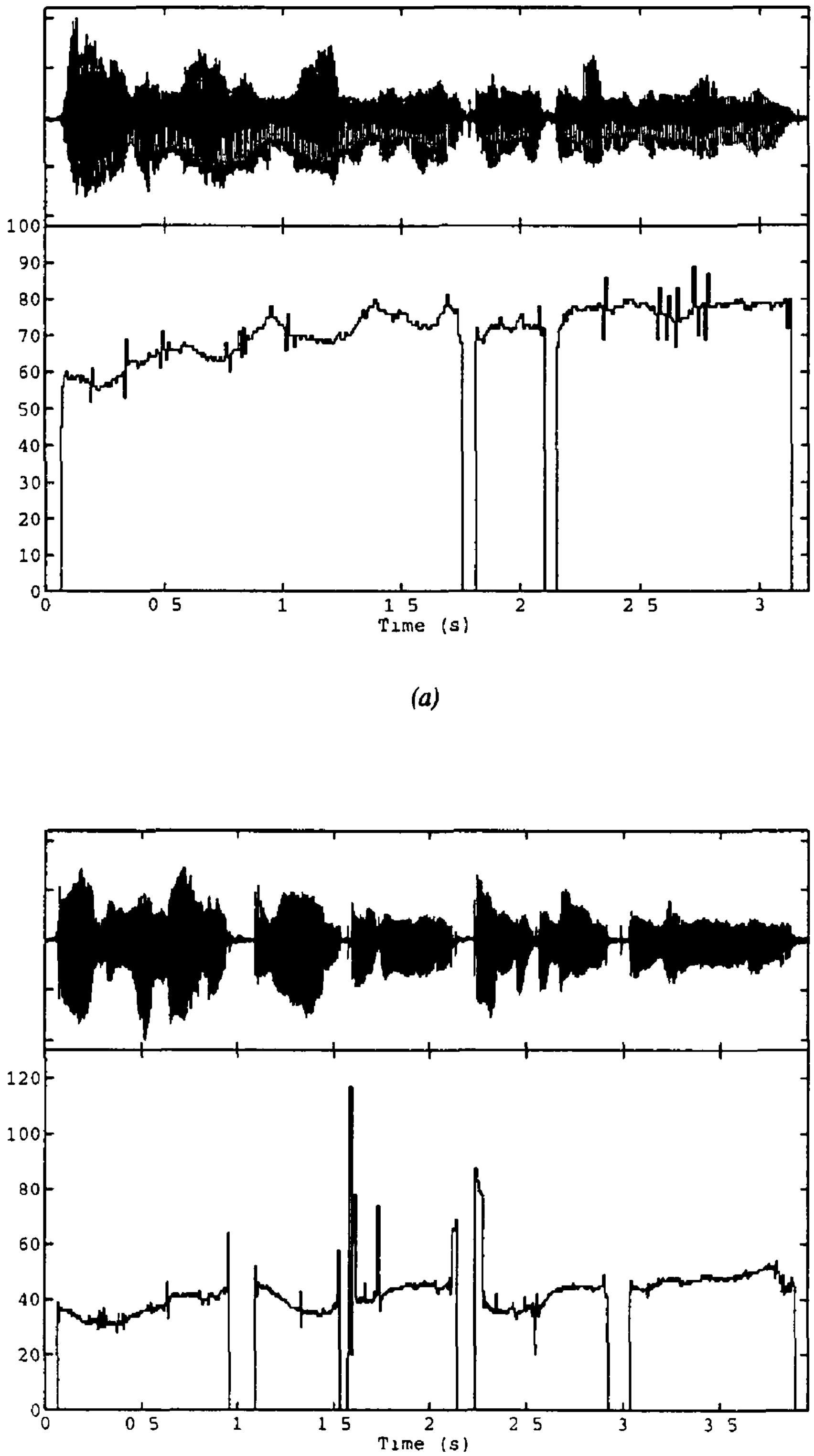


(a)



(b)

Fıg 5 19 Pıtch contours for the bandpass filtered noıseless recordıngs (a) male subject, (b) female subject from top to bottom - speech, manually estımated pıtch contour and PMLED estımated pıtch contour



(a)

(b)

Fig 5 20 Pitch contour estimated using PMLED for the office recordings (a) male subject, (b) female subject top - speech, bottom - estimated pitch contour signal Consequently, the GCIs are easier to identify precisely It must also be noted that the accuracy of the PMLED algorithm may be greater than is indicated by these results When marking the GCIs manually, the operator tends to place the closure point at the minimum of the negative peak in the speech signal associated with the closure Often, this is not the point of maximum vocal tract excitation Thus, the manual marking system may be biased in its estimation of the GCIs This may explain the +1 skew visible on Fig 5 17 (a)

To investigate the performance of the algorithm on narrowband speech, a typical bandpass filter was applied to the segments before processing with PMLED The characteristics of the filter are shown in Fig 5 18 The filter was of Butterworth design, consisting of a first order highpass section, with cut-off at 100 Hz, and a second order lowpass section, with cut-off at 3600 Hz The results of PMLED performed on the filtered speech segments are shown in Fig 5 19 As might be expected, the performance of the algorithm on the male segment is reduced For the male segment, significant low frequency energy is removed by the filter Also, the filter introduces considerable phase distortion at these frequencies Thus, much of the pitch information in the speech signal is removed or distorted This leads to inaccurate identification of the GCIs and an increase in pitch dither In contrast, the algorithm performs well on the filtered female segment In this case, the only significant difference from the manually marked pitch contour is the pitch doubling at the final offset. On re-examination of the speech signal at this point, it is unclear which pitch value is correct In all likelihood, either pitch would produce good re-synthesis in a coding system

Finally, the new system was applied to segments of male and female speech recorded in a normal office environment The pitch contours calculated by PMLED over the segments "Early one morning a man and a woman ambled along a one mile lane" are shown in Fig 5 20 The results show reasonable pitch contours in both cases Some dither can be seen but it rarely exceeds ±10 % Particularly well captured in the female segment are the longer than normal pitch periods occurring at voicing onsets and offsets

5.6 CONCLUSION

This chapter has proposed a re-formulation of an existing algorithm for GCI detection from the speech signal The deficiencies of the previous method have been explained and the improved accuracy of the new approach illustrated The performance of the new method has been analysed for various types of voicing under noise and reverberation The results have shown that the method is both accurate and robust to noise Furthermore, the method is reliable in normally reverberant conditions up to source-receiver distances of approximately 50cm

The new algorithm provides a more accurate and robust method for glottal closure detection This facilitates the use of pitch-synchronous analysis techniques under normal recording conditions These techniques, in turn, provide greater accuracy and reliability in the analysis of voiced speech One such technique, inverse filtering for extraction of the glottal excitation is examined in the next chapter

CHAPTER 6

GLOTTAL WAVEFORM EXTRACTION

6 1 INTRODUCTION

This chapter assesses the performance of two algorithms for glottal waveform estimation The techniques under investigation are Closed Phase Inverse Filtering [Berouti, 1976, Wong et al , 1979, Deller, 1981] and Pitch Synchronous Iterative Adaptive Inverse Filtering [Alku, 1992a,b,c] Closed Phase Inverse Filtering (CPIF) is a well established method, having been studied in a number of investigations [Krishnamurthy and Childers, 1986] This chapter aims to determine its robustness to noise and reverberation, and to determine its suitability for use in a glottal excited speech coding system Iterative Adaptive Inverse Filtering (IAIF) is a newer technique, developed by Alku The method shows promise in that, unlike Closed Phase Inverse Filtering, it does not require *a priori* identification of the Glottal Closure Instant Thus, it is likely to be more robust than the closed phase procedure This chapter compares the performance of the two algorithms under various conditions of noise and reverberation

For purposes of speech transmission, the estimated glottal waveform must be parameterised The approach chosen in this investigation is parameterisation by fitting of a time-domain glottal waveform model One of the most successful models, the LF model [Fant et al , 1985] is used for this purpose The LF model has shown good results in glottal waveform analysis [Gobl, 1988, 1989] and in speech synthesis experiments [Childers et al , 1987, Childers and Wu, 1990, Carlson et al , 1990] The accuracy of the LF model in representing the estimated glottal waveforms is examined in this chapter The effects of noise and reverberation on the accuracy of the proposed fitting procedure are also assessed

This chapter in split into five sections Section two describes the inverse filtering and glottal fitting algorithms The third section details the experiments carried out to determine the performance of the algorithms Section four discusses the LF data obtained during the experiments in the light of previous studies on voice source dynamics Lastly, section five concludes the chapter

6.2 DESCRIPTION OF THE SYSTEMS

This section describes the algorithms under investigation in this chapter The first sub-section describes the CPIF algorithm The second sub-section explains the IAIF procedure The third and final sub-section describes the techniques used to fit the LF model to the glottal waveforms estimated by the inverse filtering algorithms

6.2.1 Closed Phase Inverse Filtering

Closed Phase Inverse Filtering operates on the assumptions that for a few milliseconds after glottal closure the glottis is closed and that, during this time, there is no excitation of the vocal tract Thus, during the closed phase, the speech signal consists of the decaying vocal tract resonances alone Linear Prediction (LP) analysis performed over this time will therefore only identify the vocal tract filter

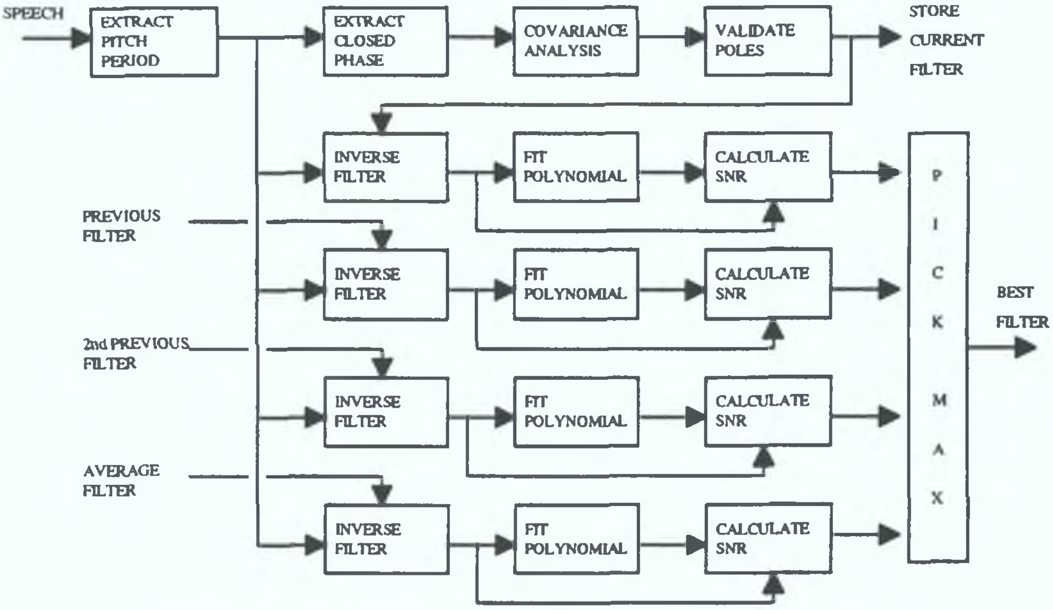


Fig. 6.1. Schematic diagram of the Closed Phase Inverse Filtering algorithm.

and will exclude any components due to the glottal excitation. The glottal waveform can then be determined by applying the inverse of the closed phase filter to the entire pitch period. As has been explained in Section 3.3.1, the method does not work in all cases, for example - there may be some residual excitation during the so-called "closed" phase, the closed phase may be too short to perform accurate Linear Prediction analysis, or the presence of noise and reverberation may cause inaccurate estimation of the vocal tract filter. The effects of these limitations are investigated in the performance analysis contained in the next section.

A schematic diagram of the CPIF algorithm used in this investigation in shown in Fig. 6.1. Input speech is recorded using phase linear equipment at a sampling frequency of 8 kHz. Glottal Closure Instant (GCI) identification precedes Closed Phase Inverse Filtering. Inaccurate GCI identification might introduce artefacts into the inverse filtering process which would put CPIF at a disadvantage compared with IAIF. Therefore, in these experiments, GCI identification is performed manually to ensure that the results of these experiments depend only on the accuracy of the inverse filtering algorithms. The effect of automatic GCI identification on the glottal extraction procedure is examined in the next chapter.

Once the GCIs have been identified, the inverse filtering algorithm proceeds by assuming that a short closed phase immediately follows the GCI. Based on the results of preliminary experiments, the closed phase is assumed to be 30 samples long and to start 2 samples after the closure instant. If the next pitch period is less than 60 samples long then the closed phase is taken as ending half way between this GCI and the next. The vocal tract filter is estimated by covariance analysis [Wong et al., 1979] performed over the assumed closed phase. Covarience analysis was chosen in preference to autocorrelation analysis since it is more accurate over short time windows. The resulting vocal tract filter is validated by removing poles that cannot be attributed to the vocal tract resonances [Childers and Lee, 1991]. Thus, poles at frequencies less than 250 Hz and with bandwidths greater than 500 Hz were removed from the filter.

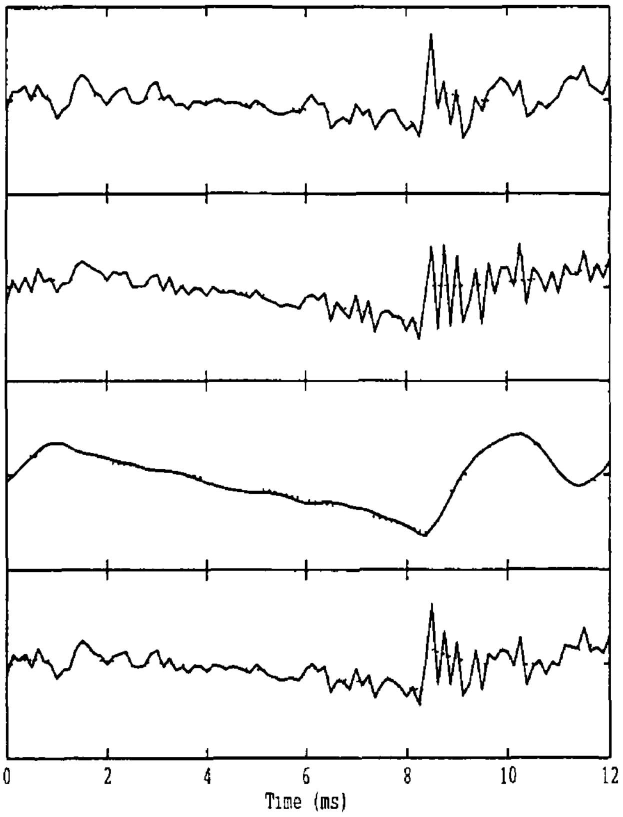


Fig 6 2 Polynomial fit to inverse filtered speech solid - inverse filtered, dotted - polynomial fit from top to bottom - current estimate, previous estimate, second previous estimate, averaged estimate

In preliminary experiments the estimated vocal tract filter often proved to be inaccurate This was generally due to noise or excitation during what was presumed to be the "closed" phase The basic inverse filtering algorithm was expanded by adding a new procedure which uses vocal tract filter estimates from previous pitch periods

In manual inverse filtering experiments, the filter is adjusted for maximum formant cancellation in the presumed closed phase [Fant, 1993] However, the true glottal flow always contains components related to uncompensated vocal tract modes that depend on the time-varying nature of the speech production system within a single glottal cycle The setting for maximum formant cancellation in the closed phase is generally used since it provides a pragmatic basis for synthesis In this investigation, the inverse filter is adjusted to achieve maximum formant cancellation throughout the entire glottal cycle That is, the estimated glottal waveform should be a smooth function of time except at the GCI Based on this principle, the most accurate vocal tract filter is viewed as the one which produces the smoothest glottal waveform estimate

The new multiple filter procedure involves inverse filtering each pitch period with four candidate vocal tract filters These candidates are the filter estimated over the current period, the filter estimated over the previous period, the filter estimated over the second previous period and a filter constructed by averaging the pole locations of these three filters The current pitch period is inverse filtered using these four candidate filters to obtain four glottal waveform estimates Two second order polynomials are fitted to each of the estimated glottal waveforms One polynomial is fitted between the previous and the current GCIs The second polynomial is fitted between the current GCI and the end of the closed phase The accuracy of each candidate inverse filter is calculated as the Signal to Noise Ratio (SNR) between the estimated glottal waveform and the fitted polynomials The candidate vocal tract filter giving the maximum SNR is taken as the smoothest glottal estimate and used as the inverse filter for the current pitch period An example of the inverse filtered and fitted waveforms can be seen in Fig 6 2 Clearly the filter calculated over the current closed phase is inaccurate In this case, the second previous filter produced the maximum SNR and was chosen as the best inverse filter for use in the current period

In early experiments, a fifth vocal tract filter estimate was tested This was a filter estimated over a concatenation of the current and two previous closed phases The technique was originally proposed by Chan and Brookes [Chan and Brookes, 1989] Unfortunately, the estimates proved to be inaccurate due to discontinuities at the concatenation points This technique was therefore abandoned

6.2.2 Iterative Adaptive Inverse Filtering

Proposed and developed by Alku, Iterative Adaptive Inverse Filtering (IAIF) operates on the principle that the overall spectral tilt of the speech signal can be attributed to the glottal waveform In the case of vowels, speech production can be viewed to consist of a glottal excitation, filtered by the vocal tract and radiated at the lips The spectra of these three components can be seen in Fig 6 3 Alku's contention is that the tilt of the speech spectrum is due to the combined glottal excitation and lip radiation spectra The vocal tract transfer function is itself wide-sense flat with some high energy regions corresponding to the formants The IAIF algorithm operates by repeatedly removing the glottal and radiation effects using low order Linear Prediction analysis and inverse filtering This removes the overall spectral tilt of the speech and allows accurate estimation of the vocal tract filter using high order Linear Prediction analysis The estimated vocal tract filter is used to inverse filter the original speech signal to obtain the differentiated glottal flow

A schematic diagram of the IAIF technique is shown in Fig 6 4 In the first iteration, the effect of the glottal excitation and the lip radiation is modelled by computing a Linear Prediction analysis of order 1 Referring to Fig 6 3, this means that the combined glottal (a) and lip spectra (c) are estimated by calculating a crude envelope to the speech spectrum (d) The combined glottal flow and lip radiation functions correspond to the differentiated glottal flow or glottal waveform Thus, the spectral effect of the glottal waveform can be cancelled by inverse filtering the speech signal using the first order Linear Prediction estimate A preliminary model for the vocal tract filter is obtained by performing 10th order Linear Prediction analysis on the inverse filtered signal The first estimate for the glottal waveform is produced by inverse filtering the original speech signal using this vocal tract filter estimate

In the second iteration, the spectral effect of the glottal waveform is re-estimated by 4th order Linear Prediction analysis performed over the glottal waveform estimated in the first iteration As before, the spectral contribution of the glottal waveform is cancelled from the speech signal by inverse filtering The vocal tract filter is re-estimated using 10th order Linear Prediction over the inverse filtered speech The final glottal waveform estimate is obtained by inverse filtering the original speech with this

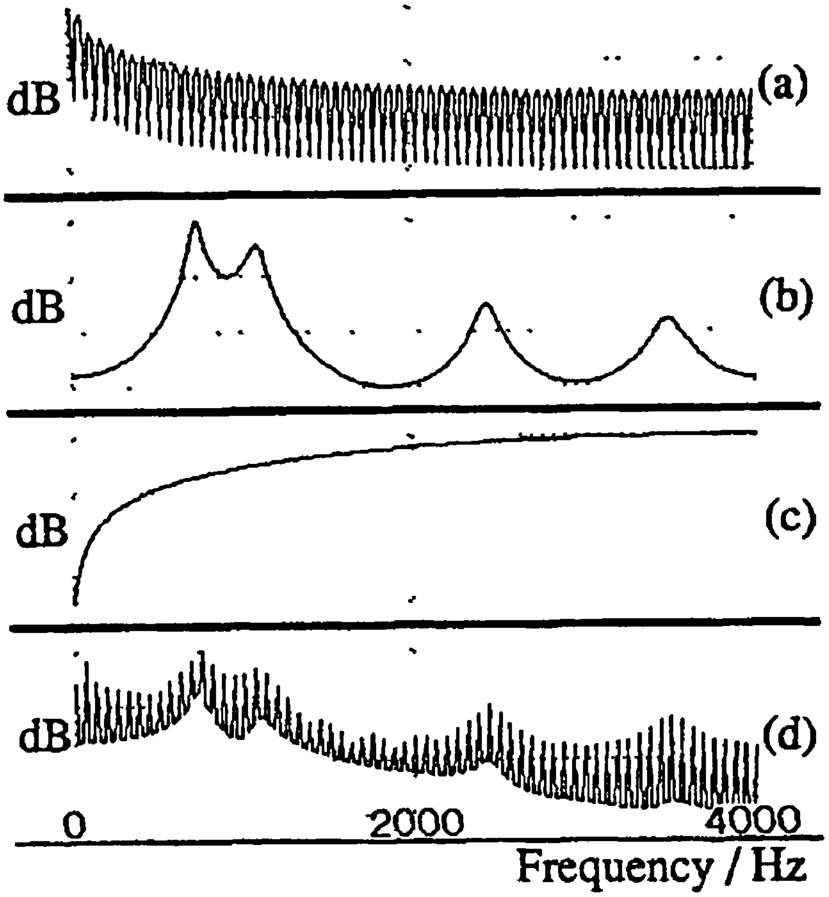


Fig 6 3 Components of speech production (a) spectrum of the glottal excitation, (b) transfer function of the vocal tract, (c) transfer function corresponding to the lip radiation effect, (d) speech spectrum [after Alku, 1992b]

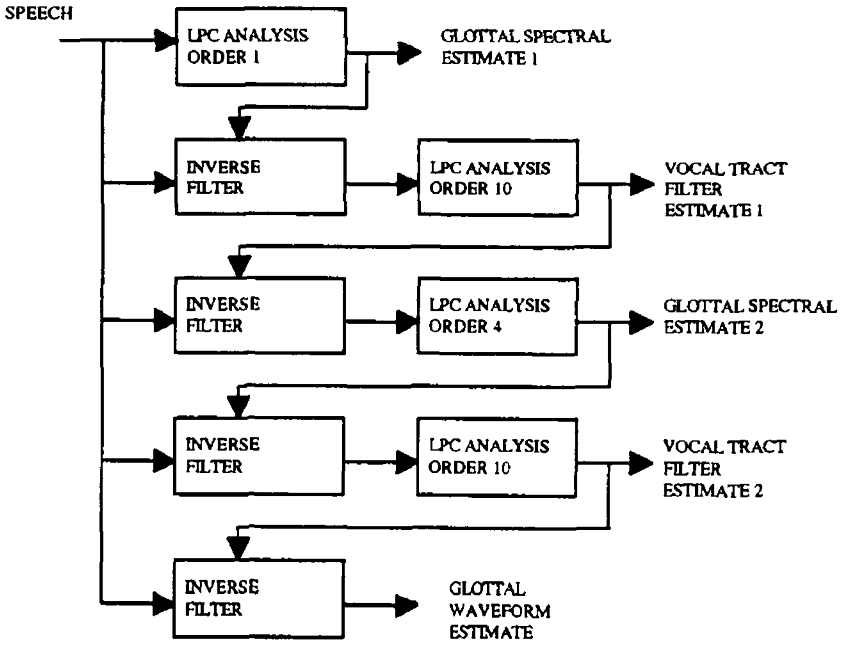


Fig 6 4 Schematic diagram of the Iterative Adaptive Inverse Filtering algorithm.

vocal tract filter estimate In all cases, the LP analysis is carried out by the autocorrelation method with Hamming windowing of the data.

The method used in this investigation is slightly altered from that proposed by Alku Firstly, for the proposes of fitting the LF model the differentiated glottal flow is estimated, whereas Alku's method estimates the glottal flow This was remedied by removing the integration steps for converting the differentiated flow to the flow Secondly, Alku's method used asynchronous IAIF to determine the pitch period before applying pitch synchronous IAIF to estimate the glottal waveform In this investigation, the pitch contour was provided to the IAIF algorithm manually In this way, the accuracy of the glottal waveform estimates produced by IAIF can be directly compared with those produced by CPIF

6.2.3 Glottal Model Fitting

Developed by Fant, Liljencrants and Lin, the LF model represents the differentiated glottal flow using a time-domain waveform model A typical LF waveform is depicted Fig 6 5 The waveform parameters of the model are - t_0 instant of glottal opening, t_p instant of maximum flow, t_e instant of glottal closure, t_a effective return phase duration, t_c end of the glottal cycle and E_e rate of flow change occurring at glottal closure For convenience, t_c is set equal to t_0 of the next glottal cycle This implies that the model lacks a closed phase In practise this is not a drawback since for small values of t_a the tail of the exponential curve will fit closely to the zero line providing, for all intents and purposes, a closed phase The fundamental period T_0 is therefore equal to t_c-t_0

The model consists of two parts The first part represents the glottal open phase by an exponentially growing sinusoid

$$g(t) = E_o e^{\alpha t} \sin\omega_g t \qquad t_o < t \le t_e \qquad (6\ 1)$$

where ω_g is the pitch period in rad/s, E_0 is a scale factor and α controls the growth of the sinusoid

The second part of the model is an exponential segment representing the residual flow after the instant of glottal closure This return phase is represented by

$$g(t) = -\frac{E_e}{\varepsilon t_a}\left(e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)}\right) \qquad t_e \le t \le t_c \qquad (6\ 2)$$

where ε controls the slope of the return phase As shown in Fig 6 5, the parameter t_a is the projection of the gradient of $g(t)$ at the glottal closure instant onto the time axis For small t_a,

$$\varepsilon \approx 1/t_a \qquad (6\ 3)$$

otherwise, ε can be iteratively determined from

$$\varepsilon t_a = 1 - e^{-\varepsilon(t_c-t_e)} \qquad (6\ 4)$$

Finally, the entire waveform is generated according to the constraint that there is zero net gain of flow during a fundamental period

$$\int_0^{T_o} g(t) = 0 \qquad (6\ 5)$$

Using these relations, the waveform parameters can be converted to the corresponding synthesis parameters and the waveform constructed Another alternative set of parameters, the analysis parameters, is frequently used in the study of glottal waveform dynamics These are defined as

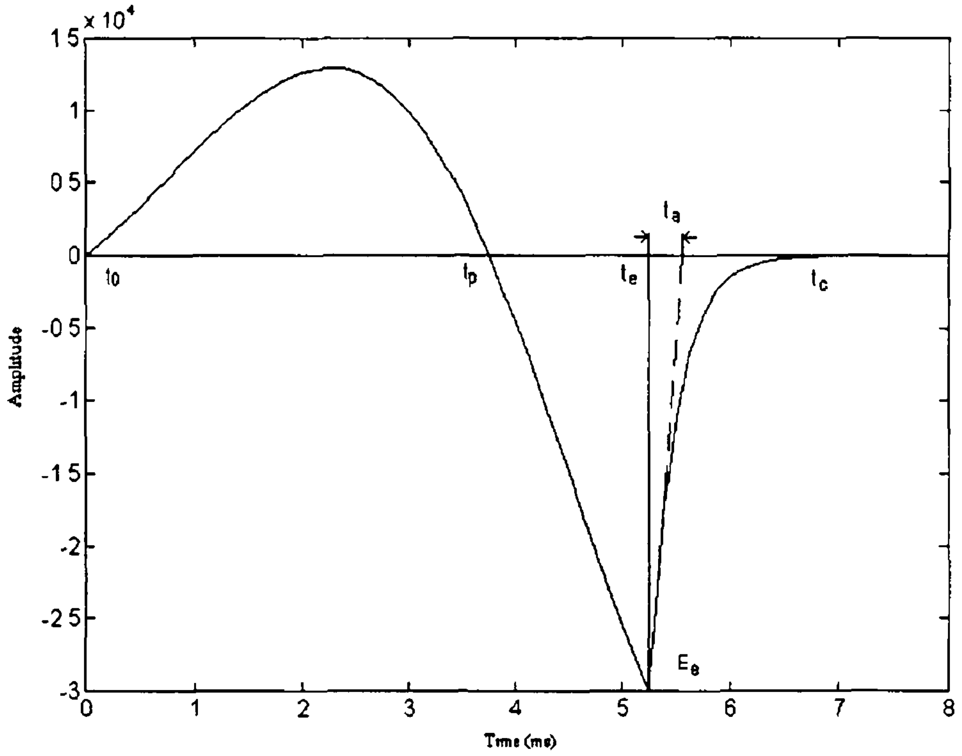


Fig 6 5 LF model representing the differentiated glottal flow

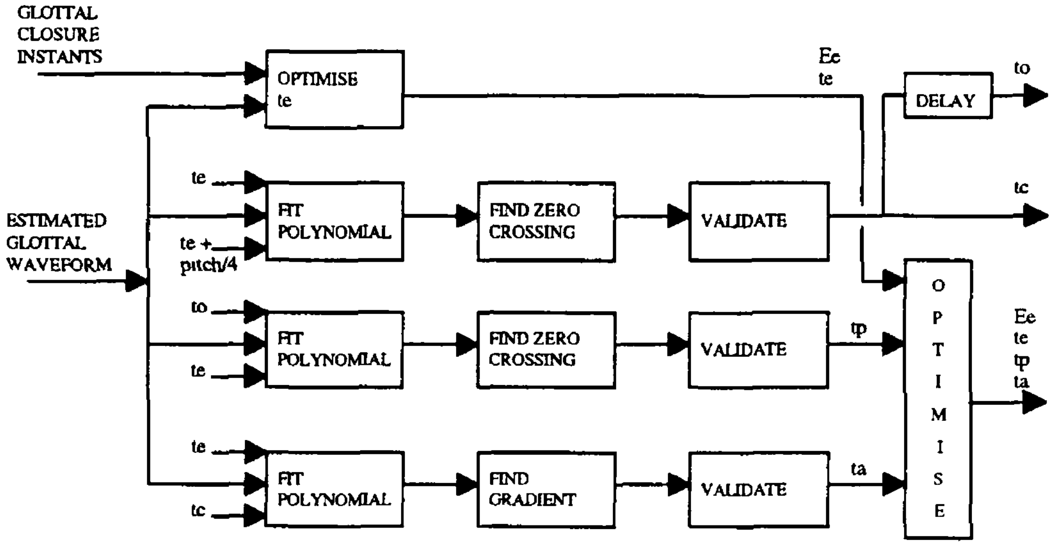


Fig 6 6 Schematic diagram of the LF model fitting algorithm.

$$r_g = \frac{\omega_g}{2\pi F_0} = \frac{T_0}{2t_p}, \quad r_k = \frac{t_e - t_p}{t_p}, \quad r_a = \frac{t_a}{T_0}, \quad o_q = \frac{t_e}{T_0} \qquad (6\ 6)$$

where r_g is the glottal frequency, r_k is the skewness factor, r_a is the dynamic leakage and o_q is the open quotient The analysis parameters are related to the spectrum of the LF waveform [Fant and Lin, 1988] The glottal gain E_e controls the overall level or intensity of the spectrum The dynamic leakage r_a controls the roll-off of the glottal spectrum Due to the exponential waveshape of the return phase its spectrum approximates that of a first order lowpass filter with a cut-off frequency of $F_a = 1/(2\pi\ \ t_a) = F_0/(2\pi\ \ r_a)$ This means that the larger r_a, the lower the cut-off frequency and the greater the high frequency energy reduction Together the parameters r_g and r_k determine the skewing of the glottal waveform and so control the level of the lower harmonics and the depth of the zeros in the source spectrum

A schematic diagram of the LF fitting procedure used in this investigation is shown in Fig 6 6 The LF model is fitted to a single fundamental period of the estimated glottal waveform based on minimisation of the mean square error between the two signals The fitting procedure is carried out in three stages Firstly, the period start and end points are identified Secondly, starting values for the LF waveform parameters are found by using a polynomial approximation to the LF model Thirdly, the mean square error between the estimated glottal waveform and the LF waveform is minimised using a multi-dimensional optimisation routine applied to the LF parameters

As for the inverse filtering algorithms, the fitting algorithm requires that reasonably accurate GCI identification has been performed beforehand The parameter t_0, marking the start of the cycle, is set equal to the end of the previous period In the case of onsets, t_0 is taken as one pitch period before the first closure The starting value for the glottal closure instant t_e is optimised by searching for the minimum of the estimated glottal waveform within +/-8 samples of the closure instant identified manually The amplitude of the minimum is taken as the starting value for the gain parameter E_e The parameter t_c marking the end of the glottal cycle is found by fitting a second order polynomial between the closure point t_e and the halt point, which is set one quarter of the way between the current closure and the next The polynomial smoothes out any noisy fluctuations in the estimated glottal waveform and so allows accurate determination of the point at which the glottal waveform has zero amplitude The zero crossing of the polynomial is taken as the end of the glottal cycle t_c The end point is limited so that it is at least one sample after t_e and before the halt point

The starting values of the remaining parameters are determined by using a second order polynomial approximation to the LF model Unlike the LF model, a polynomial waveform can be fitted to a curve in the least square sense without iteration Therefore, polynomial fits to the estimated glottal waveform are quickly computed and can be used to provide starting values for the LF waveform parameters An example of such a fit is shown in Fig 6 7 A second order polynomial is fitted to the open phase, between the glottal opening instant t_0 and the glottal closure instant t_e The zero crossing point of this curve gives the initial estimate for the instant of maximum flow t_p If the zero crossing does not occur between $(t_0+t_e)/2$ and t_e then it is arbitrarily set to $(t_0+t_e)/2+2$ In a similar manner, another second order polynomial is fitted to the return phase, between the GCI t_e and the end of the cycle t_c The gradient of the polynomial at the GCI is used to calculate the starting value for the parameter t_a The

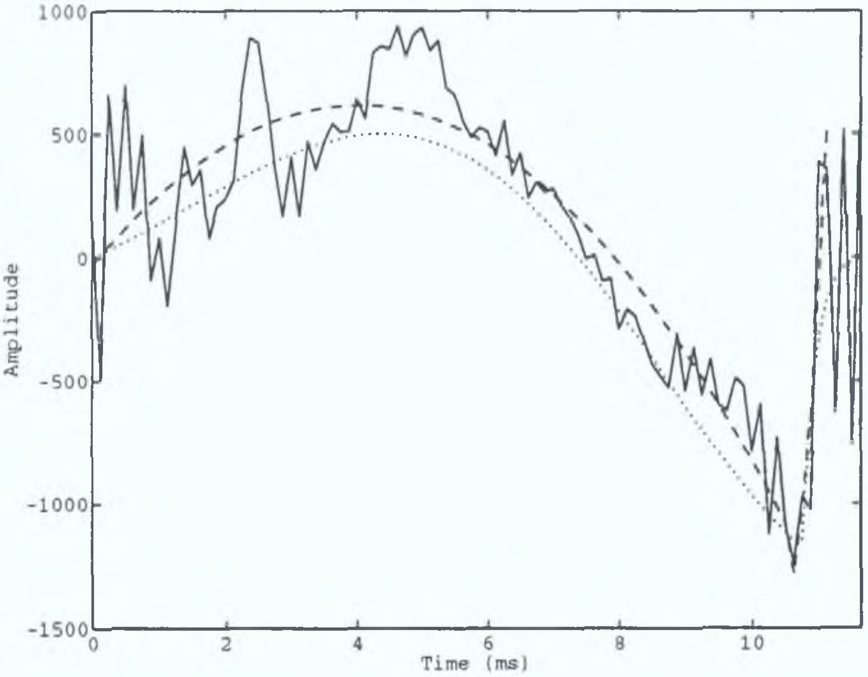


Fig. 6.7. Polynomial and LF model fit to extracted glottal waveform: solid line - extracted glottal waveform; dotted line - second order polynomial fit; dashed line - LF model.

value is limited to be greater than or equal to zero, and less than or equal to t_c-t_e. If t_a is equal to t_c-t_e then the return phase is made linear rather than exponential. This ensures that the LF waveform is reasonably smooth.

The parameters t_p, t_e, t_a and E_e are optimised using the Simplex multi-dimensional gradients based mean squared error minimisation procedure [Press et al., 1986]. At each iteration, the parameters are validated and the LF waveform generated. The mean squared error between the LF model and the estimated glottal waveform is calculated. Based on the results of this calculation, the waveform parameters are adjusted so as to minimise the error. The parameters are optimised to the nearest sampling instant and the resulting LF waveform is stored.

Obviously, the optimisation procedure is computationally complex, requiring roughly 1 hour to process 1 second of speech using Matlab on a 33 MHz 486. This makes the optimisation approach unsuitable for direct implementation in conventional speech coding applications. However in this investigation, the goal was to determine the speech quality which could be achieved by glottal based coding. Therefore, the normally severe runtime constraints imposed on coding systems have been relaxed. Additionally, the computational complexity of the approach could be substantially reduced by the use of a more intelligent optimisation algorithm and a faster formulation of the LF model [Qi and Bi, 1994].

The fitting algorithm pays special attention to the accuracy of the parameters t_p, t_e, t_a and E_e while the parameters t_o and t_c are set quite arbitrarily. This is believed to have little impact on the overall quality of the LF fit Examining the LF waveform, Fig 6 5, it can be seen that the waveform changes very rapidly around t_p, t_e and t_a In contrast, the waveform is quite smooth in the region of t_0 and t_c Thus, the precise values of t_0 and t_c are not only difficult to determine, but also make little difference to the LF waveshape As well as this, the ear is mainly sensitive to the pitch, controlled by t_e, and the spectral shape of the glottal waveform, controlled by t_p and t_a [Gobl and Ni Chasaide, 1992] Hence, accurate determination of these parameters is necessary for high quality speech synthesis Exact identification of the cycle start and end points, t_0 and t_c , is not as crucial for re-synthesis

6 3 PERFORMANCE STUDY

This section describes the experiments carried out to determine the performance of the glottal waveform extraction algorithms Four main investigations were undertaken Firstly, the effect of the LF fit optimisation routine was assessed Secondly, the impact of the new multiple filter procedure on the accuracy of CPIF was examined Thirdly, the CPIF, IAIF and LF fitting algorithms were investigated by analysing the results obtained for various types of voiced speech Fourthly, the robustness of the inverse filtering and fitting algorithms was assessed by performing glottal waveform extraction on continuous speech under noiseless, noisy and reverberant conditions

In order that the accuracy of the extraction algorithms be quantified, it was necessary that suitable error measures be defined However, the true glottal flow is difficult to determine during natural speech Therefore, the accuracy of the extracted glottal waveform cannot be determined directly Assuming that the LF model can represent all possible glottal waveform shapes, the accuracy of the inverse filter can be assessed by determining the nearness of the LF fit to the estimated glottal waveform The nearness of the fit was quantified as the LF SNR This is defined as the SNR of the LF fit relative to the estimated glottal waveform averaged over all pitch periods of the signal

For speech coding applications, the accuracy of the glottal estimate is really not that important What is important, is that speech re-synthesised from the extracted LF waveform is similar to the original input speech In order to investigate this, speech was re-synthesised from the extracted LF waveforms and compared to the input signal

To ensure high quality synthesis, the vocal tract filter and glottal gain were re-optimised for each pitch period of the speech signal The input and output signal were segmented using a Hamming window applied from t_0-1 25 ms to t_c+1 25 ms and an 8th order Linear Prediction vocal tract filter was determined by ARX estimation [Astrom and Eykhoff, 1971] between the excitation, the LF waveform, and the desired output, the original speech The glottal gain E_e was re-optimised by minimising the difference in the energy between the re-synthesised signal and the original signal Using the new vocal tract filter and glottal gain, the speech signal was re-synthesised and compared to the original Note that the re-synthesis process is explained in greater detail in Chapter 7

The quality of the re-synthesised speech was quantified as the re-synthesised SNR This is defined as the SNR of the re-synthesis speech relative to the original, averaged over all pitch periods of the signal

Three types of speech data were used in the experiments - noiseless, noisy and reverberant The noiseless data consists of the sentence "We were away a year ago", spoken by a male and female subject, and recorded under noiseless anechoic conditions The noisy data was produced by adding various intensities of white noise to the noiseless recordings Similarly, the reverberant data was produced by convolving the noiseless recordings with simulated room impulse responses For more information on the generation of this data, the reader is referred to Appendix C Of particular relevance to the following discussion is the fact that the male speech has been informally assessed as being slightly breathy

The performance study is described in the next four sub-sections The first sub-section describes the investigation of the LF fit optimsation procedure The second sub-section describes the experiments undertaken to determine the impact of the multiple filter procedure The third sub-section assesses the accuracy of the inverse filtering and LF fitting algorithms across various phonetic categories of male and female speech Lastly, sub-section four investigates the performance of the algorithms in processing continuous natural speech under noiseless, noisy and reverberant conditions

6.3 1 LF Optimisation

The effects of the LF optimisation procedure were investigated by performing LF fitting with and without optimisation The speech material consisted of the noiseless male and female recordings Both CPIF and IAIF were used to perform the inverse filtering operation

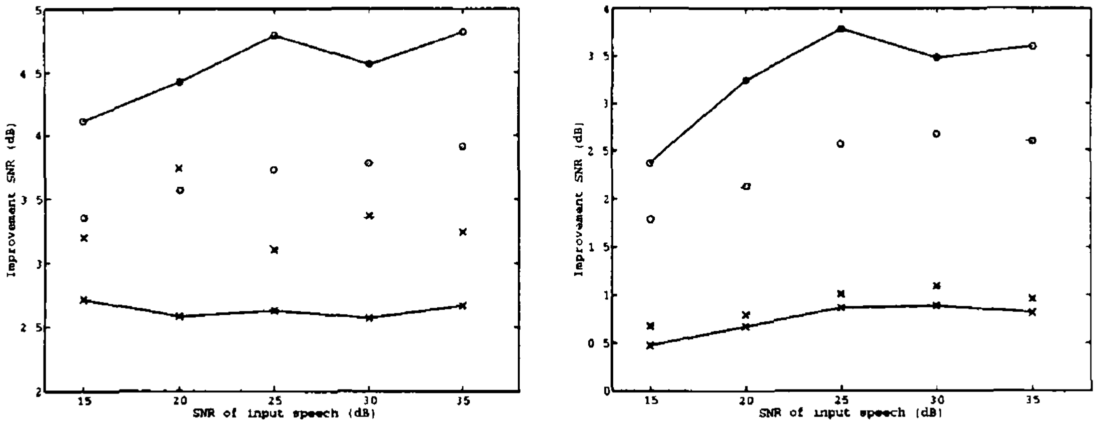
Table 6 1 shows the improvements in the LF SNR obtained by using optimisation Optimisation produces an average improvement of 3 79 dB in the LF SNR This, in turn, causes a 1 dB average improvement in the re-synthesis of the male speech and an average improvement of 3 dB for the female speech These improvements in accuracy are significant However, the results also indicate that the polynomial LF approximation is reasonably accurate and could be used in applications where the computational burden precludes optimisation

	SUBJECT	CPIF	IAIF
LF	Male	2 78	3 32
LF	Female	4 93	4 11
Re-synthesised	Male	0 99	1 04
Re-synthesised	Female	3 40	2 68

Table 6 1 *Improvement in LF and re-synthesised SNR due to LF fit optimisation, noiseless recordings*

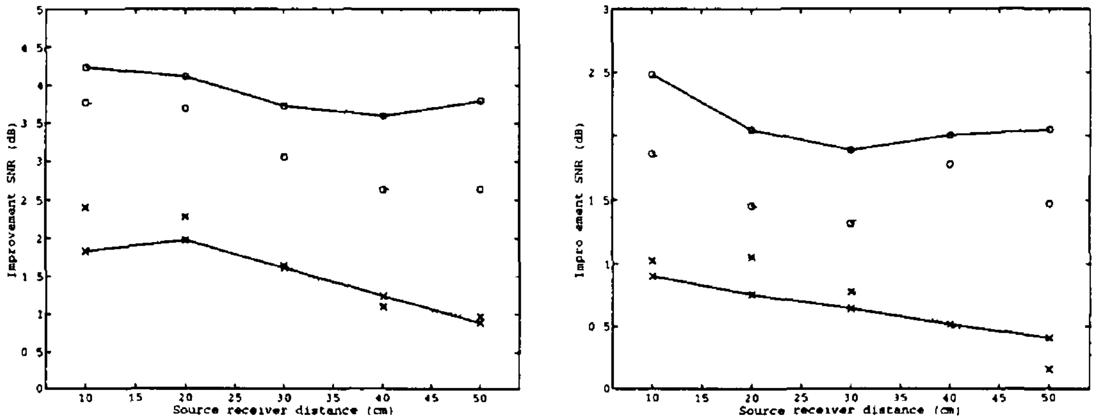
The robustness of the fit optimisation procedure was tested by performing LF extraction, with and without optimisation, on speech corrupted by the addition of noise and reverberation Figs 6 8 and 6 9 show the improvement in LF and re-synthesised SNR due to fit optimisation when processing noisy and reverberant speech

Clearly the improvement in SNR due to optimisation becomes less with increasing distortion in the speech signal Increased distortion in the speech signal leads to reduced smoothness in the glottal waveform estimates This complicates the error function of the LF fit with respect to the glottal parameters As a result, under conditions of increased distortion, the optimisation routine finds the



(a) (b)

Fig 6 8 Improvement in SNR due to LF fit optimisation, noisy speech (a) LF SNR, (b) re-synthesised SNR solid - CPIF, dotted - IAIF, x - male subject, o - female subject



(a) (b)

Fig 6 9 Improvement in SNR due to LF fit optimisation, reverberant speech (a) LF SNR (b) re-synthesised SNR solid - CPIF dotted - IAIF x - male subject, o - female subject global minimum of the error function less frequently. This has the effect of reducing the improvements in SNR brought about by the use of fit optimisation.

For the female subject, the results show a greater improvement due to optimisation than for the male. This may be explained by the breathiness of the male voice. The breathy noise means that the male glottal waveform estimates lack smoothness, resulting in the optimisation procedure settling for local, rather than global, minima of the error function.

As might be expected, the LF SNR shows a greater improvement due to optimisation, than the re-synthesised SNR. The LP coefficients determined by ARX estimation provide some compensation for inaccuracies in the extracted glottal waveform. Thus, an improvement in the LF fit has less effect on the quality of the re-synthesised speech. Nevertheless, the improvement in the re-synthesised speech is significant and supports the contention that improved LF fitting to the inverse filtered speech is equivalent to more accurate fitting of the LF model to the true glottal excitation.

Overall, the improvement in accuracy due to optimisation is significant and supports the use of optimisation when the computational requirement is not prohibitive. Additionally, the results suggest that the polynomial LF approximation is reasonably accurate and could be used independently of optimisation in some glottal extraction applications.

6.3.2 Multiple Closed Phase Filter Estimates

To investigate the effect of the new multiple filter procedure, the LF SNR and re-synthesised SNR were determined for glottal waveforms extracted by CPIF using only the current filter estimate and using the multiple filter procedure.

The improvements in SNR obtained due to the use of multiple filter procedure are presented in Table 6.2. In the case of the male speech, the new procedure leads to an improvement of approximately 1 dB in both the LF and re-synthesised SNRs. Little improvement is seen in either SNR for the female subject. This difference is probably due to the breathiness of the male voice. The presence of this noise makes each closed phase filter estimate less reliable. Thus the robustness provided by using multiple filters gives an improvement in SNR for the male subject, even under noiseless conditions. In contrast, the extra robustness is not needed for processing the modal female voice.

Table 6.3 shows the percentage of times that each of the inverse filters was chosen by the automatic procedure. As might be expected, the current estimate is chosen most often, followed by the previous, the second previous and the average filter. Comparing the male and female results shows that the current estimate is chosen more often in the case of the female speech. This fact concurs with the small change in SNR due to the use of multiple filters experienced for the female subject.

	SUBJECT	CPIF
LF	Male	1.01
LF	Female	0.39
Re-synthesised	Male	1.01
Re-synthesised	Female	-0.03

Table 6.2. *Improvement in LF and re-synthesised SNR due to the multiple closed phase filter procedure, noiseless recordings.*

	Male subject	Female subject
Current estimate	33.6	35.0
Previous estimate	30.9	28.4
Second previous estimate	24.8	24.6
Averaged	10.7	12.0

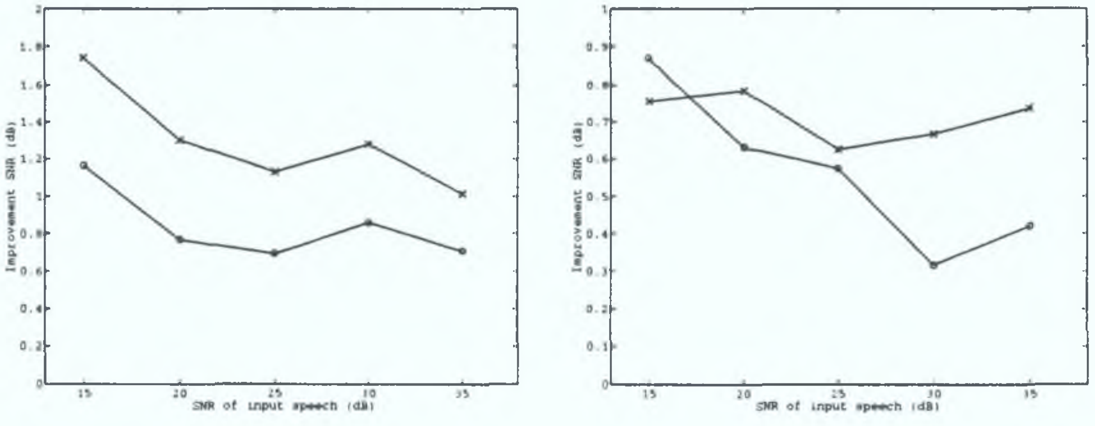
Table 6.3. *Frequency of filter selection during the multiple closed phase filter procedure, noiseless recordings.*

The robustness of the multiple filter procedure was tested by performing LF extraction on speech corrupted by the addition of noise and reverberation. Figs. 6.10 and 6.12 show the improvement in the LF and re-synthesised SNR due to use of the multiple filter procedure for noisy and reverberant speech, respectively. Figs. 6.11 and 6.13 show the percentage of times each filter was chosen by the multiple filter procedure when processing the noisy and reverberant speech, respectively.

In almost all cases, the SNR improvement increases with increasing distortion. As the amount to distortion increases, the inverse filter estimated over the current closed phase becomes less reliable. Thus, it becomes more likely that one of the previous filters, or the average filter, is more accurate than the current estimate. This fact is corroborated by three of the graphs showing the percentage times each filter is chosen. At low distortion levels, the current filter estimate is chosen most often by the automatic procedure and the averaged filter estimate is chosen least often. In contrast, at the highest distortion levels studied, the average filter estimate is chosen most frequently and the current estimate second.

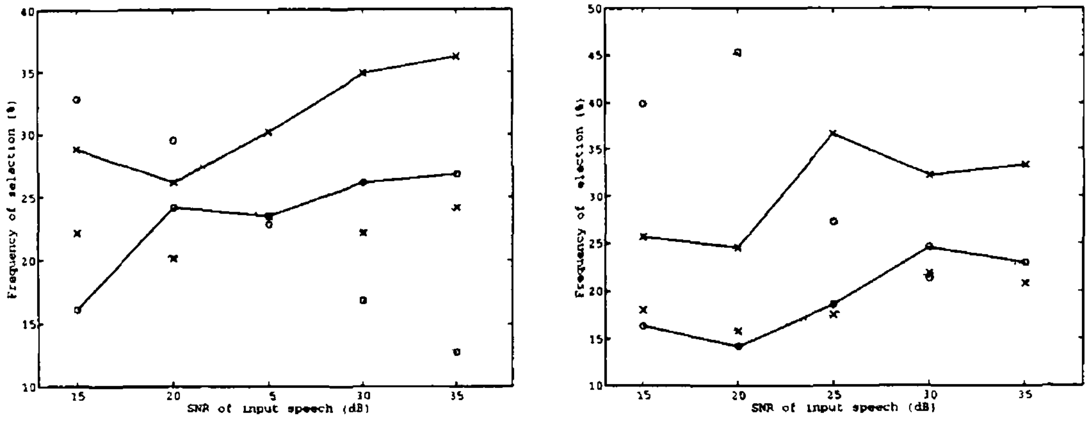
Again, the SNR improvement is greater for the male speech than for the female speech. As before, this is probably linked to the intrinsic distortion associated with breathiness of the male voice.

Overall, the multiple filter procedure provides a useful improvement to the accuracy of the CPIF algorithm. The procedure is particularly effective in making the closed phase method more robust to distortions in the speech signal. The procedure carries little computational overhead and is recommended for use in all CPIF applications.



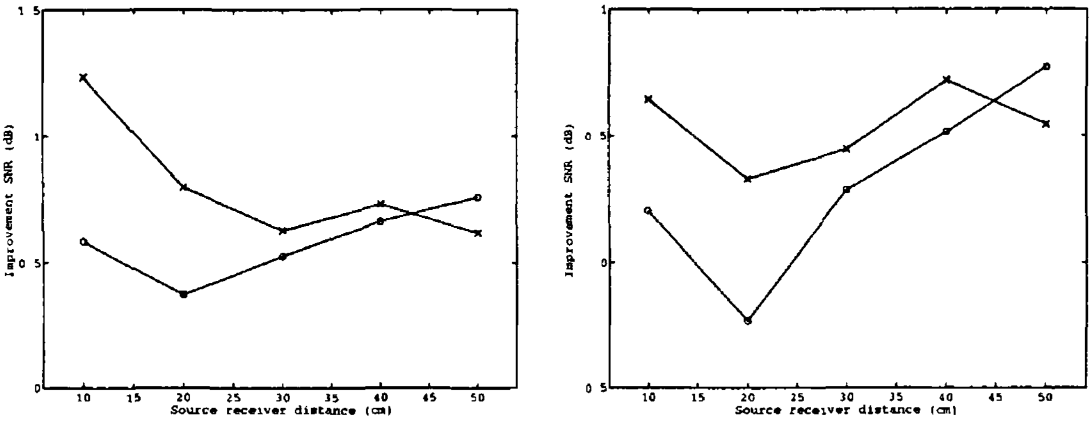
(a) *(b)*

Fig. 6.10. *Improvement in SNR due to multiple filter procedure, noisy speech: (a) LF SNR; (b) re-synthesised SNR: x - male subject; o - female subject.*



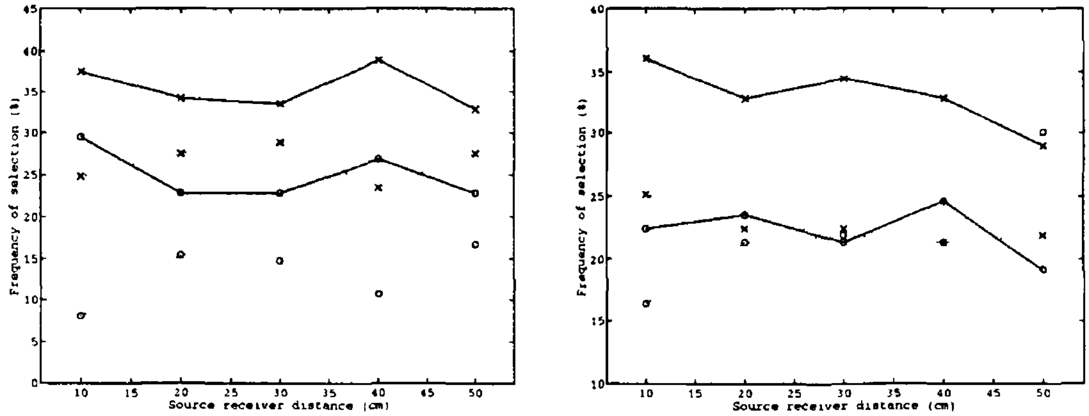
(a) (b)

Fig 6 11 Frequency of filter selection, noisy speech (a) male subject, (b) female subject solid x - current, solid o - previous, dotted x - second previous, dotted o - averaged



(a) (b)

Fig 6 12 Improvement in SNR due to multiple filter procedure reverberant speech (a) LF SNR, (b) re-synthesised SNR x - male subject, o - female subject



(a) (b)

Fig 6 13 Frequency of filter selection, reverberant speech (a) male subject, (b) female subject solid x - current, solid o - previous, dotted x - second previous, dotted o - averaged

6.3 3 Phonetic Categories

The performance of the inverse filtering and LF fitting algorithms was tested by processing segments of speech recorded under noiseless conditions

Figs 6 14 (a) and (b) show the male vowel [ə], obtained from the end of "ago", inverse filtered using the CPIF and IAIF algorithms, respectively The glottal waveform estimates provided by the inverse filtering algorithms are very similar Both are reasonable, with the IAIF algorithm providing the best formant cancellation and a slightly smoother waveform estimate The LF fitting procedure performs well in both cases, accurately capturing the dynamics of the estimated glottal waveforms

The results obtained when processing the male vowel [ɪ] are shown in Fig 6 15 Again, both inverse filtering algorithms provide good glottal waveform estimates The CPIF estimate is noisier than that produced by IAIF However, the IAIF estimate shows incomplete formant identification at the start of the open phase As before, the LF model gives a close fit to the estimated glottal waveforms

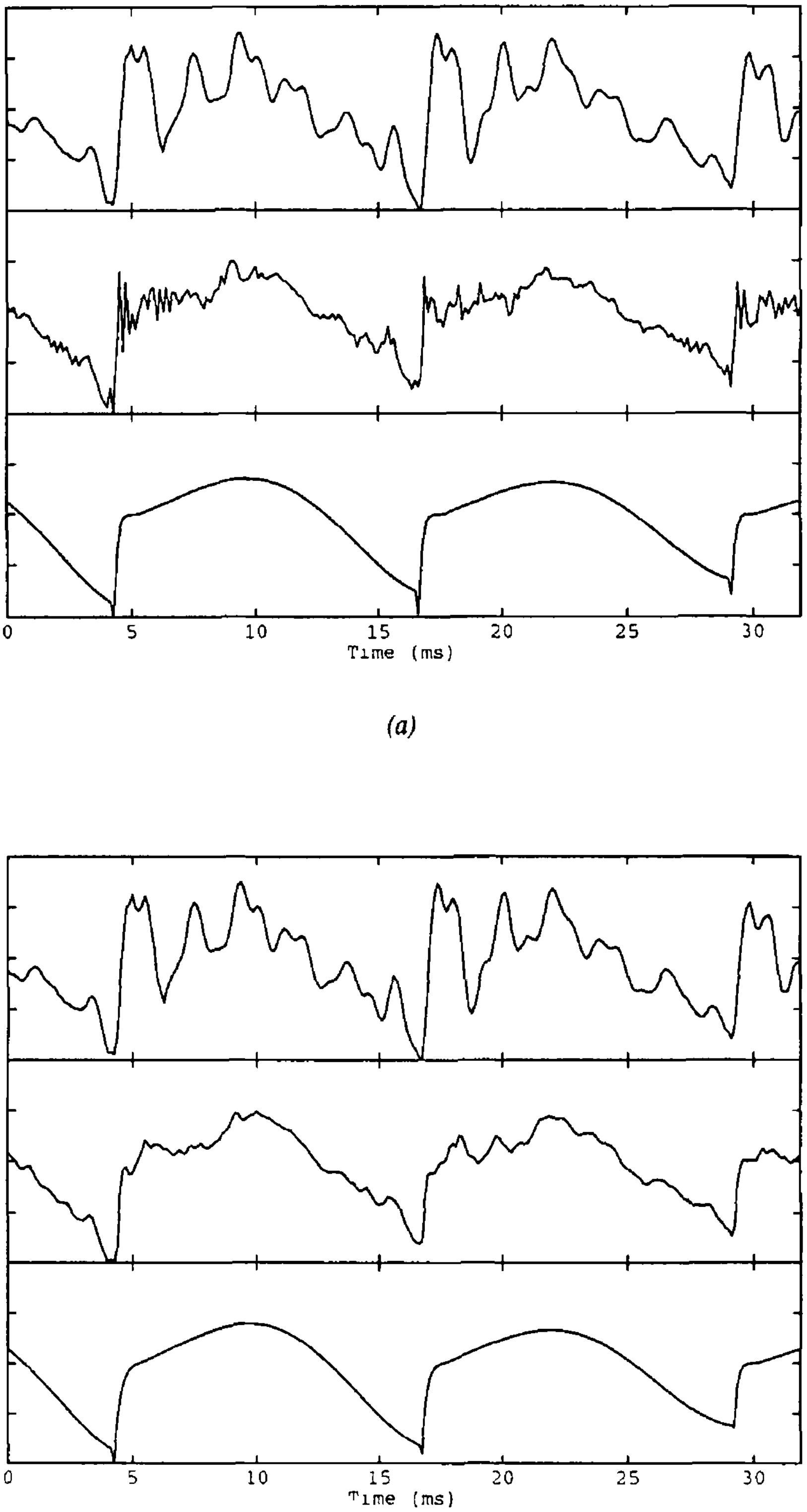
The results of inverse filtering of the same vowel [ɪ] phonated by a female subject, are presented in Fig 6 16 In this case, poor results would be expected from the CPIF algorithm due to the short duration of the closed phase However, CPIF and IAIF perform equally well Again, IAIF provides a slightly smoother waveform estimate, particularly at the start of the closed phase Regardless of this, the fitted LF waveforms are very similar

The results obtained from the female vowel [æ] are shown in Fig 6 17 In this test, the results obtained by CPIF and IAIF are extremely similar The only difference is during the open phase, when the CPIF estimate is almost flat and the IAIF estimate contains some residual formant ripple The fitted LF waveforms match almost exactly

Inverse filtering of the vowel [u] is difficult due to its low first formant Fig 4 18 shows CPIF and IAIF applied to [u] as phonated by a male subject The difficulties of the inverse filtering operation can be clearly seen in the poor glottal waveform estimates In both cases, a large dip occurs in the middle of the open phase This leads to poor LF model fitting in the fourth cycle of the IAIF glottal waveform estimate Overall, the CPIF algorithm performs slightly better than IAIF The IAIF algorithm separates the source and vocal tract effects in the spectral domain In the case of a vowel with a low first formant, discrimination of the F0 and F1 peaks is difficult and this approach results in incomplete F1 cancellation and poor glottal estimation

Voiced fricatives present further problems for inverse filtering algorithms Accurate formant estimation is difficult to achieve due to excitation noise occurring during the glottal cycle Fig 6 19 shows the results obtained by applying the inverse filtering algorithms to the male voiced fricative [v] The waveform estimates produced by the two algorithms are surprisingly similar Excitation noise during the closed phase was expected to cause inaccurate formant identification in the case of the CPIF algorithm It may be that the closed phase in this segment is long enough to allow the noisy effects to be averaged out Higher pitched voiced fricatives might cause greater problems The fitted waveforms illustrate an important property of the LF model, that is, the ability to represent sinusoidal waveforms as well as those with sharp discontinuities at the closure instant

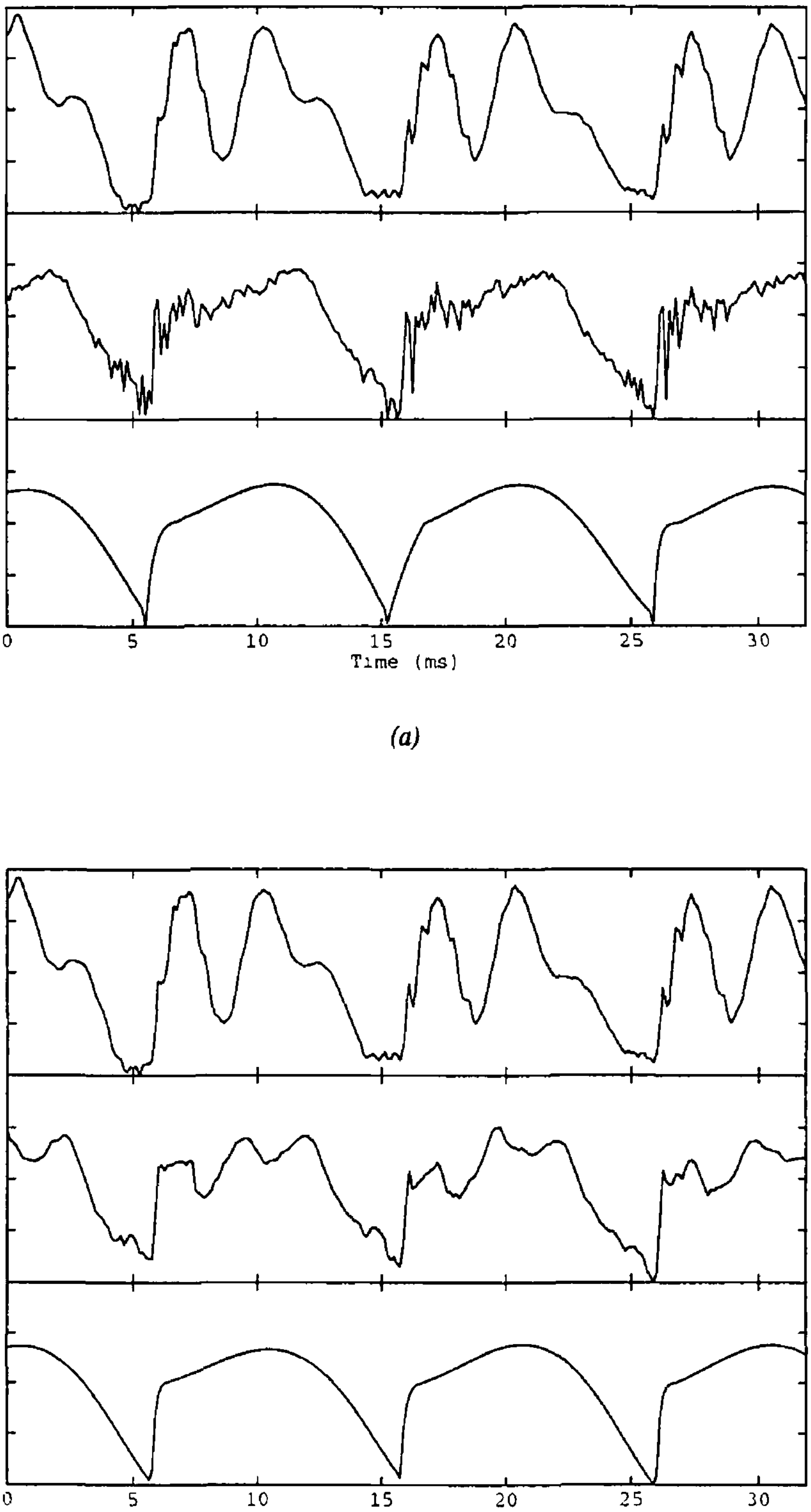
Voiced plosives are another category of sound for which inverse filtering is difficult Voiced plosives are produced by rapid removal of an airflow obstruction within the vocal tract At the start of a



(a)

(b)

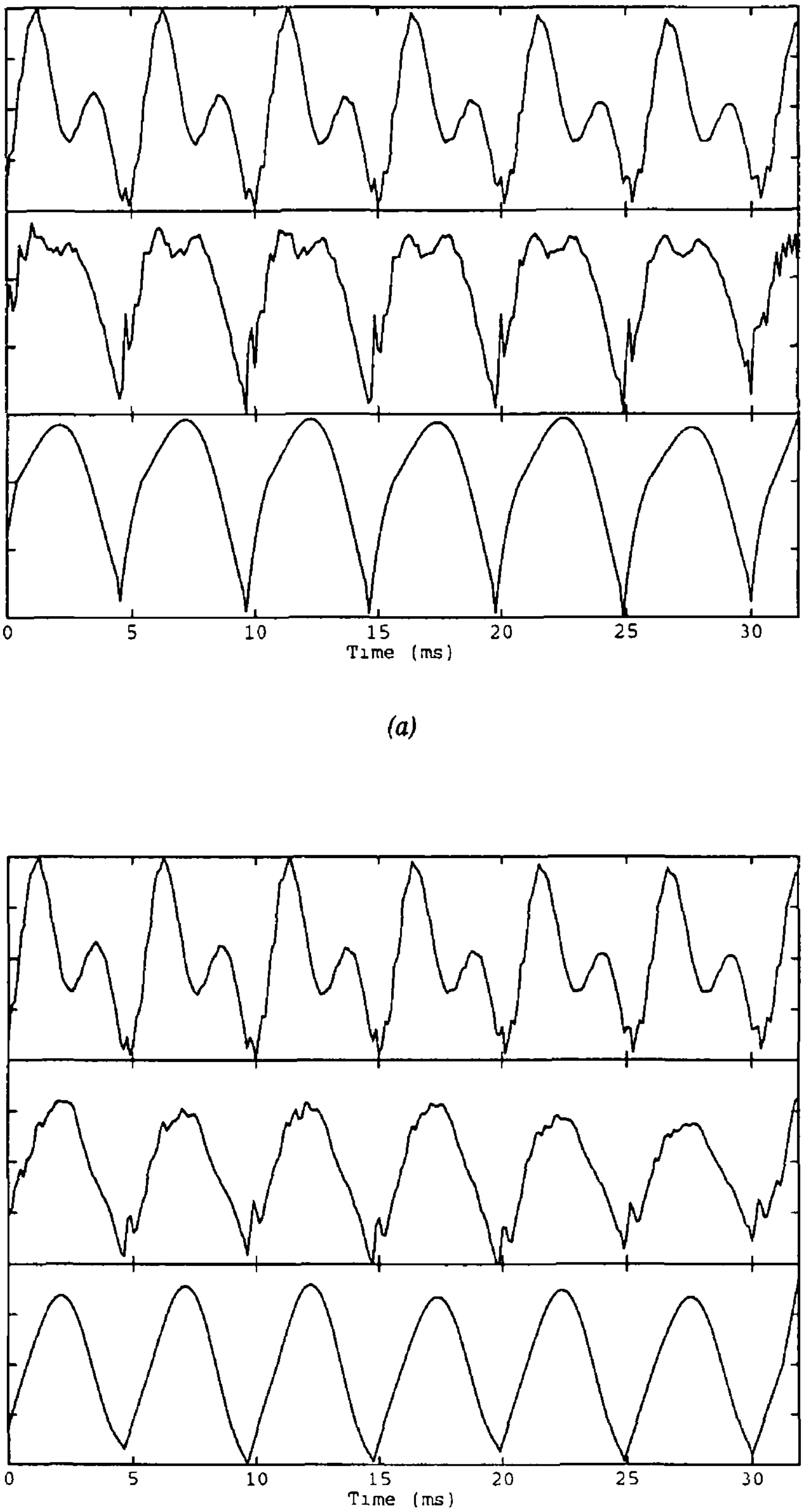
Fig 6 14 Glottal extraction applied to male [ə] (a) CPIF, (b) IAIF top - speech middle - estimated glottal waveform, bottom - fitted LF waveform.



(a)

(b)

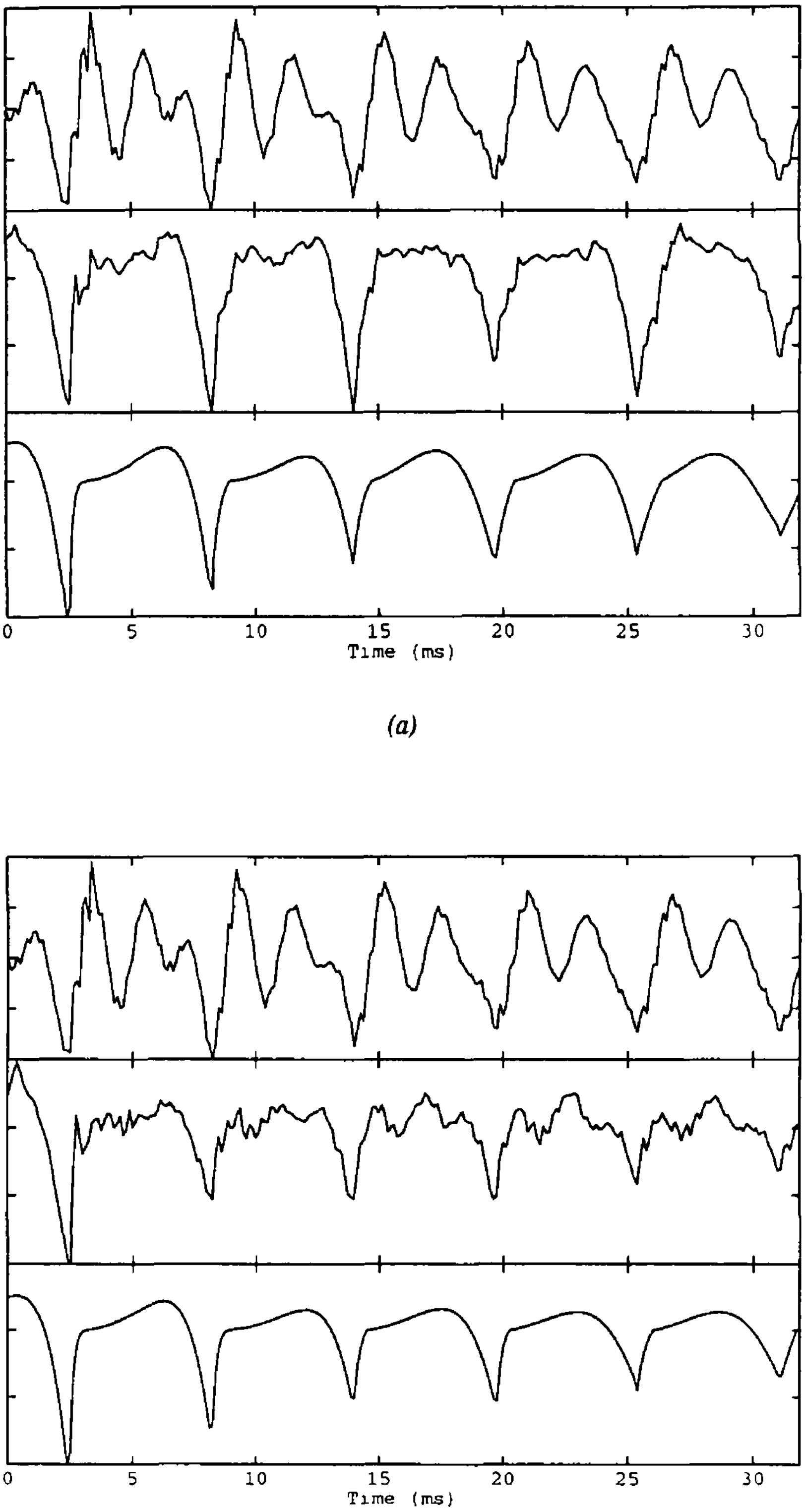
Fig 6 15 Glottal extraction applied to male [i] (a) CPIF, (b) IAIF top - speech, middle - estimated glottal waveform bottom - fitted LF waveform.



(a)

(b)

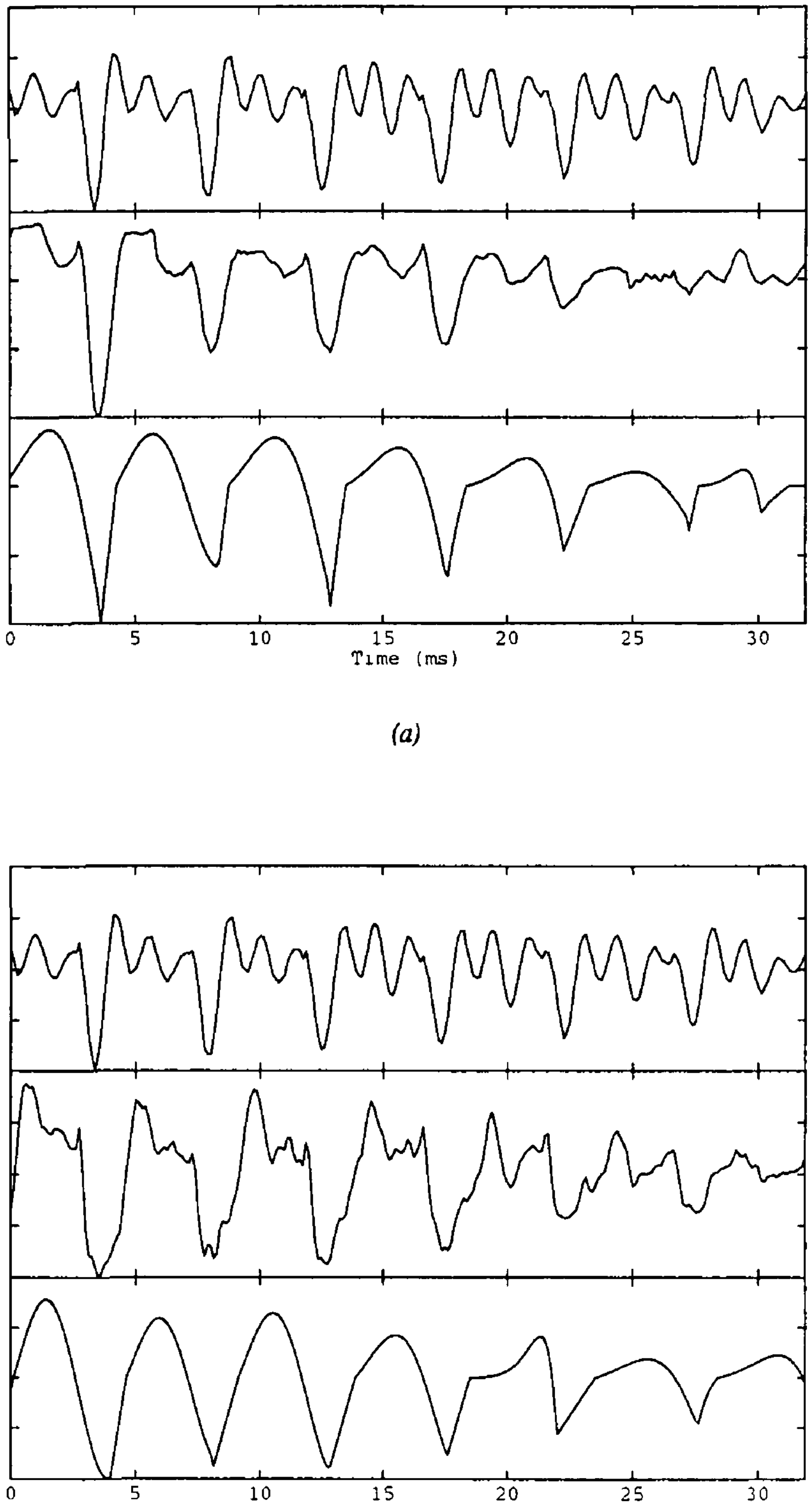
Fig 6 16 Glottal extraction applied to female [i] (a) CPIF, (b) IAIF top - speech, middle - estimated glottal waveform bottom - fitted LF waveform.



(a)

(b)

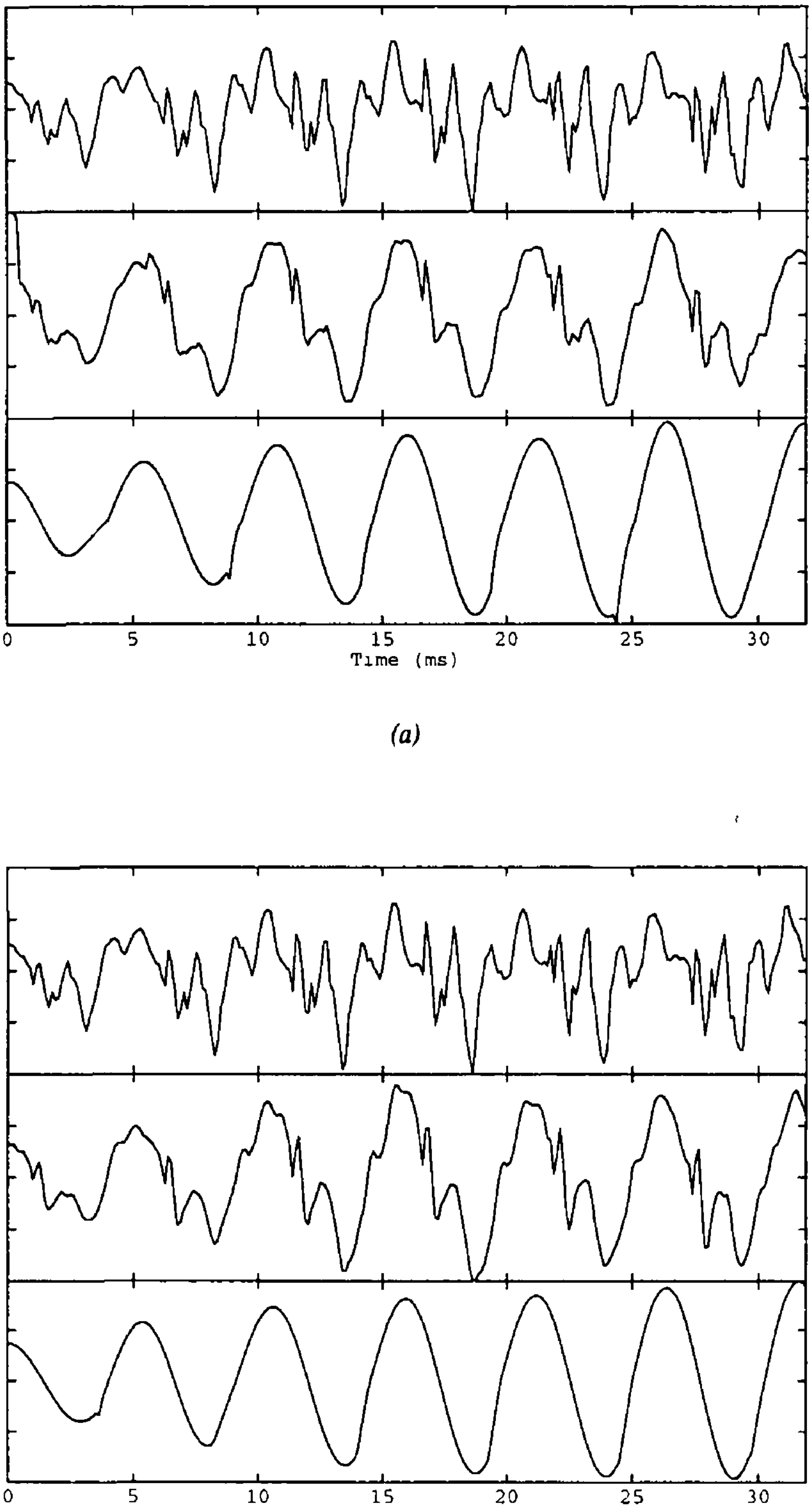
Fig 6 17 Glottal extraction applied to female [æ] (a) CPIF, (b) IAIF top - speech, middle - estimated glottal waveform, bottom - fitted LF waveform.



(a)

(b)

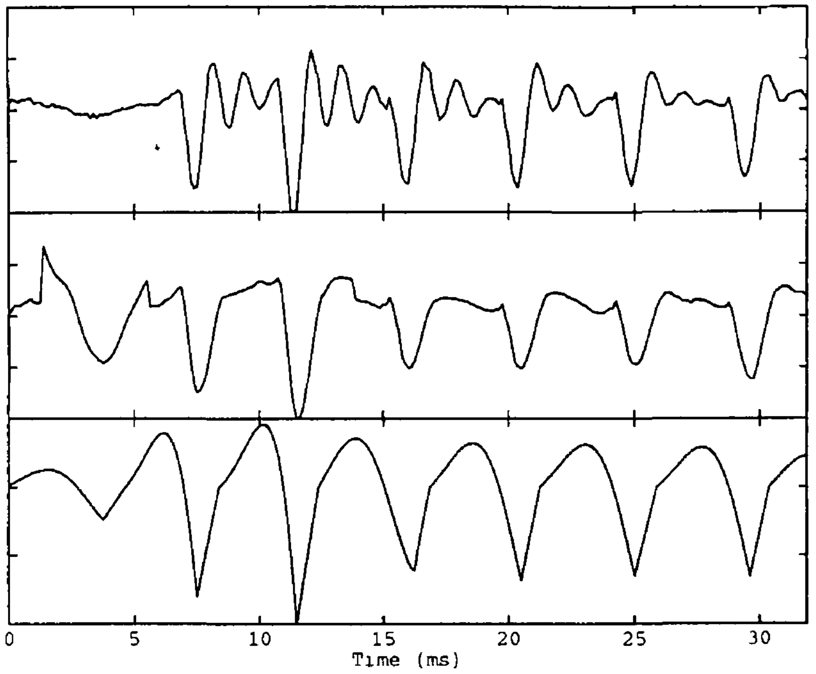
Fig 6 18 Glottal extraction applied to male [u] (a) CPIF, (b) IAIF top - speech, middle - estimated glottal waveform, bottom - fitted LF waveform.



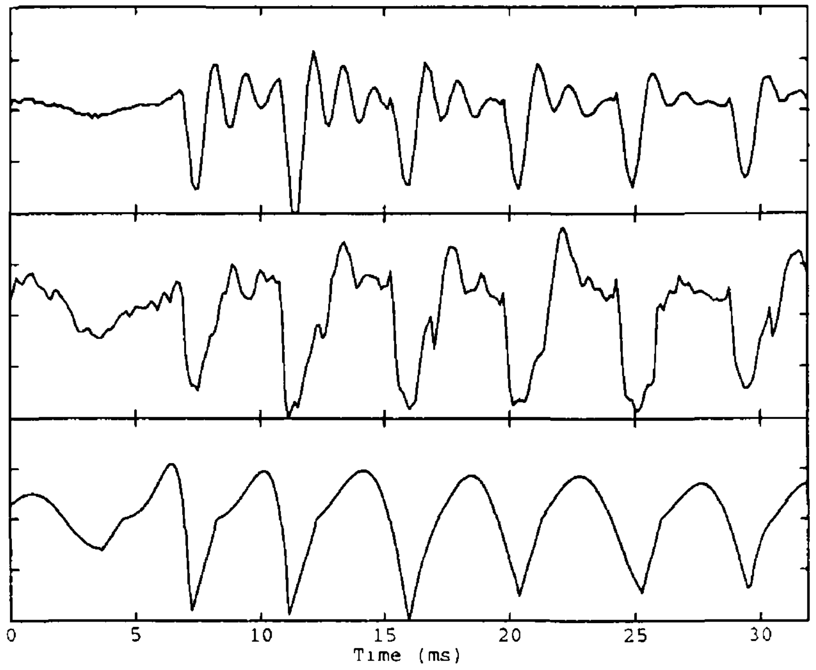
(a)

(b)

***Fig 6 19** Glottal extraction applied to male /v/ (a) CPIF (b) IAIF top - speech, middle - estimated glottal waveform bottom - fitted LF waveform.*

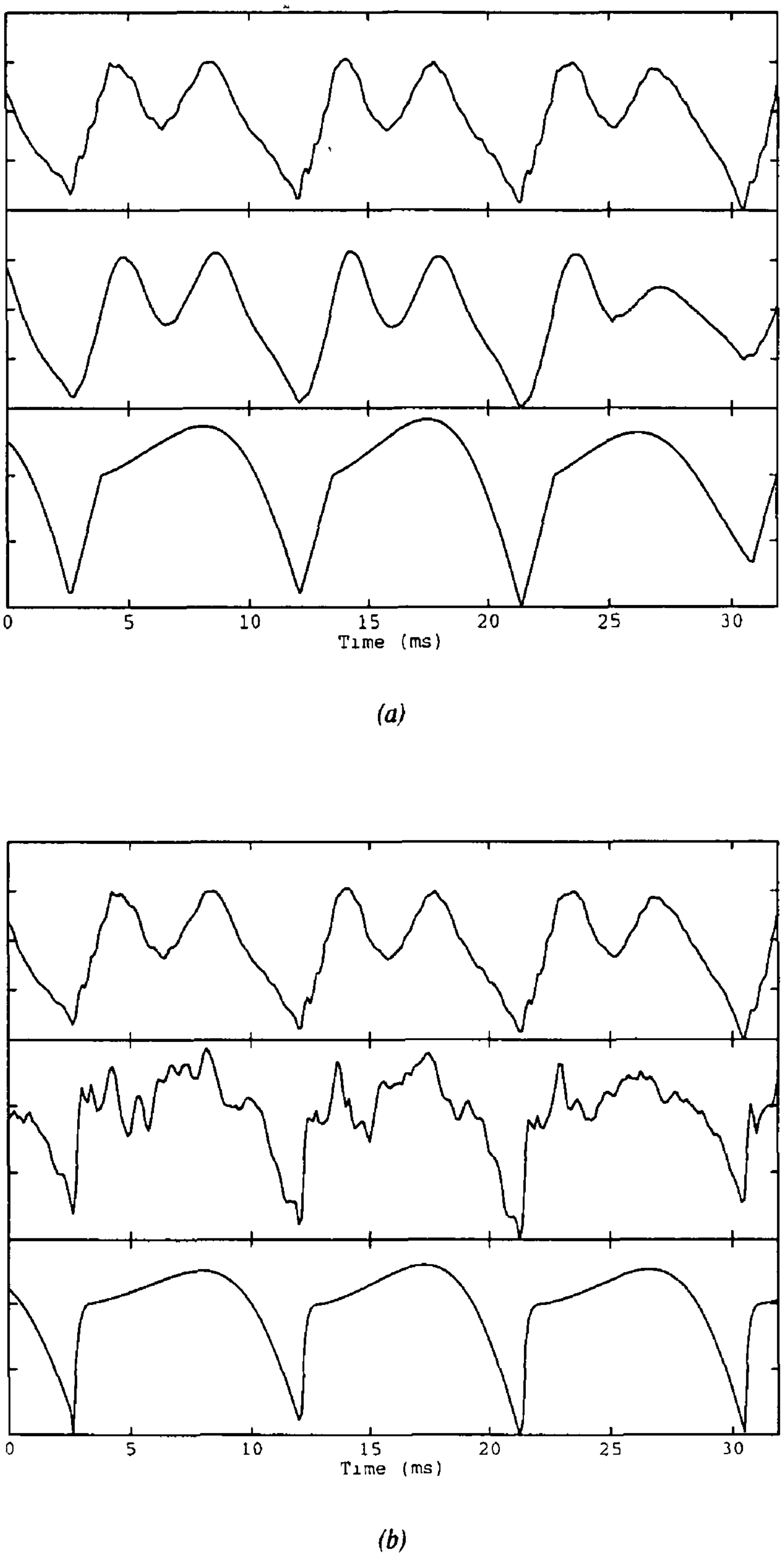


(a)



(b)

Fig 6 20 Glottal extraction applied to male [b] (a) CPIF, (b) IAIF top - speech middle - estimated glottal waveform, bottom - fitted LF waveform.



(a)

(b)

Fig 6 21 Glottal extraction applied to male [m] (a) CPIF, (b) IAIF top - speech, middle - estimated glottal waveform bottom - fitted LF waveform.

plosive, the excitation is sudden and so may not be well represented by the LF model. Fig. 6.20 shows the inverse filtering algorithms applied to the voiced plosive [b]. It is difficult to determine which algorithm performs best - both perform poorly at the start of the phoneme. Examining the first two glottal pulses, it can be seen that the LF model is inappropriate for representing the first excitation pulse at the start of the phoneme.

Inverse filtering of nasals is often inaccurate due to the presence of zeros in the vocal tract transfer function. This is illustrated in Fig. 6.21 which shows the results of applying the algorithms to a male [m]. CPIF fails completely - the inverse filtered waveform is little different to the original speech. The LF waveform is, effectively, just fitted to the original speech signal. IAIF does better - a reasonable glottal waveform is extracted and LF fitting is performed correctly. Interestingly, the LF waveform generated from the IAIF glottal estimate and from the CPIF estimate are very similar, even though CPIF fails. Presumably this is coincidental. However, it should be noted that the speech signal often preserves some of the details of the glottal excitation.

In general, CPIF and IAIF, coupled with LF modelling, provide accurate glottal waveform extraction for male and female vowels. With the exception of vowels with a low first formant, IAIF outperforms CPIF, providing better formant cancellation and more reasonable glottal waveform estimates. In the case of voiced fricatives and plosives, the inverse filtering algorithms perform reasonably well. However, in these cases, the LF model is incapable of fully representing the dynamics of the excitation. For nasal sounds, CPIF fails completely while IAIF produces useful results.

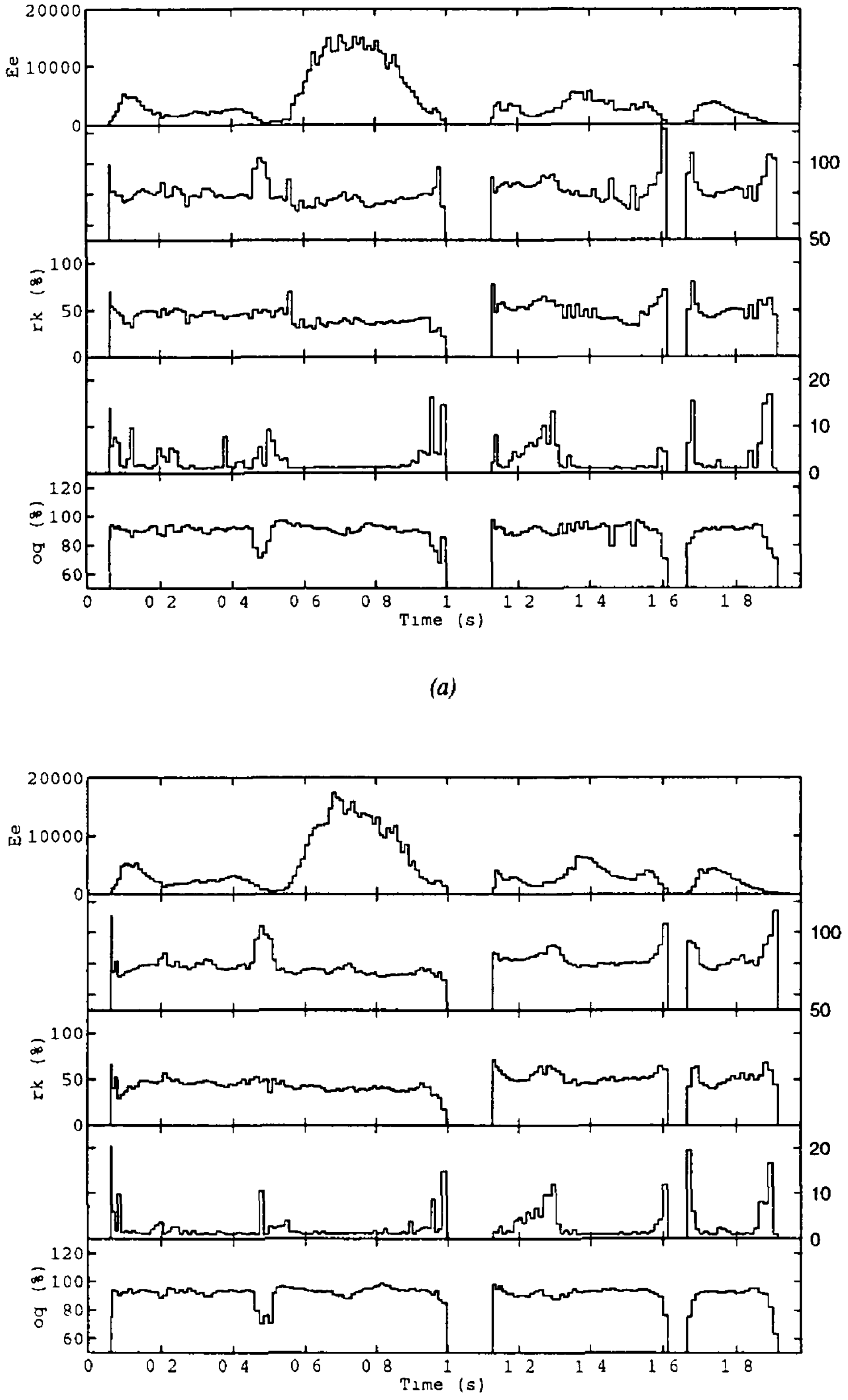
6.3.4 Continuous Speech

The performance of the inverse filtering and LF fitting algorithms was further investigated by applying them to a sentence of all-voiced speech recorded by a male and a female subject. The speech was recorded under noiseless conditions and was subsequently corrupted by the addition of white noise and reverberation.

Fig. 6.22 (a) and (b) show the LF parameter tracks estimated for the noiseless male recording by CPIF and IAIF, respectively. The parameter tracks produced by the two algorithms are generally very similar. However, those produced by IAIF are much smoother than those generated by CPIF. During steady voicing regions, the parameters estimated by CPIF change rapidly on a period-by-period basis, whereas those estimated by IAIF are nearly constant. Notably, at transients such as onsets and offsets, the parameters estimated by IAIF change as rapidly as those produced by CPIF. This suggests that IAIF is capable of identifying rapid changes in the glottal waveform but produces more consistent results when the rate of change is low.

The LF parameter tracks estimated for the female speech are shown in Fig. 6.23. Again, the general shape of the tracks are similar. In this case, the variation of the parameters estimated by IAIF is little different from that of those estimated by CPIF.

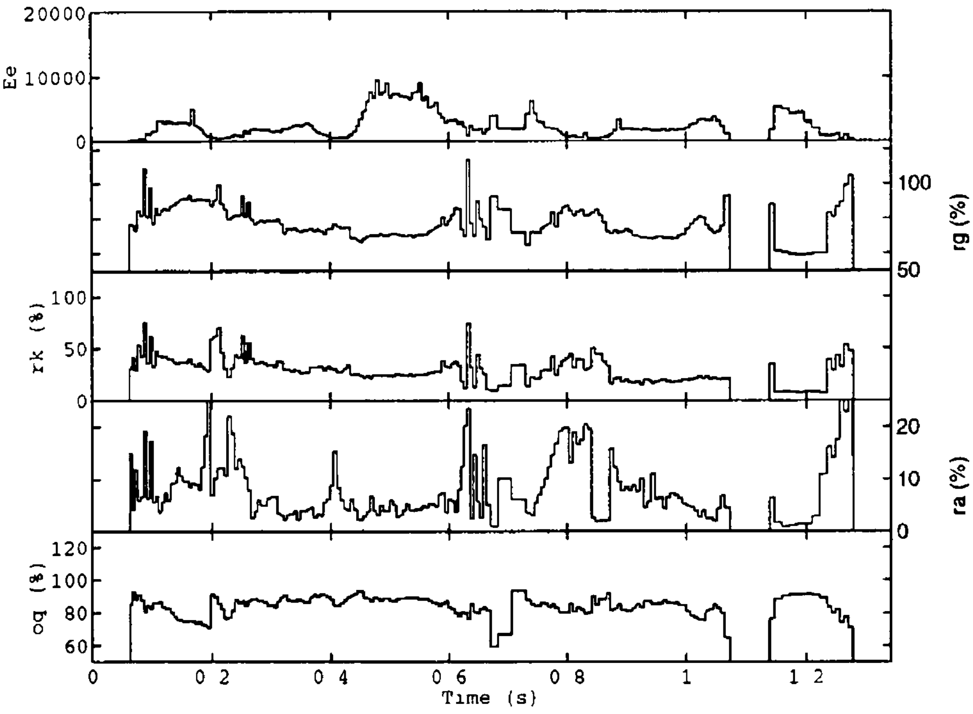
Figs. 6.24 and 6.25 show the distribution of the LF parameters extracted for the male and female subjects by CPIF and IAIF, respectively. Table 6.4 gives the mean and standard deviation of the LF analysis parameters. Clearly, the distributions produced by CPIF and IAIF are very similar. This indicates that neither algorithm has a bias for producing waveforms of a particular shape. Notably, for



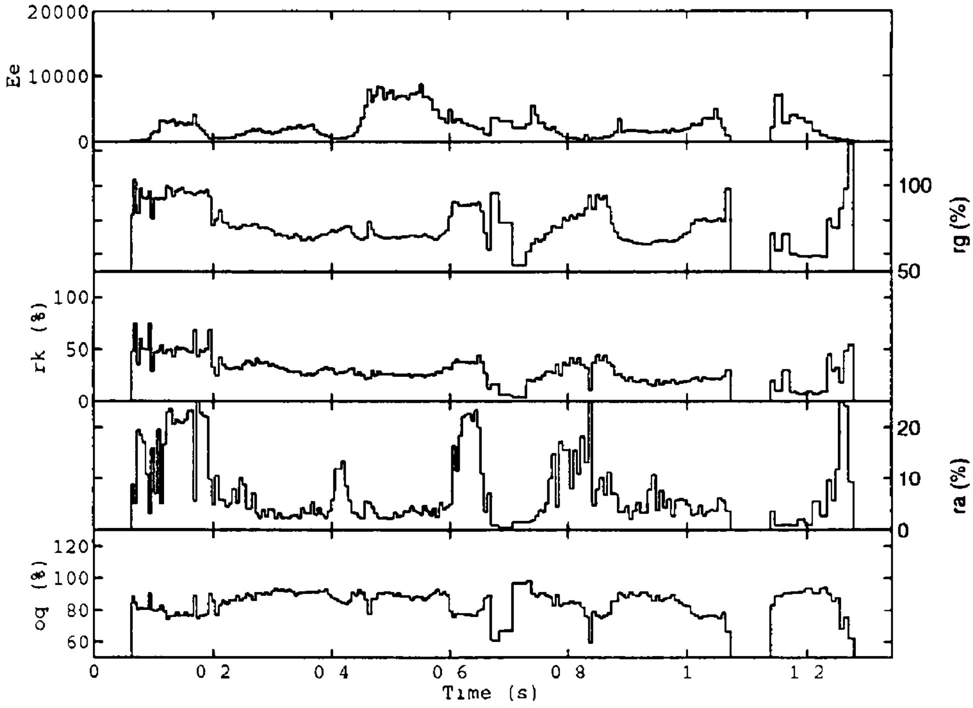
(a)

(b)

***Fig 6 22** LF parameter tracks extracted for male subject noiseless recordings (a) CPIF (b) IAIF*

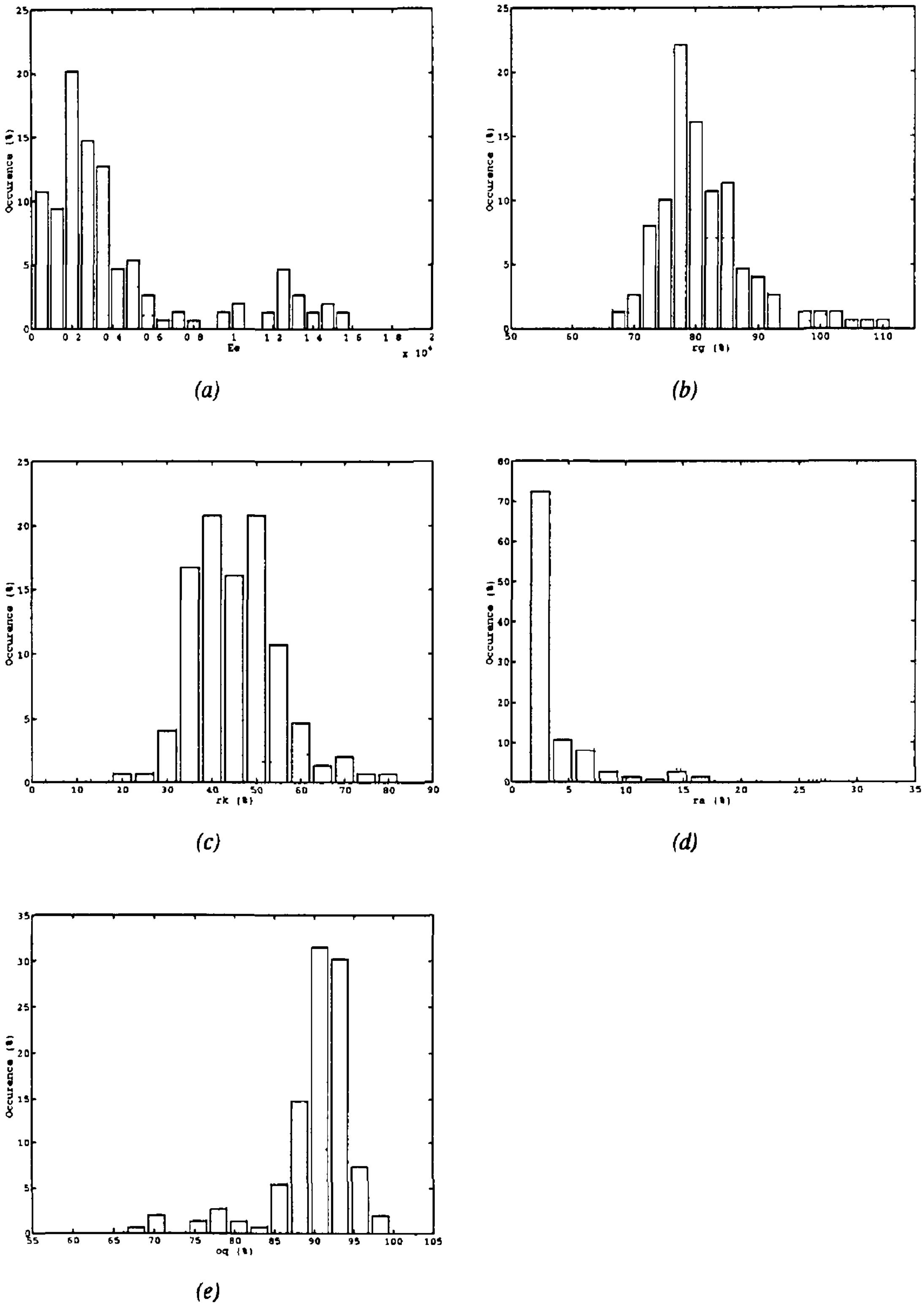


(a)



(b)

Fig 6 23 LF parameter tracks extracted for female subject, noiseless recordings (a) CPIF, (b) IAIF



(a)

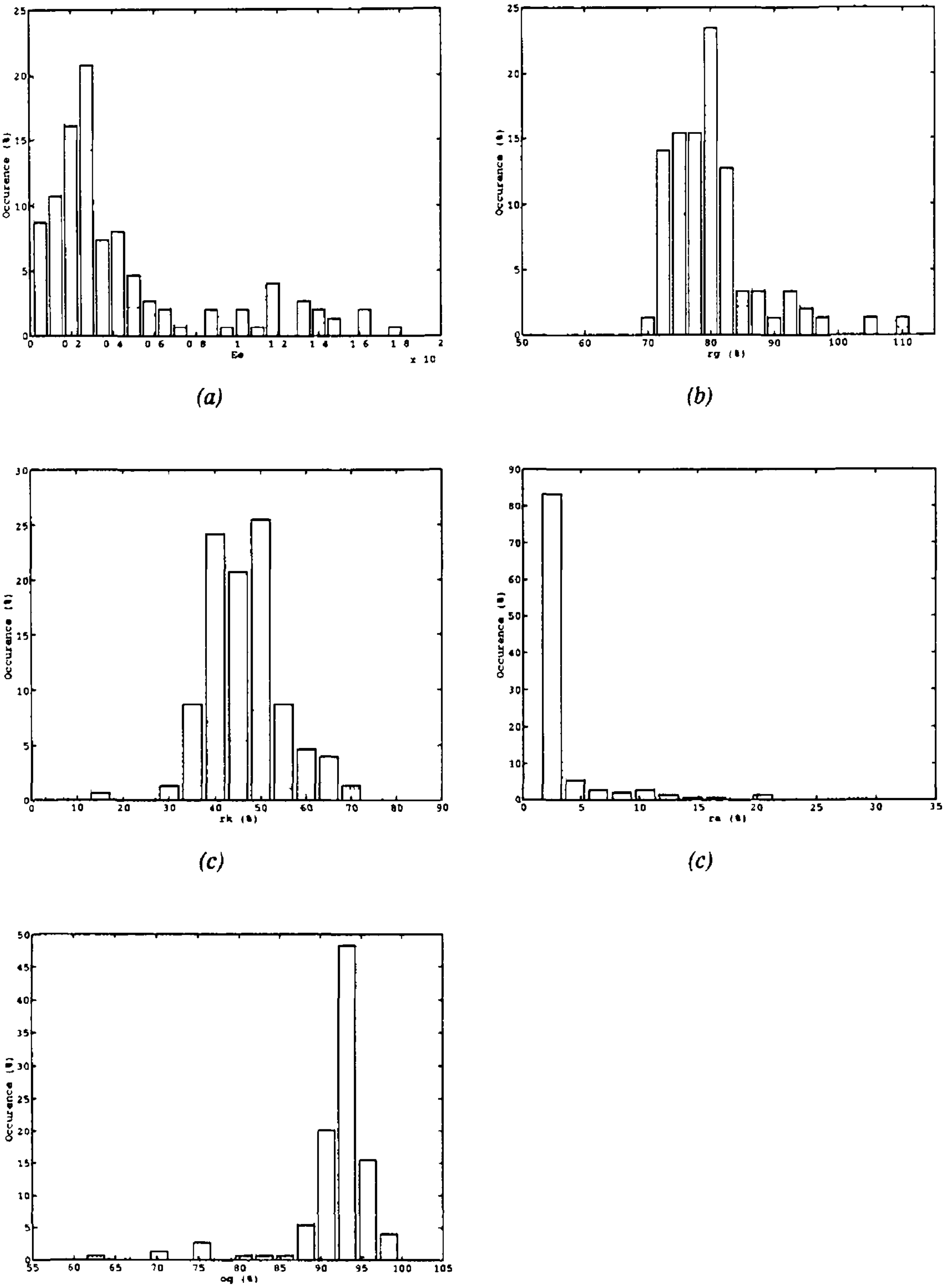
(b)

(c)

(d)

(e)

Fig 6 24 Distribution of LF parameters extracted by CPIF (a) glottal gain E_e *(b) glottal frequency* r_g*, (c) skew factor* r_k*, (d) dynamic leakage* r_a*, (e) open quotient* o_q *solid - male subject, dotted - female subject*



(a)

(b)

(c)

(c)

(e)

Fig 6 25 Distribution of LF parameters extracted by IAIF (a) glottal gain E_e*, (b) glottal frequency* r_g*, (c) skew factor* r_k*, (d) dynamic leakage* r_a*, (e) open quotient* o_q *solid - male subject, dotted - female subject* the male subject, CPIF produces a greater standard deviation for three out of the four parameters In contrast, for the female subject, IAIF produces greater standard deviation for all the parameters

	CPIF				IAIF			
	rg	*rk*	*ra*	*oq*	*rg*	*rk*	*ra*	*oq*
Male subject	81 8 6	44 10 5	2 9 3 1	89 5 5	80 7 4	47 8 2	2 7 3 4	92 5 1
Female subject	78 9 2	30 12 1	8 0 5 8	84 5 6	78 11 0	31 12 2	8 0 6 9	85 6 9

***Table 6 4** LF waveform parameters (%) automatically estimated from noiseless recordings within each row, top line - mean, bottom line - standard deviation*

The values calculated for the LF and re-synthesised SNR are shown in Table 6 5 Examining the results, it can be seen that CPIF performs better than IAIF for the female subject However, IAIF outperforms CPIF for the male subject It was originally thought that CPIF would perform better for the male subject than for the female subject Since male speech is of lower pitch than female, the closed phase tends to be longer and so it was thought that the closed phase filter estimates would be more accurate for the male subject It is postulated that the breathiness in the male voice leads to poor vocal tract filter identification during the closed phase and to inaccurate inverse filtering Notably, IAIF seems to be robust to the breathiness of the male voice Apparently, the longer analysis window serves to average out the effects of the noise In addition, the high fundamental frequency in the case of the female recording may prevent accurate discrimination of the first formant in the case of IAIF

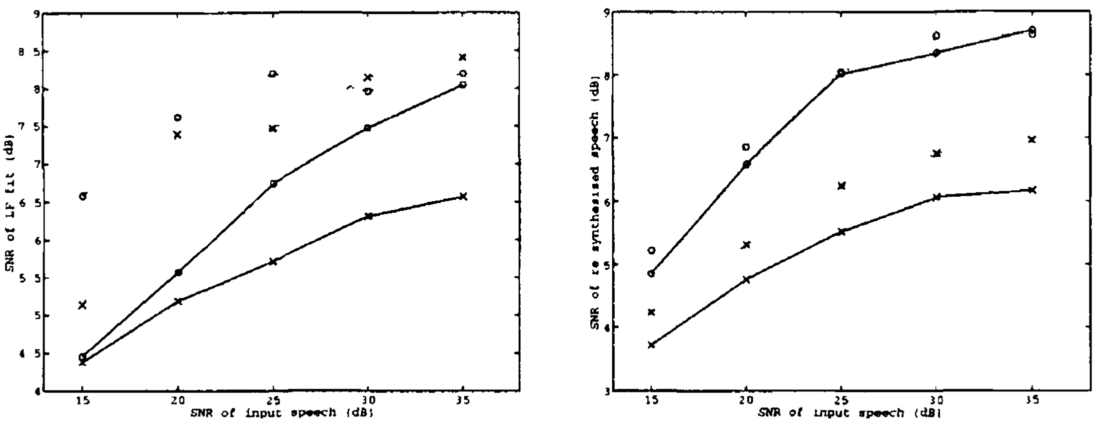
	SUBJECT	CPIF	IAIF
LF	Male	6 84	8 57
LF	Female	8 51	8 09
Re-synthesised	Male	6 76	7 56
Re-synthesised	Female	8 71	8 47

***Table 6 5** LF and re-synthesised SNR, noiseless recordings*

The performance of the algorithms was tested on speech degraded by noise White noise was added to the noiseless recordings to produce speech data with signal to noise ratios (SNR) of 35, 30, 25, 20, and 15 dB

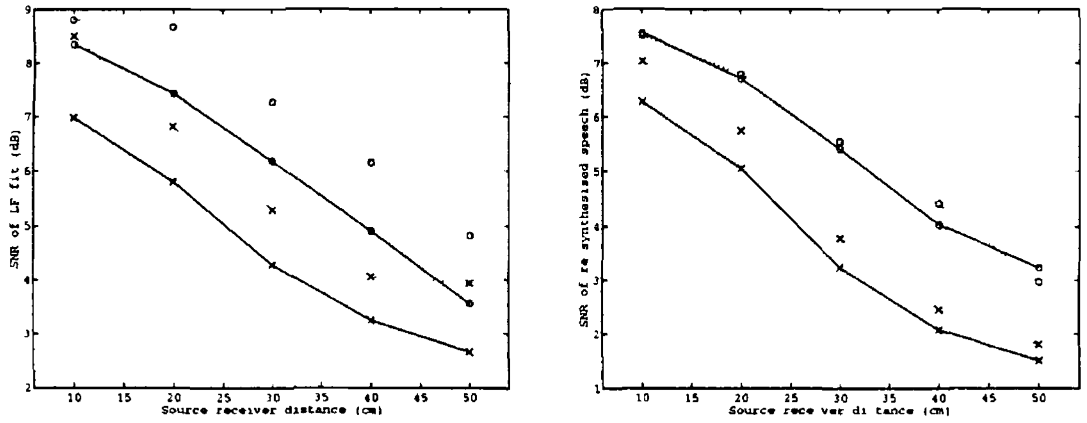
Fig 6 26 (a) shows the variation of LF SNR with the nominal SNR of the input speech For both the male and female speech, IAIF outperforms CPIF at all noise levels Again, the CPIF shows poor performance in processing the male speech compared to the female speech The re-synthesised SNR is shown in Fig 6 26 (b) In the case of the male speech, IAIF outperforms CPIF by a constant 1 dB In contrast, for the female speech IAIF only gains a 0 5 dB advantage at the highest input noise level tested

Comparing the SNR results for the LF fit and the re-synthesised speech it can be seen that the accuracy of the LF fit is governed by the choice of inverse filtering algorithm In contrast, the quality of the re-synthesised speech is governed by the speech material This can be explained by considering the nature of the various algorithms Since the IAIF algorithm estimates the vocal tract filter over the entire



(a) (b)

6 26 Variation of mean periodic output SNR with input noise (a) LF SNR (b) re-synthesised SNR solid - CPIF, dotted - IAIF, x - male subject, o - female subject



(a) (b)

6 27 Variation of mean periodic SNR with reverberation (a) LF SNR, (b) re-synthesised SNR solid - CPIF, dotted - IAIF x - male subject, o - female subject pitch period, it is less susceptible to noise than CPIF Therefore, under conditions of noise, it outperforms CPIF in terms of the accuracy of the extracted waveform In contrast, for the quality of the re-synthesised speech, the breathiness of the male speech becomes the limiting factor The breathiness cannot be reproduced by the LF model Hence, the quality of the re-synthesised female speech exceeds that of the re-synthesised male speech at all noise levels Nevertheless, for the male and female subject individually, the greater accuracy of the LF waveform extracted by the IAIF algorithm leads to a higher SNR for the speech re-synthesised from the IAIF glottal excitation estimates

Reverberation was added to the noiseless recordings to produce speech data corresponding to recordings made in a normal room at source-receiver distances of 10, 20, 30, 40 and 50 cm The glottal extraction algorithms were then applied as before

The LF SNR was calculated and is plotted in Fig 6 27 (a) As under conditions of white noise, IAIF outperforms CPIF, in this case by a margin of approximately 1 dB at 50 cm As before, the longer analysis window of IAIF makes it more robust to distortion Both algorithms show a graceful degradation in performance with increasing source-receiver distance, around 1 dB per 10 cm This is encouraging for the application of these techniques to speech coding, where drastic errors in speech reproduction must be avoided

As before, speech was re-synthesised from the LF data Fig 6 27 (b) shows how the re-synthesised SNR varies with respect to the simulated source-receiver distance In the main, the re-synthesised SNR follows the trend determined by the quality of the LF fit IAIF performs slightly better than CPIF and the degradation with source-receiver distance is gradual

Overall, the experiments indicate that, under noisy and reverberant conditions, IAIF performs more accurate LF extraction than CPIF This leads to higher quality re-syntheses of the speech material from the glottal waveform extracted by IAIF Both algorithms show increasing errors in glottal extraction with increasing distortion Fortunately, in both cases, the performance degradation is gradual

6.4 DISCUSSION

In this section the LF data obtained in these experiments is compared with that obtained in previous studies by other authors A number of earlier investigations extracted the LF parameters from voiced speech using manual inverse filtering and waveform fitting Comparison of the results herein with those obtained manually allows the accuracy of the automatic algorithms to be established Furthermore, the automatic inverse filtering and fitting procedures facilitate the processing of much larger amounts of speech material than is possible manually Thus the LF data extracted in these experiments constitutes a substantial database for investigating the nature of the glottal excitation

Encouragingly, the LF parameter tracks extracted by the automatic algorithms in this study show similar trends to those identified manually in an investigation by Gobl [Gobl, 1988] In his paper, Gobl reports the results of manual inverse filtering and LF fitting experiments performed on 12 seconds of continuous Swedish speech recorded by three male subjects Gobl notes that when voicing is terminated, for example by a pause, the value of E_e decreases and the values of r_a and r_k increase Furthermore, Gobl states that maximum values of r_a are normally found at the termination of the sentence These effects can be clearly seen in Figs 6 22 and 6 23 Gobl also records higher than normal values for r_a and r_k at voice onsets These features are again prominent in the results recorded by the automatic algorithms In addition, Gobl finds evidence for glottal waveform variation due to the stress and phonetic content of the utterance This conclusion is supported by the parameter tracks presented above

The LF parameter distributions for the male recording fall within the range recorded by Gobl in the same study However, the mean values deviate from those presented by Gobl In another study, Gobl investigated the relationship between glottal waveform shape and voice quality [Gobl, 1989] In this work, he recorded a British male phonetician reproducing the phrase "Say babber again" using four different voice types - modal, breathy, whispery and creaky Manual inverse filtering and LF model fitting was performed on the vowel [æ] extracted from the nonsense word babber The means and standard deviations of the extracted LF parameters are reproduced in Table 6 6 The mean parameters obtained in this study for the male subject, Table 6 4, correlate most strongly with those of the breathy voice type This finding confirms the earlier subjective supposition that the male speech is of breathy type and supports the accuracy of the automatic inverse filtering and LF fitting algorithms In all likelihood, the differences between the mean breathy parameter values and those measured here for the male subject occur due to the differing phonetic content of the data sets

	rg	rk	ra	oq
Modal voice	117 7 4	34 1 0	1 0	55 2 2
Breathy voice	88 2 4	41 1 9	2 5 0 6	81 1 2
Whispery voice	94 3 8	32 2 9	7 1 2	70 3 8
Creaky voice	113 11 0	20 2 4	0 8 0 5	53 4 1

Table 6 6 *LF waveform parameters (%) manually estimated from the male vowel [æ] - mean and standard deviation [after Gobl, 1989]*

In a cross-language study, Gobl and Ní Chasaide extracted LF parameters from female speech by manual inverse filtering [Gobl and Ni Chasaide, 1988] In the text of the document, they state that the female voice displays a r_a 2-4 times higher than the male Also they found that r_g values should be 10-20 % lower than the male and r_k values tend to be either the same or slightly higher These correction factors make the parameter values recorded for the female subjects in this investigation Table 6 4 compatible with Gobl's results for the modal male voice, Table 6 6 In another study Karlsson [Karlsson, 1988] manually extracted the LF parameters for two vowels phonated by seven female subjects The mean results are shown in Table 6 7 The values of the r_k and r_a parameters are very close to those reported for the female subjects in this study However, the r_g value is significantly greater than, and the o_q value is much less than, those reported here This seems to be due to a difference in the LF fitting procedure Karlsson's model allows the instants t_c and t_0 to be placed at different times, creating a period of zero flow between glottal cycles As stated earlier in this investigation the instant t_c from the previous period is set equal to the instant t_0 from the current period This leads to lower values for t_p and t_e in Karlsson s study This, in turn, increases the value of r_g and decreases the value of o_q in

Karlsson's study, relative to those in this investigation Allowing for this discrepancy, Karlsson's data seems to confirm the modal nature of the female speech

	rg	*rk*	*ra*	*oq*
Female subjects	103	30	9 9	64

Table 6 7 *Mean LF waveform parameters (%) manually estimated from two vowels recorded by seven female subjects [after Karlsson, 1988]*

Table 6 8 shows the correlation coefficients calculated between the various glottal parameters over the noiseless recordings, both male and female These relationships are of interest, not only for providing an understanding of the underlying dynamics of the voice source, but also for the purposes of vector quantisation of the glottal parameters

Surprisingly, the fundamental frequency F_0 displays a negative correlation with the glottal gain E_e The negative correlation observed may be due to the small size of the test data The female speech is more highly pitched and softer than the male Thus, there is some correlation between the presence of a high fundamental frequency and low glottal amplitude, cf Fig 6 24 (a) In general, it is expected that, at normal voicing levels, the pitch and amplitude are controlled by the speakers so as to provide semantic information to the listener and are therefore determined independently of each other, based on the linguistic content of the utterance

In an investigation using CPIF and an adaptive non-linear least-squares LF fitting algorithm, Strik and Boves [Strik and Boves, 1992] measured the correlations between a number of LF parameters in continuous spontaneous male speech They found the following correlations with pitch period T_0 r_g 0 04, r_k 0 22, and r_a 0 18 The correlation between T_0 and r_g is close to that determined by CPIF in this investigation As well as this, the correlation coefficient for T_0 and r_k approximates that found in this study However, the correlation between T_0 and r_a determined by Stik and Boves is opposite to that found herein This difference could be due to the differing phonetic content of the data.

The timing parameters r_g, r_k and r_a all show negative correlations with the glottal gain E_e This can be explained by considering the physiological process of glottal fold vibration Increased glottal gain corresponds to more rapid glottal closure To achieve this, the folds must move together more quickly This leads to greater skewing and reduced r_g and r_k Additionally, faster closure leads to firmer sealing of the vocal folds and reduced leakage, corresponding to a shorter return time and decreased r_a Previously, these effects have been noted by Gobl, Ni Chasaide and Pierrehumbert [Gobl, 1988, Gobl and Ni Chasaide, 1988, Pierrehumbert, 1989] Linked to this is the strong covariation between r_g and r_k, both of which control glottal skewing Similarly, there is a reasonably strong correlation between r_g and r_a In general, greater skewing is associated with faster closure and less dynamic leakage

The strong negative correlation between the fundamental frequency F_0 and the open quotient o_q has been previously identified in semi-automatic experiments by Lobo and Ainsworth [Lobo and Ainsworth, 1988] The open quotient o_q also shows a strong negative correlation with the glottal frequency r_g and the return time r_a Assuming relatively constant skew, a longer opening phase (i e an increased t_p) leads to a longer overall open phase and so to an increased t_e Hence, o_q is inversely related to r_g In the case of breathy speech, the glottis barely closes As a result the return time t_a extends until the opening instant of the next period t_0, therefore o_q+r_a=1 This effect also occurs, to lesser extent, in modal speech Thus, an overall negative correlation exists between the open quotient o_q and the dynamic leakage r_a, and between the open quotient o_q and the fundamental frequency F_0

	Fo	*Ee*	*rg*	*rk*	*ra*
Ee	-0 21 -0 23	-	-	-	-
rg	-0 04 0 08	-0 33 -0 27	-	-	-
rk	-0 28 -0 25	-0 17 -0 07	0 73 0 68	-	-
ra	0 43 0 51	-0 42 -0 34	0 35 0 52	-0 02 0 05	-
oq	-0 35 -0 40	0 28 0 30	-0 48 -0 62	0 24 0 15	-0 56 -0 67

***Table 6 8** Correlation coefficients calculated between LF parameters top line - CPIF, bottom line - IAIF*

6 5 CONCLUSION

This chapter investigates the performance of two algorithms for glottal waveform extraction - Closed Phase Inverse Filtering and Iterative Adaptive Inverse Filtering The glottal waveform estimates provided by the two inverse filtering algorithms were parameterised using the LF model via a time-domain least squared error fitting procedure

Two error measures were proposed for determining the accuracy of the extracted glottal waveform The LF SNR was defined as the mean of the SNR calculated between the fitted LF waveform and the estimated glottal waveform, averaged over each pitch period of the signal Similarly, the re-synthesised SNR was defined as the mean of the SNR calculated between the speech signal re-synthesised from the LF waveform and the original speech, averaged over each pitch period In all experiments, a strong correlation was found between both of these measures

The LF fitting technique proceeded by matching a polynomial LF approximation to the glottal estimate Based on this initial fit, a multi-dimensional optimisation routine was employed to find the parameters which minimised the mean squared error between the LF waveform and the glottal estimate The effects of optimising the LF fit were investigated by determining the LF and re-synthesised SNRs for the glottal waveform extracted with and without optimisation In all cases, optimisation gave an improvement in SNR However, the fit provided by the polynomial LF approximation also gave good results Thus, optimisation is useful but the parameters obtained from the polynomial fit are appropriate for use when the optimisation routine is too computationally complex to be employed

A new multiple filter procedure was proposed to improve the robustness of CPIF The scheme involved using filters estimated over the current, previous and second previous closed phases and an average of the three The inverse of each filter was applied to the current pitch period and a polynomial was fitted to the estimated glottal waveforms The filter giving the least squared error between the polynomial and the estimated glottal signal was chosen as best for that period In experiments, CPIF was performed on natural speech with and without the multiple filter procedure The results showed that the multiple filter procedure improved the LF SNR and re-synthesised SNR Furthermore, the improvement due to the procedure increased with increasing distortion in the speech signal The increasing impact of the multiple filter procedure was emphasised by results showing that the number of times the average filter is chosen increases dramatically with increasing distortion

The performance of the CPIF and IAIF algorithms was compared in tests on various categories of speech sounds for both male and female subjects The findings indicated that, for most voiced sounds, both CPIF and IAIF provide adequate glottal waveform extraction However, apart from during vowels with a low first formant, IAIF performs slightly better than CPIF, providing more effective formant cancellation and smoother glottal waveform estimates The zeros introduced into the transfer function of the vocal apparatus due to nasalisation cause CPIF to fail In contrast, IAIF provides useful results During voiced fricatives and plosives, although the inverse filtering algorithms seem to perform satisfactorily, the LF model lacks the flexibility to accurately represent the excitation signal

In tests on the performance of the inverse filtering algorithms applied to noisy and reverberant speech, IAIF produced the best results in terms of the LF and re-synthesised SNRs This is probably due to its longer analysis window The greater window length reduces the effect of distortions by averaging them out over a greater number of samples For both algorithms, the degradation in performance with increasing distortion is gradual and graceful, a necessary criterion for applications in speech coding

The statistics of, and relationships between, the LF parameters extracted in this study show close similarity with those determined in manual investigations by other authors This finding further supports the accuracy of the automatic algorithms In addition, the results herein provide a useful database for study of the glottal excitation

The investigation described in this chapter proposes a new LF fitting procedure, provides a more robust method of CPIF, quantifies the effect of LF optimisation and establishes IAIF as superior to CPIF in terms of formant cancellation and robustness to distortion These findings provide a basis for future studies of the glottal excitation during voiced speech Furthermore, the reliability of the inverse filtering methods is such that they may be considered for incorporation in a speech coding scheme This possibility is examined in the next chapter

CHAPTER 7

GLOTTAL EXCITED SPEECH CODING

7 1 INTRODUCTION

This chapter describes experiments carried out to determine the transmission rate and speech quality achievable by a Glottal Excited Linear Prediction (GELP) coding system for voiced speech

GELP systems synthesise speech by applying a parameterised glottal waveform to a Linear Prediction vocal tract filter This approach has the potential to produce higher quality speech at a lower bit rate than conventional systems for two reasons Firstly, the parameters which describe the glottal excitation are slower time-varying than those of conventional LP residual models [Cheng and O'Shaughnessy, 1993] Thus, GELP systems should be capable of transmitting the voiced excitation more efficiently Secondly, the speech production model incorporated in GELP more closely approximates the human speech production mechanism [Pinto et al , 1989, Childers and Wu, 1990] Therefore, higher quality speech is likely to be provided by GELP

Unfortunately, there remain two obstacles to the widespread use of glottal based speech coding Firstly, glottal estimation by inverse filtering is sensitive to phase distortion in the speech signal [Holmes, 1975, Markel and Gray, 1976] Secondly, the fitting of glottal waveform models to the estimated glottal excitation can be computationally expensive [Fant et al , 1985] The investigation described herein attempts to establish an upper bound on the performance achievable using inverse filtering based GELP systems Consequently, throughout this work the recording channel is assumed to be phase linear and the normally severe run-time constraints imposed on coding systems are relaxed

The GELP system proposed in this chapter estimates the glottal excitation during voiced speech by inverse filtering Two inverse filtering techniques are studied - Closed Phase (CPIF) and Iterative Adaptive Inverse Filtering (IAIF) The estimated glottal excitation is parameterised by fitting the LF model in the time-domain The LP filter coefficients are then determined by ARX estimation The pitch, LF and filter parameters are quantised and transmitted to the decoder The GELP receiver re-synthesises the speech signal by re-generating the LF excitation and applying it to the LP filter The coding system is designed to operate at a low bit rate 2 4 - 4 8 kb/s, and to incur medium delay, approximately 0 1 s overall

The development of the quantisation scheme for the GELP coder is described in detail The filter parameters are quantised as Line Spectral Pairs, while the pitch quantisation scheme is based on that of U S Federal Standard 1016 CELP The LF parameters are quantised differentially Based on LF distributions obtained from processing natural speech, optimum non-linear quantisers are generated by minimisation of the distortion rate In order to determine the best bit allocation for the LF parameters, the speech quality and bit rate provided by a number of configurations were tested using a knockout procedure The coded speech quality was assessed by an objective quality measure, the Bark Spectral Distortion (BSD) [Wang et al , 1992] The BSD measure is based on the perceptual properties of the human auditory system and has shown good correlation with subjective quality measures, such as the Mean Opinion Score

The performance of the finalised GELP coding system was assessed in comparative tests with the standard speech coders - LPC-10, CELP and GSM All the systems were applied to the same test data and the output speech quality was assessed by the BSD measure Additionally, the robustness of the systems was determined by processing noisy and reverberant speech material

The chapter is divided into five sections Section two describes the coding and BSD systems used in this investigation Section three describes the development of the GELP quantisation scheme Section four details the experiments carried out to determine the performance and robustness of the coders Lastly, section five concludes the chapter

7.2 DESCRIPTION OF THE CODING SYSTEMS

This section describes the systems used in the investigation The operation of the GELP coder is explained, together with that of the three conventional systems - LPC-10, CELP and GSM Also, the method used for computation of the BSD speech quality measure is described

The section is split into five sub-sections, each covering one of the algorithms under consideration - GELP, LPC-10, CELP, GSM and BSD

7.2.1 GELP System

A schematic diagram of the GELP encoding system is shown in Fig 7 1 The system consists of a number of sub-processes First of all, the speech signal is segmented into 240 sample (30 ms) frames to allow block processing Automatic GCI identification is performed by Pre-emphasised Maximum Likelihood Epoch Detection (PMLED), see Chapter 5 This sub-process determines the pitch of the signal and passes a vector of Glottal Closure Instants to the inverse filtering algorithm From the incoming speech signal the inverse filtering algorithm produces an estimate of the glottal excitation Herein two algorithms are tested for this purpose, these being CPIF and IAIF, see Sections 6 2 1 and 6 2 2 respectively The LF model is fitted to each period of the estimated glottal waveform by minimisation of a mean square error criterion, see Section 6 2 3 The optimised LF parameters are passed to the quantiser and the re-generated LF waveform is used to calculate the LP coefficients for the frame The LF and speech signal vectors from the current frame are pre-emphasised, Hamming windowed and applied to an ARX procedure [Astrom and Eykhoff, 1971] This sub-process returns the all-pole coefficients which, when applied to the re-generated LF excitation provide the least squared error between the re-synthesised speech and the original The pre-windowing helps to ensure the stability of the estimated filter and minimises discontinuities between frames For stability, any poles outside the unit circle are reflected back inside After ARX estimation the glottal gain E_e of the LF excitation is optimised so as to match the energy of the re-synthesised and original speech This is done for two reasons Firstly, the gain of the inverse filter may be different to that of the ARX estimated filter Secondly, the human ear is very sensitive to the energy contour of the speech signal Thus, the coder must ensure that the energy of the re-synthesised signal closely matches that of the original recording

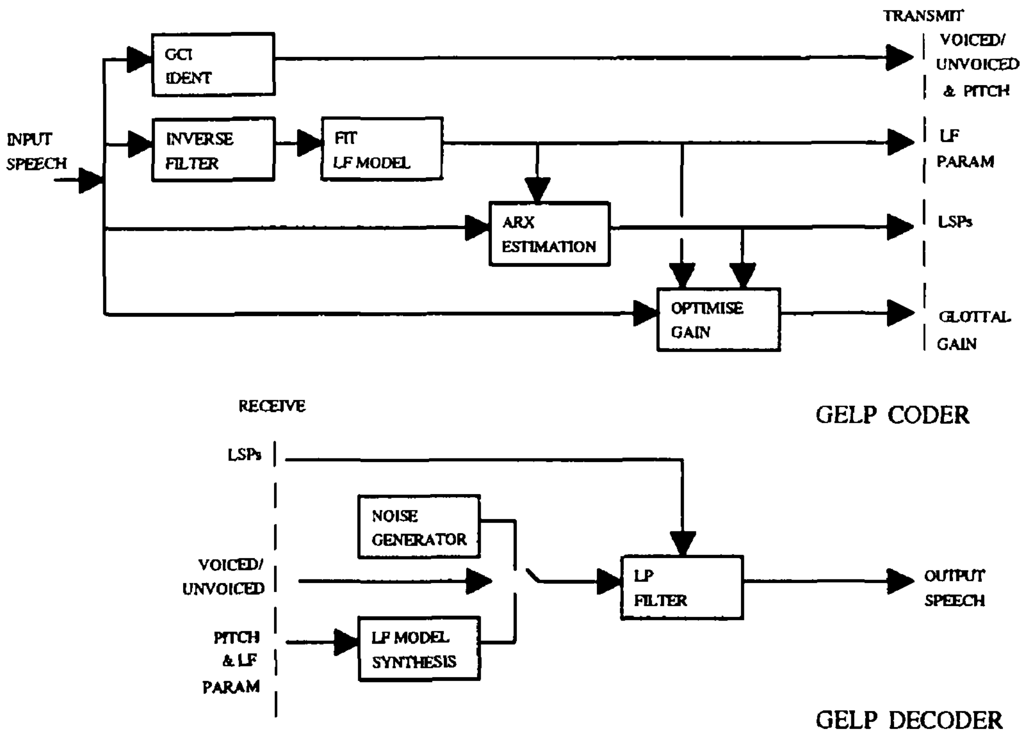


Fig 7 1 Schematic diagram of the GELP coding system.

The gain is optimised via a Simplex search of the gain quantisation levels, starting at the value closest to that originally determined during LF fitting To avoid discontinuities, the energy of the two signals is calculated from the start of the current period until half way through the next period

At the GELP receiver, the bit stream from the transmitter is decoded and the pitch period, LF parameters and LP coefficients determined The LF excitation is re-generated, based on the pitch and LF parameters, and is passed to the LP filter for re-synthesises of the speech signal

In this study only the coding of voiced speech is considered The strategy employed for unvoiced speech is similar to that of LPC-10 and consists of a simple binary voicing decision at the encoder and a white noise excitation in the decoder Although the approach is sensitive to voicing decisions errors, it does provide for low rate transmission

7 2 2 LPC-10 System

The basic LPC vocoding strategy has been used in many speech coding applications The vocoding system used in this investigation is based on the U S Federal Standard 1015 LPC-10e [Tremain, 1982, Campbell and Tremain, 1986], developed by the U S Department of Defence in the 70s The system operates at a total transmission rate of 2 4 kb/s and achieves a Mean Opinion Score of approximately 2 5 [Jayant, 1990] The system used in this study is a UNIX C implementation which was released into the public domain by the U S National Security Agency

Fig 7 2 shows a block diagram of the LPC-10 system Speech is partitioned into 180 sample (22 5 ms) frames A voicing decision is made using the Average Magnitude Difference Function (AMDF) [Ross et al , 1974], zero crossing rate, energy measures, LP reflection coefficients and prediction gains In the case of an unvoiced decision, speech is re-synthesised using a white noise

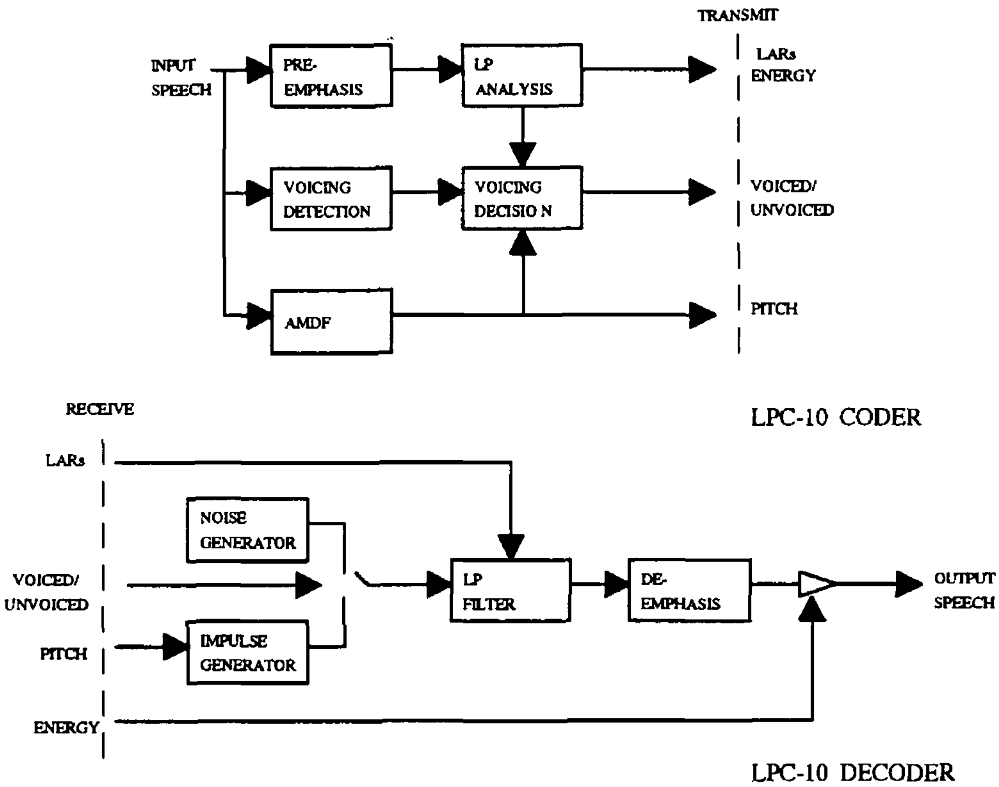


Fig 7 2 Schematic diagram of the U S Federal Standard 1015 LPC-10 speech coding system [after Tremain, 1982]

excitation to a 4th order all-pole filter In the case of a voiced decision, a train of bandpass filtered impulses at the pitch period, determined by the AMDF, is used to excite a 10th order all-pole filter The coefficients of the LP filter are obtained by covariance analysis over the pre-emphasised speech signal The voicing decision, pitch and LP coefficients are quantised and transmitted to the receiver Additionally, the energy of the frame is forwarded to the decoder

The decoder proceeds, depending on the voicing indicator, either by generating a white noise excitation or by generating an impulse train at the pitch period The appropriate signal is passed through the synthesis filter determined by the received LP coefficients, de-emphasised and scaled to match the energy of the original speech signal

The quality of the speech produced by LPC-10 is limited by its excitation model The impulse train is a very coarse approximation to the true glottal waveform Partly as a result of this, LPC-10 coded voiced speech has a "buzzy" quality Also, the binary voicing decision leads to poor sound quality for phonemes which naturally have a mixed excitation, for example voiced fricatives Moreover, the scheme produces extremely distorted speech when the voicing decision is incorrect

7.2 3 CELP System

The CELP coding system used in this study was developed by the U S Department of Defence U S Federal Standard 1016 [Campbell et al, 1991, Fenichel, 1991, 1992] The system was proposed in 1989 and is based on analysis-by-synthesis techniques proposed by Atal and Schroeder in the mid 80s [Schroeder and Atal, 1985] The system achieves a Mean Opinion Score of 3 0 at a bit rate of 4 8 kb/s

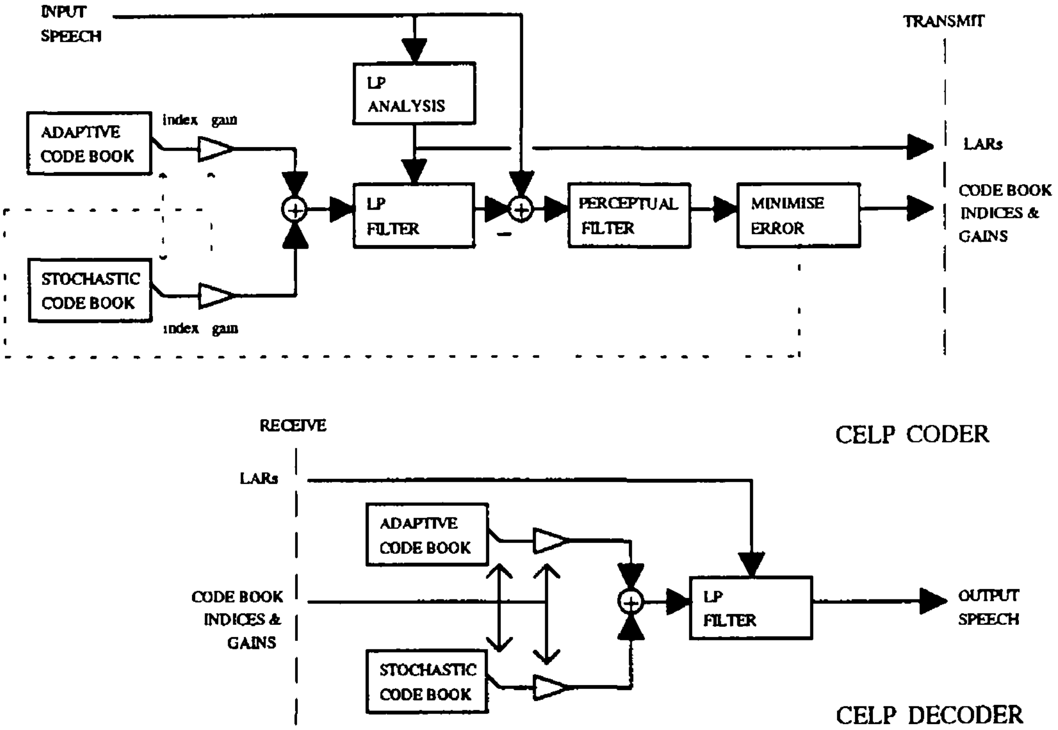


Fig 7 3 Schematıc dıagram of the U S Federal Standard 1016 CELP speech codıng system [after Fenıchel, 1991]

[Jayant, 1990] The ımplementatıon used for these experıments was wrıtten ın UNIX C and was placed m the publıc domaın by the U S Natıonal Securıty Agency

Schematıcs of the CELP coder and decoder are shown ın Fıg 7 3 The coder uses a 240 sample (30 ms) frame sıze wıth four equal sub-frames CELP analysıs consısts of three basıc functıons - short-term Lınear Predıctıon analysıs, long-term adaptıve codebook search and stochastıc codebook search Lınear Predıctıon analysıs ıs performed once per frame by open-loop, 10th order autocorrelatıon wıth a 30ms Hammıng wındow The adaptıve codebook ıs searched exhaustıvely to determıne the delay and gaın whıch, when applıed to the prevıous excıtatıon and passed through the synthesıs filter, mınımıses a perceptually weıghted predıctıon error crıterıon The predıctıon error sıgnal ıs calculated as the dıfference of the re-synthesısed sıgnal and the orıgınal speech The predıctıon error ıs perceptually weıghted by passıng ıt though a modıfied versıon of the LP filter Thıs perceptual weıghtıng filter ıs desıgned to de-emphasıse errors whıch are masked by the speech sıgnal and whıch are, therefore, ınaudıble to the lıstener The resultıng perceptually weıghted predıctıon error sıgnal ıs squared and summed to provıded the perceptually weıghted predıctıon error crıterıon The stochastıc codebook search proceeds ın a sımılar manner Each of 512 stochastıc excıtatıons ıs scaled, added to the adaptıve excıtatıon and passed through the LP filter Agaın, the optımum ınnovatıon ıs chosen as the entry gıvıng the mınımum perceptually weıghted predıctıon error A sparse (70% zero), overlapped and ternary valued (-1,0,+1) codebook allows fast computatıon The ındıces of the optımum codebook entrıes the gaın for each codebook and the coefficıents of the LP filter are quantısed and transmıtted to the receıver

The CELP decoder generates the adaptıve excıtatıon by delayıng prevıous ınnovatıons and scalıng them accordıng to the transmıtted adaptıve gaın Sımılarly, the stochastıc excıtatıon ıs created by selecting the appropriate entry from the stochastic codebook and scaling it by the stochastic gain The speech signal is re-synthesised by summing the excitations and applying them to the LP filter

CELP produces digital cellular quality speech at low bit rates The analysis-by-synthesis scheme allows the coder to select the excitation which will sound best to a human listener Unfortunately, the exhaustive codebook searches make the system extremely computationally complex - the CELP encoder operates at roughly 30 MIPS [Rabiner, 1994] Also, the sound quality produced by the system is said to be slightly "reverberant" or muffled This may be because the adaptive codebook can only build up a periodic excitation gradually over several sub-frames

7.2.4 GSM System

GSM is the current pan-European cellular telephony standard The GSM speech coding system operates at a bit rate of 13 kb/s and is a Regular Pulse Excitation - Long Term Prediction - Linear Prediction (RPE-LTP) scheme [ETSI, 1989] The GSM system used in this study is a public domain implementation written in UNIX C

A schematic diagram of the GSM system is presented in Fig 7 4 The coder uses a frame size of 160 samples (20 ms), divided into four sub-frames GSM analysis consists of three main functions - short term Linear Prediction analysis, Long Term Prediction analysis and Regular Pulse Excitation analysis Eighth order Linear Prediction analysis is performed over each pre-emphasised frame using the autocorrelation method The quantised LP coefficients are used to construct an inverse filter which is applied to the pre-emphasised speech signal to give the short-term residual Each sub-frame of the residual is coded using a Long Term Predictor and a Regular Pulse Excitation The delay and gain of the Long Term Predictor is determined by finding the maximum of the cross-correlation between the current residual and the re-synthesised residual from previous sub-frames The quantised delay and gain parameters are used to generate the Long Term Prediction signal which is subtracted from the short-term residual to give the long-term residual This long-term residual is subjected to Regular Pulse Excitation analysis Each sub-frame of the long-term residual is filtered and down-sampled into four interleaved candidate sub-sequences The candidate sub-sequence, or grid position, with the maximum energy is selected for quantisation by Adaptive Pulse Code Modulation The resulting quantised Regular Pulse Excitation is reconstructed and added to the Long Term Predictor output to give the re-synthesised residual for this sub-frame The APCM coded Regular Pulse Excitation, the grid position, the Long Term Prediction lag and the LP coefficients are transmitted to the receiver

The GSM decoder reconstructs the Regular Pulse Excitation sub-sequence from the APCM information and a sub-frame of the long-term residual is created by up-sampling based on the received grid position The output from the Long Term Predictor, a lagged and scaled version of previous excitations, is added to the long-term residual This excitation signal is applied to the decoded LP filter to re-synthesise the speech signal

The GSM system closely models the residual obtained by passing the LP inverse filter over the input speech signal This ensures a close match between the re-synthesised and original waveforms This, in turn, means that the speech quality produced by the GSM system is very high However it must

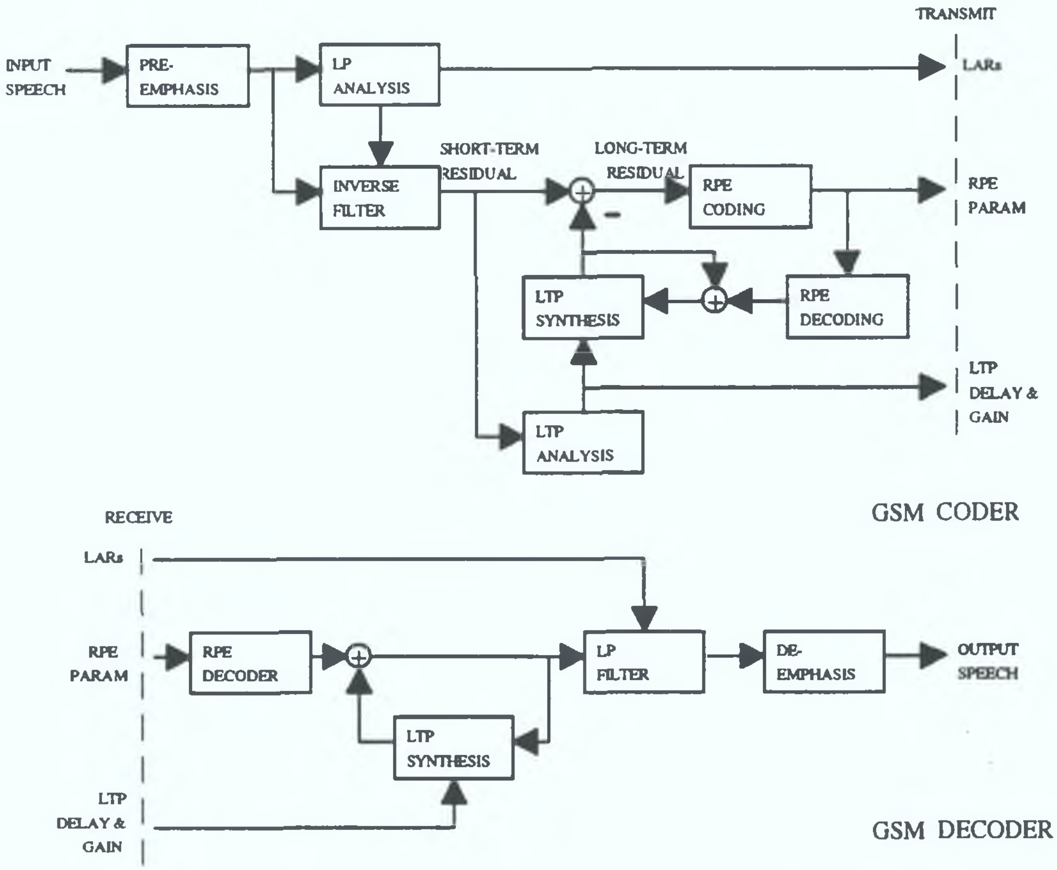


Fig. 7.4. Schematic diagram of the GSM speech coding system [after ETSI, 1989].

be noted that the improvement in quality relative to CELP and LPC-10 is achieved at the expense of a much higher bit rate.

7.2.5 BSD Algorithm

The Bark Spectral Distortion is a perceptually based objective measure for evaluating speech quality. The measure, computed from original and coded versions of an utterance, exhibits a monotonic relationship with the Mean Opinion Score (MOS) and has proven extremely effective in predicting MOS scores for low bit rate coders. As such, the measure provides a consistent and accurate means of objectively evaluating coder performance.

Fig. 7.5 shows a schematic diagram of how the BSD is computed. The original $x(n)$ and coded speech $y(n)$ are separately converted to their Bark spectra $L_x(i)$ and $L_y(i)$. The Bark spectra represent the perceptual auditory characteristics of each signal. The subjective quality of the coded speech is defined as the distance between the Bark spectra.

The process of converting the speech signal to the Bark spectra is designed to model the properties of the human auditory system. Three main factors are considered in the model. Firstly, the human auditory system is known to have poorer tone discrimination at high frequencies than at low frequencies. Secondly, the human ear is not equally sensitive to stimulations at different frequencies. Thirdly, the perceived loudness of a tone is a nonlinear function of the relative acoustic level.

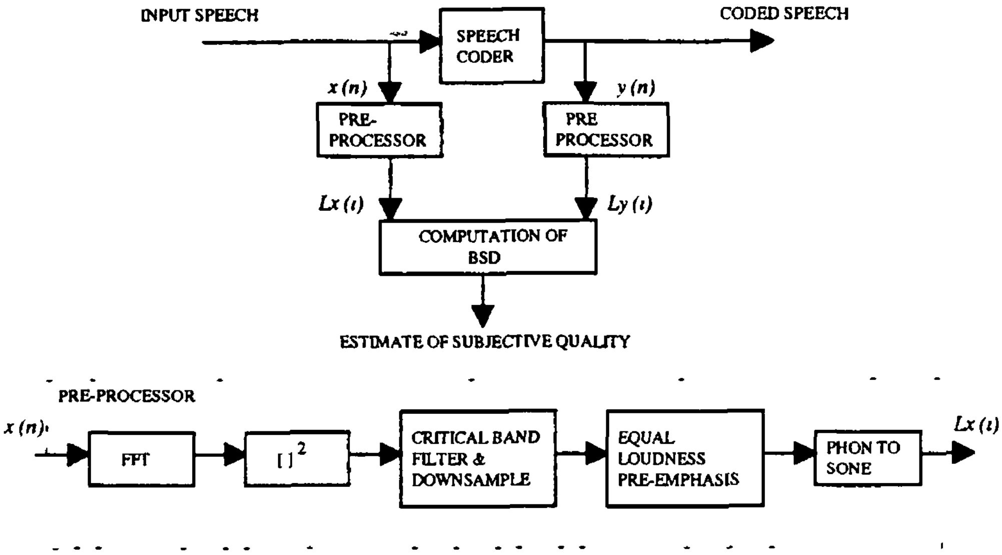


Fig 7 5 Schematic diagram of the procedure for computing the BSD [after Wang et al, 1992]

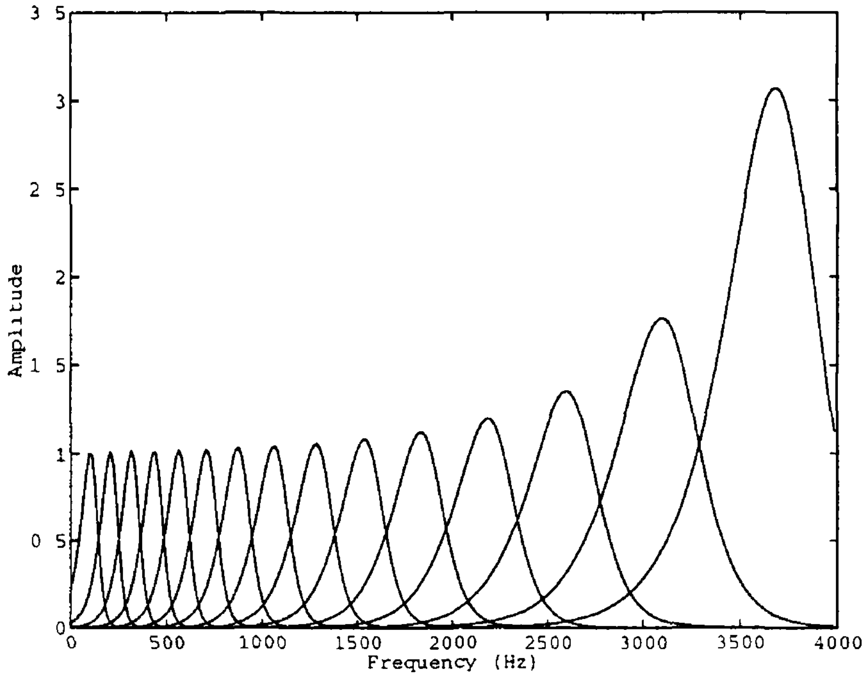


Fig 7 6 Weight functions used for calculating Bark spectrum from power spectrum [after Wang et al, 1992]

To model these perceptual effects, the critical band filtering and frequency dependent sensitivity of the human ear are mapped into the linear frequency domain as a set of fifteen pre-computed weighting functions, Fig 7 6 The conventional power spectrum of the speech signal is converted to the Bark spectrum by applying these weighted averaging functions and a non-linear loudness transformation These operations effectively represent the pre-processing of the human auditory system and allow the low dimensional Bark spectrum to capture the perceptual characteristics of the speech signal The operations are carried out on a frame-by-frame basis using an 80 sample Hamming pre-window with a 50% overlap between consecutive frames

The distortion between the original and coded speech frames is calculated as the Euclidean distance between the original and coded Bark spectra The final BSD score is obtained by calculating the mean distortion and normalising it by the mean Bark energy of the original signal Thus, overall

$$BSD = \frac{\sum_{k=1}^{N}\left[\sum_{i=1}^{15}\left(L_x^k(i) - L_y^k(i)\right)^2\right]}{\sum_{k=1}^{N}\left[\sum_{i=1}^{15}\left(L_x^k(i)\right)^2\right]} \qquad (7\ 1)$$

where N is the number of frames in the speech signal

The BSD measure allows the subjective quality of speech coding systems to be determined by objective means Thus, the speech quality provided by the GELP systems can be reliably assessed without the need for time consuming listener tests Furthermore, the BSD measure is extremely sensitive and allows incremental improvements to be made to a coding scheme Unfortunately, the BSD computation procedure has not been standardised Hence, the scores presented herein are not comparable with those published elsewhere

7.3 GELP QUANTISATION

This section describes the development of the GELP quantisation scheme The vocal tract parameters are encoded via Line Spectral Pairs This is an efficient quantisation scheme and has been used in a number of low rate systems The optimum order of the LP synthesis filter is determined by examining the variation of GELP speech quality with filter order The quantisation levels of the LSPs are selected by examining their distribution during GELP coding of natural speech The pitch quantisation scheme is based on that used in U S Fed Std 1016 CELP

The quantisation of the LF parameters is studied in some detail To provide efficient quantisation, a differential scheme is proposed whereby only the change in the LF parameters relative to the previous period is quantised and transmitted The differential LF parameter distributions are modelled by fitting standard Probability Density Functions to the distributions observed during natural speech Based on these PDF models optimum non-linear quantisation schemes are determined by minimising the induced distortion Quantisers of varying resolution are developed for all of the LF parameters The optimum bit allocation for the LF model is then determined by a knockout procedure Based on this procedure the bit allocation providing the best trade-off between speech quality and bit rate is selected for further study in the next section

The section is divided into three sub-sections. The first sub-section describes the quantisation of the LP coefficients. The second describes the quantisation of the pitch. The third sub-section details the development of the LF parameter quantisers, while the fourth sub-section covers selection of the final bit allocation scheme. The fifth, and final, sub-section presents a summary of the overall GELP quantisation scheme.

7.3.1 Filter parameters

The optimum order of the LP filter was determined by applying unquantised GELP coding to segments of natural speech. Fig. 7.7 shows the variation in BSD with filter order observed when GELP coding was applied to all-voiced noiseless male and female recordings (see Appendix C). Clearly, the best compromise between filter order and speech quality is 8th order. This is consistent with previous work which suggests that two poles are required to model each vocal tract formant and the spectral envelope [Markel and Gray, 1976]. Since the LF model captures the spectral envelope, and since there are normally four formants in the range 0-4 kHz [Fant, 1956], this implies that eight poles are required to accurately reproduce the vocal tract resonances.

The decision was taken to quantise the LP coefficients as Line Spectral Pairs [Sugamura and Itakura, 1981; Crosmer and Barnwell, 1985]. This representation has been shown to lead to efficient quantisation and is used in U.S. Federal Standard CELP. The LSPs are obtained from the z-domain representation of the all-pole transfer function of the LP filter $H(z)$

$$H(z)=\frac{1}{1-\sum_{i=1}^{M}a_i z^{-i}} \tag{7.2}$$

where M is the order of the filter and a_i are the LP coefficients. The inverse filter is given by

$$A(z)=1-\sum_{i=1}^{M}a_i z^{-i} \tag{7.3}$$

$A(z)$ can be decomposed into two (M+1) order polynomials

$$\begin{aligned}P(z)&=A(z)+z^{-(M+1)}A\left(z^{-1}\right)\\ Q(z)&=A(z)-z^{-(M+1)}A\left(z^{-1}\right)\end{aligned} \tag{7.4}$$

so that

$$A(z)=\frac{P(z)+Q(z)}{2} \tag{7.5}$$

It turns out that $P(z)$ has a real zero at z=-1 and $Q(z)$ has a real zero at z=1. All the other zeros of the polynomials are complex and are interleaved on the unit circle. These zeros comprise the LSP parameters. Although the zeros are complex, their magnitudes are known to be unity so that only their frequency or angle is required to specify each one.

Fig. 7.8 shows the distribution of LSPs obtained by applying unquantised GELP coding to the noiseless recordings. The LSP bit allocation scheme for the GELP coder is based on that used in U.S. Federal Standard CELP. Since CELP operates close to the target bit rate for GELP, it is viewed that similar spectral resolution is required in the GELP coder to provide good quality re-synthesis. As in the CELP system, four bits are allocated to the perceptually significant LSPs (two, three and four) while

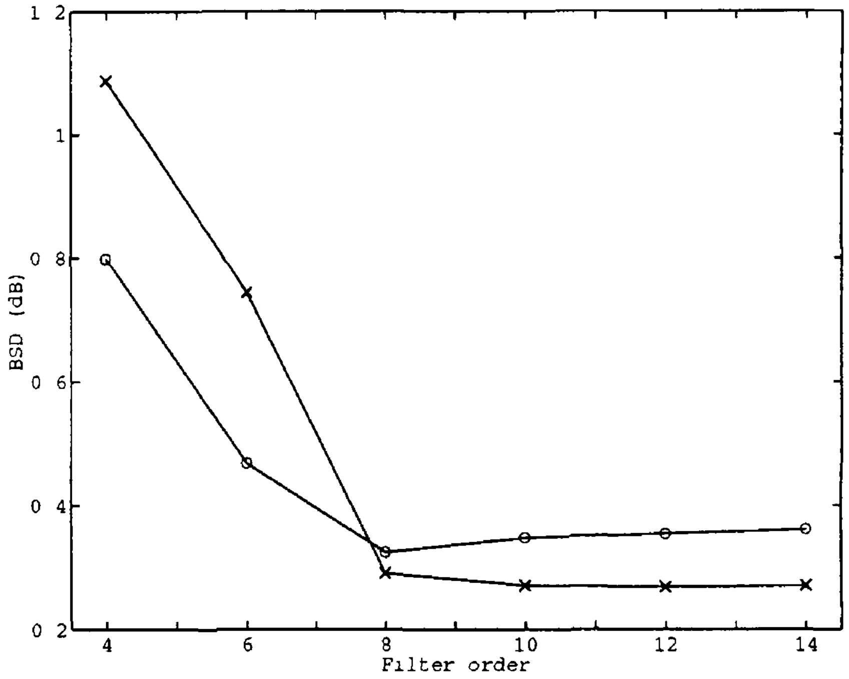


Fig 7 7 Variation of BSD with filter order, unquantised IAIF GELP coder, noiseless recordings x - male subject, o - female subject

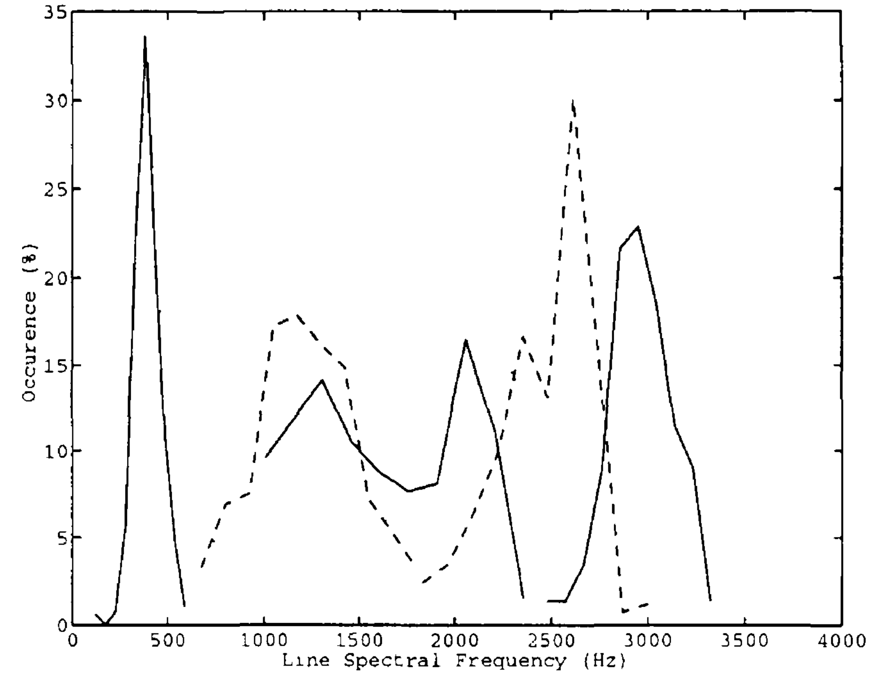


Fig 7 8 Distribution of LSPs unquantised IAIF GELP coder male and female data noiseless recordings from left to right solid - LSP 1 dotted - LSP 2, dashed - LSP 3 solid - LSP 4, dotted - LSP 5, dashed - LSP 6, solid - LSP 7, dotted - LSP 8 three bits are allocated to the remainder The quantisation levels were selected arbitrarily based on the observed LSP distributions and are shown in Table 7 1

LSP	OUTPUT LEVELS (Hz)
1	200 270 325 350 380 440 520 600
2	310 335 365 395 425 460 500 540 580 620 660 710 770 840 910 980
3	720 760 800 840 885 940 1005 1075 1150 1250 1350 1450 1550 1660 1750 1850
4	1000 1050 1130 1210 1285 1350 1430 1510 1590 1670 1750 1850 1950 2050 2150 2250
5	1670 1770 1890 2030 2200 2400 2600 2800
6	1975 2150 2275 2400 2550 2700 2900 3100
7	2610 2730 2850 2950 3050 3160 3280 3400
8	3190 3270 3350 3420 3490 3590 3710 3830

Table 7 1 Output quantisation levels for LSPs

7.3.2 Pitch Parameter

The fundamental voice source parameter is the pitch period Accurate reproduction of pitch is essential for high quality synthesis Therefore, the pitch period is transmitted for every glottal cycle and is quantised to a resolution of 8 bits, according to the scheme defined in U S Federal Standard CELP, see Table 7 2 This scheme provides fractional resolution at certain frequencies This has been shown to reduce the roughness perceived in the reproduction of voiced speech Moreover, since the LF analysis parameters are derived from the fundamental frequency, the accuracy of the re-synthesised excitation is dependent on the resolution of the pitch Thus, the pitch is quantised to a high resolution and is transmitted independently for each glottal cycle

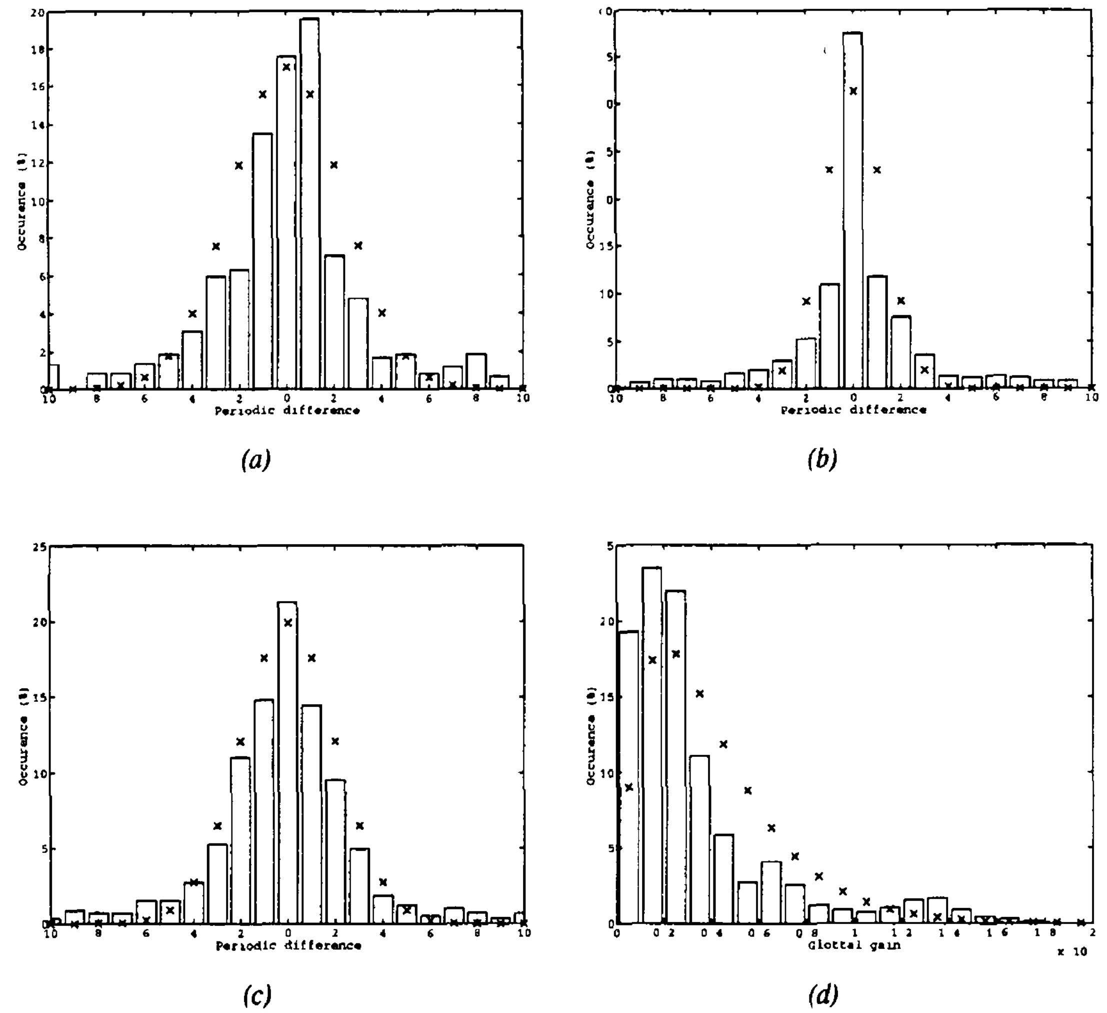
PITCH RANGE (in samples)	RESOLUTION (in samples)
20 - 25 2/3	1/3
26 - 33 3/4	1/4
34 - 79 2/3	1/3
80 - 147	1

Table 7 2 Pitch quantisation scheme [after Fenichel 1991, 1992]

7 3 3 LF Parameters

The LF analysis parameters were deemed best for transmission since they are most closely related to the spectral characteristics of the glottal excitation [Fant and Lin, 1988] Furthermore, the analysis parameters are ratios of the pitch period and so easily adapt to a changing fundamental frequency The LF analysis parameters are as follows - the glottal gain E_e, the glottal frequency r_g , the dynamic leakage r_a and the open quotient o_q (definitions are supplied in Section 6 2 3)

The distributions of the LF parameters extracted from two sentences of noiseless voiced speech, one male and one female, are shown in Figs 6 24 and 6 25 Evidently, the variance of the parameters is quite large On examining the parameter contours, Figs 6 22 and 6 23 it was noted that there is a high degree of correlation between successive glottal parameters In order to remove this redundancy, it was



(a) (b) (c) (d)

Fig 7 9 Distribution of differential LF parameters (solid) and PDF fit (x), male and female data, noiseless recordings (a) glottal frequency, r_g, *(b) dynamic leakage,* r_a, *(c) open quotient,* o_q, *(d) glottal gain,* E_e

decided that the LF parameters be quantised differentially That is, the differences between the current glottal parameters and the previous are quantised and transmitted Fig 7 9 shows the distribution of the differential LF parameters calculated over the noiseless recordings Clearly, the variance of the period-by-period difference in the parameters is significantly less than that of the parameters themselves For the purposes of coding, the default values for the LF parameters were taken as r_g = 75%, r_a = 2%, and o_q = 92 5% At voicing onsets the differential parameters are calculated relative to these defaults

Optimum independent quantisation of the differential parameters can be achieved by finding the non-linear quantiser which provides minimum distortion [Max, 1960, Jayant, 1974] The distortion D generated by a L level quantisation scheme is given by

$$D=\sum_{k=1}^{L}\int_{x_{k-1}}^{x_k} h(y_k-\xi)f(\xi)d\xi \qquad (7\ 6)$$

where x_k is the end-point of level k ($x_0=-\infty$ and $x_L=\infty$), y_k is the output value of level k, $f(x)$ is the PDF of the parameter to be quantised and $h(x)$ is the distortion measure The necessary conditions for minimum distortion are obtained by differentiating D with respect to $\{x_k\}$ and $\{y_k\}$ The result of this differentiation is the pair of equations

$$h(y_k - x_k) = h(y_{k+1} - x_k) \qquad k = 1,2,\quad ,L-1$$

$$\int_{x_{k-1}}^{x_k} h'(y_k - \xi) f(\xi) d\xi = 0 \qquad k = 1,2,\quad ,L \tag{7 7}$$

In the special case of a squared error distortion measure

$$h(x) = x^2 \tag{7 8}$$

and so

$$x_k = \frac{y_k + y_{k+1}}{2} \qquad k = 1,2,\quad ,L-1$$

$$y_k = \int_{x_{k-1}}^{x_k} \xi f(\xi) d\xi \qquad k = 1,2,\quad ,L \tag{7 9}$$

Thus y_k is the centroid (mean) of $f(x)$ between x_{k-1} and x_k These equations may be solved numerically for any given $f(x)$

In order that a quantisation scheme be generated for the LF parameters, a PDF suitable for integration must be defined Examining the distributions of the differential LF analysis parameters it can be seen that, with the exception of the glottal gain, they approximate to the Normal PDF given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \tag{7 10}$$

where σ and μ are the standard deviation and mean of the distribution, respectively The Normal PDF was matched to the observed differential LF distributions based on a minimum squared error criterion The mean was set to zero and the standard deviation of the PDF was numerically optimised by a gradients method to minimise the mean squared error between the distributions The error was calculated between the matched and measured PDFs at sample points 2% apart from -40% to +40% The Normal PDFs fitted to the measured distributions are shown in Fig 7 9 Table 7 3 lists the optimum standard deviations values determined by the least mean squared error procedure

LF PARAMETER	STD DEVIATION	MEAN SQUARED ERROR
rg	2 34	3 3e-4
ra	1 24	1 0e-3
oq	1 99	1 6e-4

***Table 7 3** Standard deviation of Normal distributions giving the best fit to the differential LF distributions*

Based on the matched PDFs, optimum non-linear quantisers were generated numerically via Eqs (7 6), (7 9) and (7 10) To ensure that the quantisation schemes could parameterise smooth LF parameter contours, it was necessary to provide a zero quantisation level for all of the differential LF parameters Obviously, this produces an odd number of quantisation levels which is unusual for such a scheme Nevertheless, the decision was taken to use 2^n+1 output levels in each of the quantisers under test The optimum quantisers are shown in Table 7 4

NO OF LEVELS	*rg*	*ra*	*oq*
3	2 85	1 52	2 43
5	1 79 4 03	0 95 2 14	1 52 3 42
9	1 04 2 15 3 45 5 27	0 55 1 14 1 83 2 80	0 88 1 82 2 93 4 48
17	0 57 1 15 1 75 2 40 3 11 3 93 4 97 6 49	0 30 0 61 0 93 1 27 1 65 2 09 2 64 3 45	0 48 0 97 1 49 2 04 2 64 3 34 4 22 5 52

Table 7 4 Output levels of the optimum non-linear quantisers for the differential LF parameters Since the distributions are symmetric only the levels greater than zero are shown.

The quantisation scheme for the glottal gain parameter E_e was developed in a similar fashion Fig 7 9(d) shows the distribution obtained for the glottal gain over the noiseless recordings The Gamma PDF was found to most closely approximate the distribution

$$f(x)=\begin{cases}(x/\beta)^{\alpha-1}e^{-(x/\beta)}/\Gamma(\alpha) & x\geq 0\\ 0 & x<0\end{cases} \qquad (7\ 11)$$

where α controls the skew of the distribution and β governs the scale

$$\Gamma(\alpha)=\int_0^{\infty} x^{\alpha-1}e^{-y}dy \qquad (7\ 12)$$

As for the other LF parameters, the PDF was fitted to the observed distribution by numerical minimisation of the mean squared error, Eqs (7 6), (7 9) and (7 11) Following the minimisation process, the near-optimal PDF parameter settings, $\alpha=2$ and $\beta=2000$, were chosen to reduce the computational burden of generating the quantisation scheme The PDF fit obtained using these parameters is shown in Fig 7 9(d) Optimum four, five and six bit non-linear quantisers were obtained by numerical minimisation of the distortion Table 7 5 shows the derived quantisation schemes

It must be noted that in quantising the glottal gain parameter, only positive levels have been considered For re-synthesis purposes, it is irrelevant whether the parameter is positive or negative Thus the polarity of the glottal gain need not be transmitted By convention, the synthesiser produces negative glottal pulses at closure In the encoder, the polarity of the glottal pulse is important for the purposes of LF fitting However, it is constant for a given recording configuration and a simple positive/negative adaption scheme can be implemented at connection time

NO OF BITS	LEVEL OUTPUTS
4	613 1306 1956 2604 3269 3964 4701 5494 6360 7321 8409 9672 11190 115753 20123
5	368 763 1118 1457 1789 2119 2449 2783 3122 3469 3824 4189 4566 4956 5362 5785 6229 6698 7193 7719 8283 8889 9547 10269 11069 11966 12992 14194 15647 17492 20041 24221
6	218 446 647 835 1014 1188 1359 1528 1695 1861 2027 2194 2360 2527 2696 2866 3037 3210 3384 3562 3741 3923 4108 4295 4485 4679 4877 5078 5284 5494 5710 5930 6155 6387 6624 6868 7119 7379 7646 7923 8209 8507 8816 9138 9475 9827 10196 10584 10994 11427 11888 12382 12911 13483 14106 14789 15548 16401 17376 18517 19894 21626 24789 15548 16401 17376 18517 19894 21626 23982 27706

***Table 7 5** Output levels of the optimum non-linear quantiser for the glottal gain, E_e (assuming a 16 bit quantiser)*

7.3.4 LF Bit Allocation

Selection of the best bit allocation scheme for the LF parameters necessitates a compromise between speech quality and bit rate In order to determine the optimum bit allocation scheme, a knockout strategy was employed In each round of the knockout, GELP coding was performed on the noiseless male and female recordings using four candidate bit allocation schemes These four candidate schemes were formed by adding one extra bit to the scheme which won the previous round The candidates differed in which parameter the extra bit was allocated to - either E_e, r_g, r_a or o_q The performance of the four candidate schemes was determined by comparing the BSD of the re-synthesised speech segments The candidate giving the lowest BSD, averaged over the male and female scores, was chosen as the winner and proceeded to the next round The knockout started with four bits allocated to the glottal gain and zero bits allocated to the LF parameters Thus, in the first round the LF parameters were fixed For the purposes of the knockout procedure the zero quantisation levels used for the LF analysis parameters are ignored, for example, replacing the fixed dynamic leakage with a three level quantiser is treated as being equivalent to adding one bit to the glottal gain quantiser To the author's knowledge the knockout strategy for determining bit allocation is unique to this investigation

The results of the knockout procedure are shown in Figs 7 10 and Table 7 6 Fig 7 10 shows the variation of mean BSD with the knockout round and Table 7 6 shows the winning bit allocation in each round Also, Fig 7 11 (a) and (b) shows the mean periodic SNR of the LF signal, relative to the inverse filtered speech, and the mean periodic SNR of the re-synthesised speech, relative to the original recording, respectively

ROUND	1	2	3	4	5	6
WINNING PARAMETER	*Ee*	*ra*	*rg*	*Ee*	*rg*	*Ee*
MEAN BSD CHANGE	-0 067	0 014	-0 035	-0 012	-0 003	-0 002

***Table 7 6** Winners of knockout rounds showing change in mean BSD obtained since the previous round*

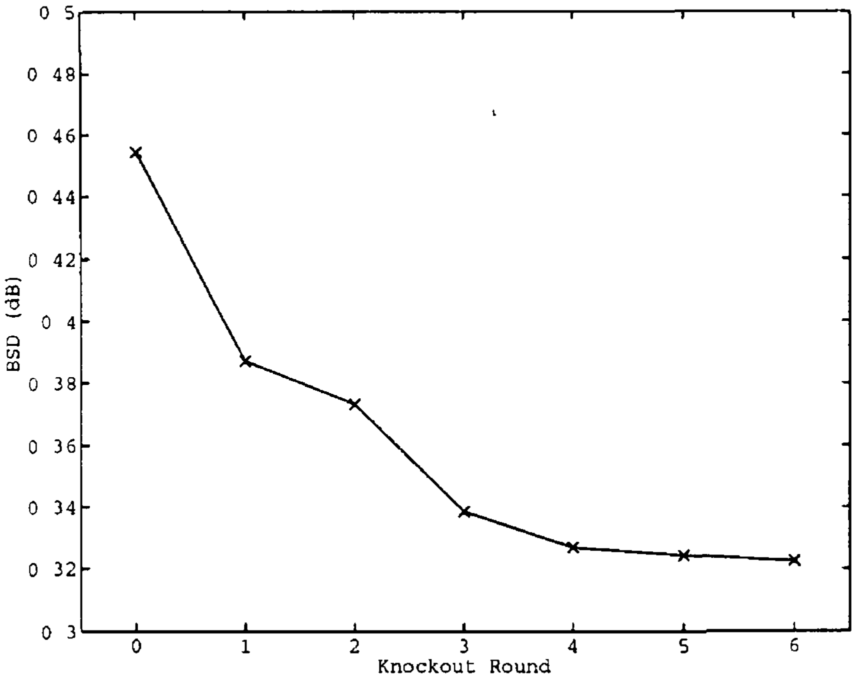
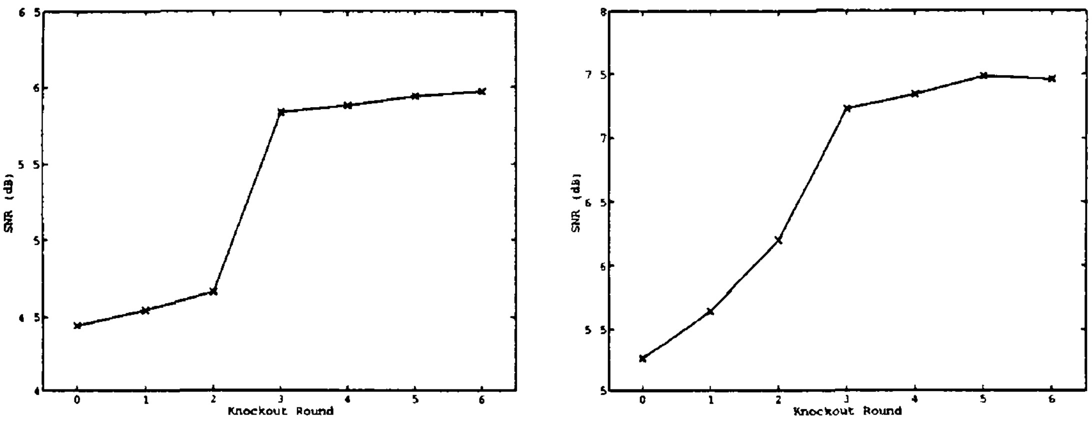


Fig 7 10 Variation of BSD with knockout round, quantised IAIF GELP coder, averaged over male and female data, noiseless recordings



(a) *(b)*

Fig 7 11 Variation of mean periodic SNR with knockout round quantised IAIF GELP coder, averaged over male and female data noiseless recordings (a) LF SNR, (b) re-synthesised SNR

Clearly, increasing the number of bits allocated to the source parameters improves the speech quality as measured by the BSD. Of note is the importance of the dynamic leakage r_a and glottal frequency r_g parameters for high quality re-synthesis. In fact, the knockout procedure shows that three level quantisation of either the dynamic leakage or the glottal frequency is more important than increasing the resolution of gain quantisation from 32 levels to 64 levels. The dynamic leakage and glottal frequency parameters are known to control the shape of the glottal magnitude spectrum [Fant and Lin, 1988]. In particular, the dynamic leakage controls the spectral roll-off and the glottal frequency determines the level of the lower harmonics. Thus, they have a direct impact on the speech spectrum and hence control the perceived quality of the re-synthesised speech. In contrast, the open quotient o_q has a minor influence on speech quality. As was mentioned earlier, the open quotient governs the glottal opening instant. At opening, the glottal waveform is relatively smooth and, therefore, changes in the open quotient parameter have little effect on the glottal waveshape or the source spectrum.

The improvements in BSD and mean periodic SNR per additional bit decrease substantially after the fourth round. The efficiency of the glottal coding appears to saturate beyond this point. A greater variety of subjects or speech material might increase the number of bits required for accurate LF coding. However, given that the differential scheme allows for a degree of adaption to the speaker characteristics, it seems reasonable to assume that these findings are typical for most speech materials. For this reason, the LF quantisation scheme of round four seems most appropriate for use in the GELP coding system. In order that an integer number of bits is allocated to each glottal cycle, the round five scheme, shown in Table 7.7, was actually implemented in the coder.

PARAMETER	NO. OF LEVELS
Ee	64
rg	5
ra	3
oq	0

Table 7.7. Chosen ten bit LF parameter quantisation scheme.

7.3.5 Overall Bit Rate

The overall bit allocation is shown in Table 7.8. Since the LF parameters are transmitted for each pitch period, the overall bit rate of the system is dependent on the fundamental frequency of the incoming speech. The variation of bit rate with fundamental frequency is shown in Fig. 7.12. The use of a variable rate system simplifies the implementation of the GELP system but makes it unsuitable for certain applications. It is the author's view that the quantisation scheme present herein could be modified for fixed rate transmission without undue difficulty, and with little or no loss in speech quality.

PARAMETERS	REFRESH RATE	NO. OF BITS
LSP	every frame	27
Voicing Decision	every frame	1
Pitch	every pitch period	8
LF	every pitch period	10

Table 7.8. Bit allocation scheme of GELP coder.

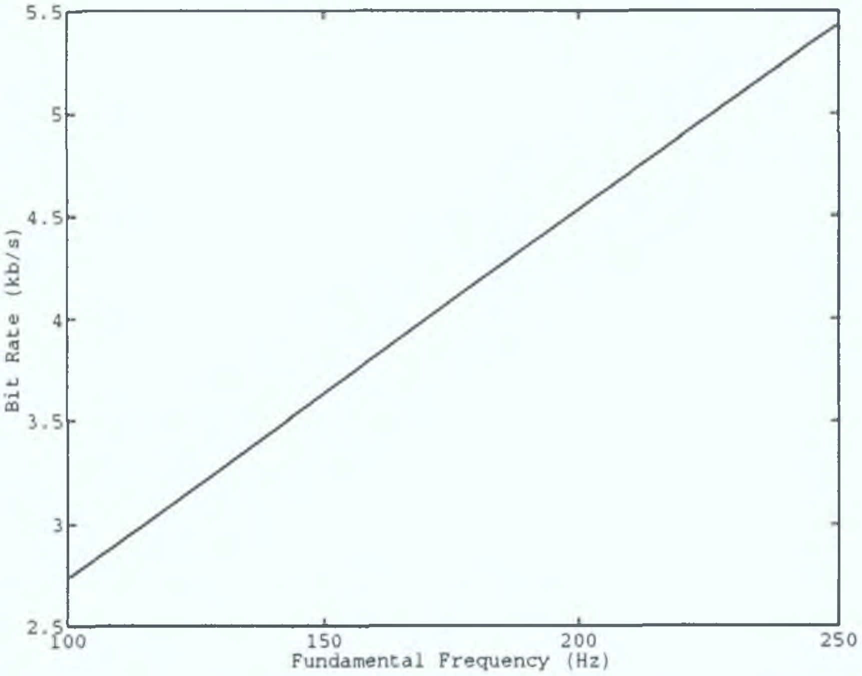


Fig. 7.12. Variation of mean bit rate with fundamental frequency.

7.4 PERFORMANCE STUDY

This section describes experiments carried out to determine the performance and robustness of the GELP system compared to the three conventional coders.

The speech quality provided by the system was assessed by calculating the BSD introduced by each of the schemes when applied to four all-voiced recordings. These consisted of a male and a female subject recorded under noiseless conditions and a male and a female subject recorded in a typical office environment. The robustness of the GELP algorithms was investigated by testing the coders on the noisy and reverberant recordings. These were generated by adding white noise and artificial reverberation to the noiseless male and female recordings.

It should be noted that the BSD speech quality scores which will be presented for the GELP system may be overly generous to all but the office recordings for two reasons. Firstly, the LF parameter distributions obtained from the noiseless recordings were employed in the design of the quantisation scheme, Section 7.3.3. Secondly, manual identification of the GCI was used in GELP processing of the noiseless, noisy and reverberant recordings. The impact of these factors on the BSD measured for the GELP systems is thought to be minimal. Firstly, since a differential coding scheme is employed, the GELP coder should quickly adapt to the voice source characteristics of any speaker. Secondly, for the distortion levels under investigation the PMLED technique has already displayed good performance, see Chapter 5, therefore the results obtained using the automatic system are likely to be little different from those obtained using manual marking. Thirdly, the BSD scores obtained for the office recordings are entirely consistent and compatible with the results based on the noiseless recordings. The office recordings (which were not employed in the design of the LF quantisation scheme) were GELP coded using automatic GCI detection. Overall, it is viewed that the experiments using the noiseless, noisy and reverberant speech provide reliable information on the quality of GELP coding. Furthermore, the use of a large data set is crucial for reaching valid conclusions regarding the performance of the GELP system.

The section is divided into two sub-sections. The first sub-section describes the performance of the coding systems in processing the noiseless and office recordings. The second sub-section considers the robustness of the system in processing the noisy and reverberant test data. Both sub-sections include a full discussion and assessment of the results. Further information on the capture of the speech data used in this section is provided in Appendix C.

7.4.1 Speech Quality

The speech quality provided by the coding systems is shown in Fig. 7.13 while Table 7.9 shows the mean pitch of the segments plus the mean and peak bit rates for the GELP coders.

	NOISELESS RECORDINGS		OFFICE RECORDINGS	
	MALE	FEMALE	MALE	FEMALE
MEAN PITCH (Hz)	89	160	113	194
MEAN BIT RATE (kb/s)	2.52	3.75	2.96	4.39
PEAK BIT RATE (kb/s)	3.33	4.53	3.93	5.73

Table 7.9. Mean pitch of speech material plus mean and peak bit rates for GELP coders.

A number of factors influence the speech quality achieved by the coding systems. Examining the results for the LPC-10 coder and relating them to the pitch of the processed speech, it can be clearly seen that the distortion introduced by LPC-10 coding reduces with increasing pitch. The standard coders examined in this study are designed for application in telecommunications systems. As such, they need only process speech in the telephone bandwidth of 300-3600 Hz. Thus, LPC-10 does not preserve the low frequency information in the speech recordings. This introduces significant distortion in the cases of the low pitched male speech but less for the female subjects.

The CELP system also removes low frequency energy from the speech signal. As for LPC-10, this causes an improvement in BSD with increasing pitch, for example, compare the male noiseless recording and the male office recording. The removal of low frequency energy in CELP is caused by a 175 Hz highpass post-processing filter. Due to their high pitch, the filter has little effect on the two female recordings. What is significant in this case, is that the office recording is the noisier of the two. Background noise is hard to account for in a low rate coding scheme, hence the BSD for CELP increases when moving between the female noiseless and office recordings.

In contrast to the lower rate systems, GSM has only a slight highpass filtering effect and, due to its greater transmission rate, is more robust to distortions of the incoming speech signal. As a result, the BSD observed for GSM is roughly constant across all of the speech material.

The two GELP systems accurately preserve the low frequency components in the speech signal. Thus, pitch has less immediate impact on their BSD scores. In all but one case, the IAIF based system

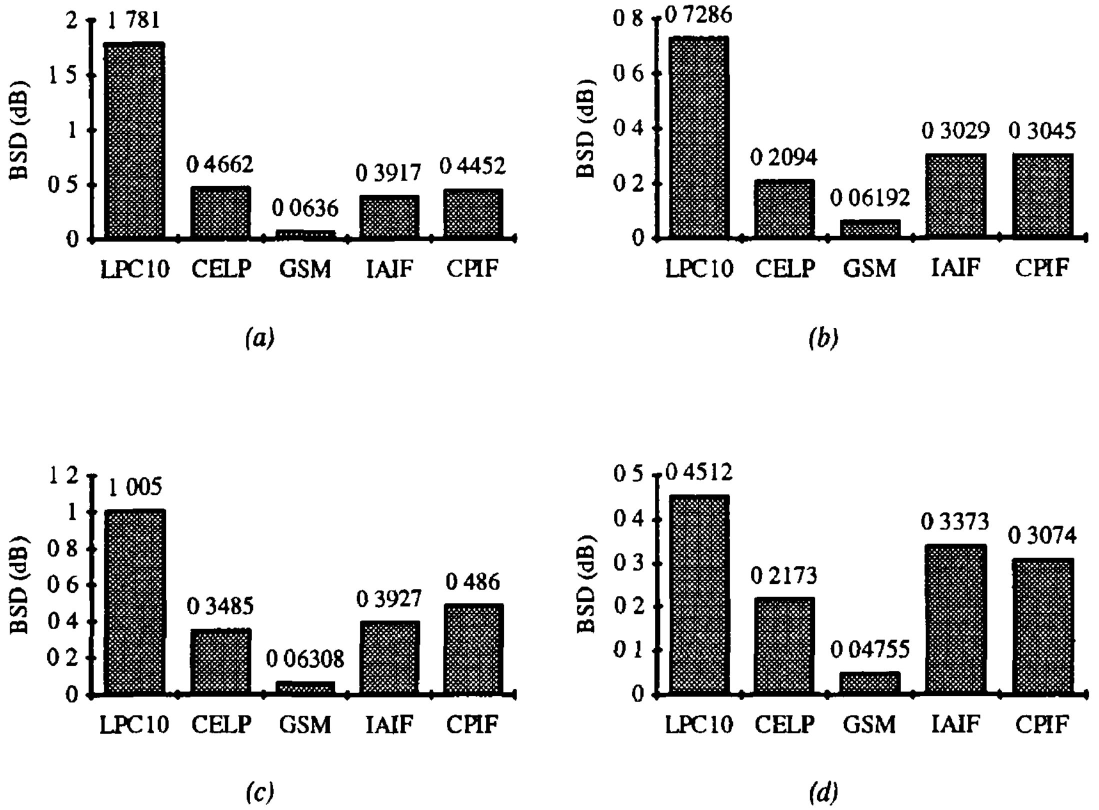
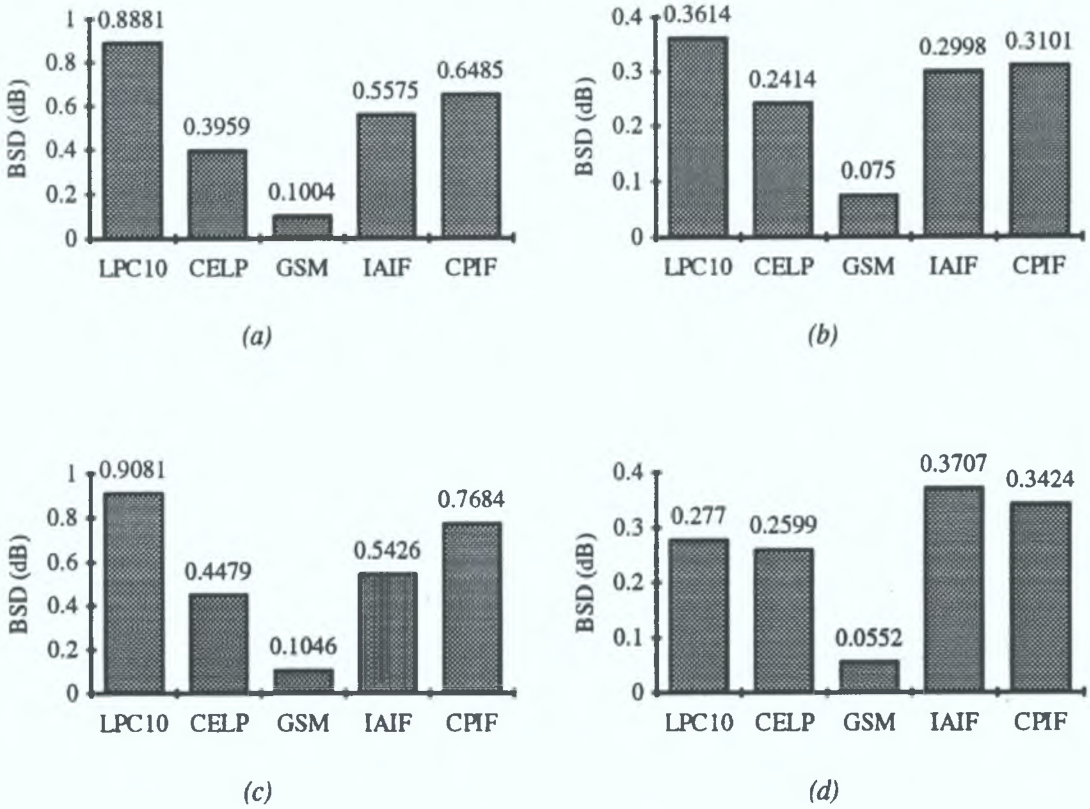


Fig 7 13 BSD of coded speech (a) male speech, noiseless recording, (b) female speech, noiseless recording, (c) male speech, office recording, (d) female speech, office recording

outperforms the CPIF system This appears to be due to the robustness of the IAIF procedure In the female noiseless recording there is an absence of input distortion, hence both techniques perform equally well In contrast, noise is present in the two male recordings - breathy noise, in the case of the noiseless recording, and ambient noise, in the case of the office recording As a result, the more robust IAIF based coder outperforms the CPIF based system for both of these recordings The only test data for which CPIF provides better speech quality than IAIF is the female office recording In this case it appears that IAIF has problems separating the glottal source and vocal tract filter in the spectral domain, due to the high pitch of the voice This effect is evident for high pitched voices, particularly during phonemes with low first formants, such as the vowel [i] Although CPIF outperforms IAIF in this case, it must be remembered that for voices of still higher pitch CPIF fails completely because the closed phase becomes too short for accurate LP analysis

Comparing IAIF based GELP with LPC-10, it can be seen that GELP produces better BSD scores for all of the speech material This performance advantage is because of the more accurate low frequency modelling in GELP due to the inclusion of a parametric glottal excitation Overall, the BSD scores obtained for GELP increase relative to LPC-10 as pitch and noise increase The advantages of glottal modelling become less significant as fundamental frequency and distortion rise For the low pitched



(a) *(b)*

(c) *(d)*

Fig. 7.14. BSD of band-limited coded speech: (a) male speech, noiseless recording; (b) female speech, noiseless recording; (c) male speech, office recording; (d) female speech, office recording.

male speech, GELP achieves a mean transmission rate comparable to that of LPC-10, while for the female speech material, the mean bit rate rises to 1.5 or 2 times that of LPC-10.

GELP only succeeds in providing better quality than CELP for the lowest pitch male subject under noiseless conditions. In all of the other tests, CELP gives adequate low frequency performance and is more robust to distortions of the incoming speech. However, the quality advantage of CELP is paid for at the cost of a higher mean bit rate. In particular, for the male noiseless data the mean GELP bit rate is almost half that of CELP.

In order to determine the advantage bestowed on the GELP systems by measuring the BSD over frequencies not represented by the standard systems, the BSD scores were re-calculated over a bandwidth typical of telephone systems. Before determining the BSD, two Chebychev filters were applied, forwards and backwards, to the input and output speech data. The filters were a tenth order 300 Hz highpass filter and an eighth order 3600 Hz lowpass filter. Both were specified to incur 60dB attenuation in the stop-band and, at most, a 0.5 dB ripple in the passband. The resulting BSD scores are shown in Fig. 7.14. Note that the bandpass filtering operation was not applied to the speech input to the coding systems. For correct GELP processing, the incoming speech material must preserve the low frequency information.

Removal of low frequency energy results in a significant improvement in the BSD scores obtained by LPC-10 Since CELP incorporates more moderate highpass filtering, its band-limited BSD scores only show an improvement for the lowest pitched male subject. Also, for the GSM coder, which incorporates a good low frequency model, band-limitation leads to higher BSD figures in all cases For all of the coders, the lowest BSD results are obtained for the female speech This effect has already been commented on by Wang et al and it prevents comparisons between the BSD scores obtained for the male and female subjects

In all but one case, the BSD values attained by the GELP systems are greater than those obtained before band-limitation This clearly points to the accurate low frequency modelling of the GELP approach As before, IAIF based GELP outperforms the CPIF based system, except for the highest pitched test data

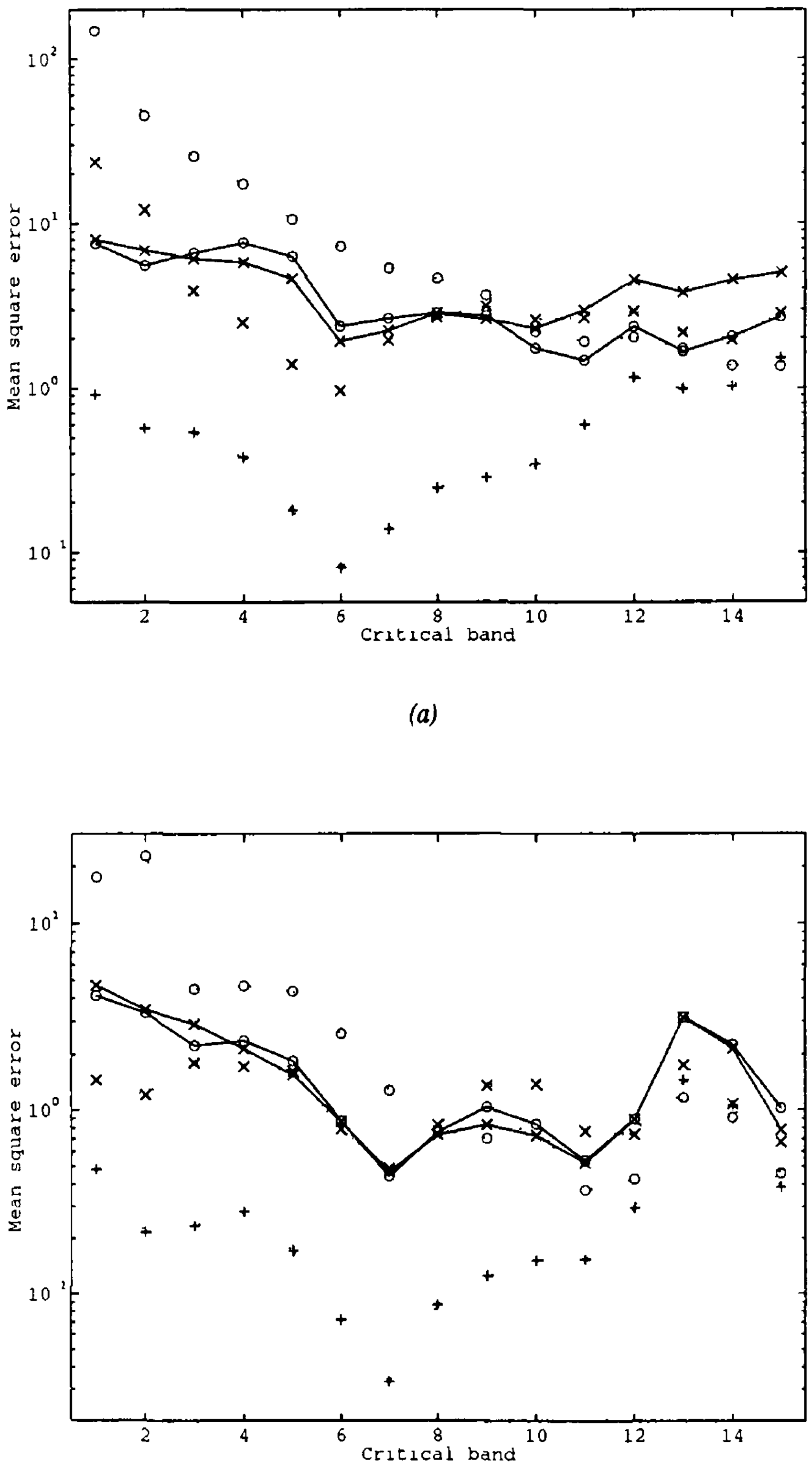
Overall, the application of the telephone bandwidth limitation to the calculation of the BSD scores leads to a deterioration in the performance of GELP relative to the standard systems However, the perceptual quality achieved by GELP still exceeds that of LPC-10, except for the highest pitch female subject. With the inclusion of band-limitation, CELP outperforms GELP in all of the test segments Under these conditions, the only advantage GELP offers over CELP is a reduced bit rate

So that the performance of the coders can be analysed more fully a new quality measure, the mean square critical band distortion, is proposed The critical band distortion is calculated as the difference in Bark spectra calculated for the original and coded versions of the recordings As in the BSD procedure, the Bark spectra are determined by applying a frequency domain weighting function and a loudness transformation to the short-term magnitude spectra of the signals The mean square critical band distortion is calculated as the squared distortion observed in each critical band averaged over the entire speech recording The mean square critical band distortion is thus a measure of the frequency-dependant perceptual distortion introduced by the coding process Fig 7 15 shows the mean square critical band distortions calculated over the test recordings and Table 7 10 shows the centre frequencies of the critical bands

BAND NUMBER	1	2	3	4	5	6	7	8
CENTRE FREQUENCY (Hz)	101	204	313	430	560	705	870	1059
BAND NUMBER	9	10	11	12	13	14	15	
CENTRE FREQUENCY (Hz)	1278	1532	1828	2176	2584	3064	3630	

***Table 7 10** Centre frequencies of critical band weighting functions*

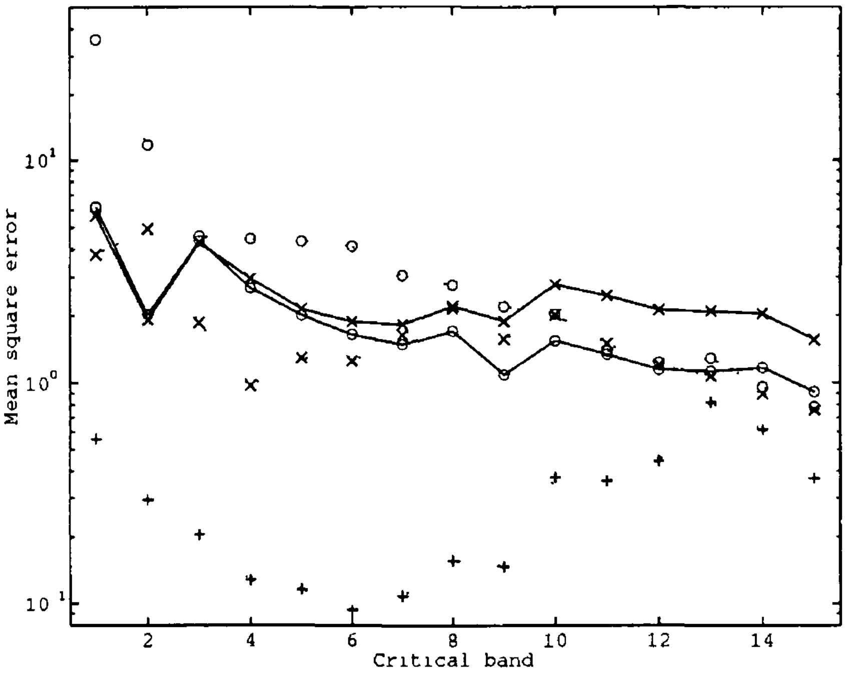
The graphs clearly show the improved speech quality of GELP relative to LPC-10 in the low frequency range For all but the female office recording, the perceptual distortion introduced by GELP is less than that induced by LPC-10 in all of the critical bands up to 8 or 9, that is up to roughly 1 kHz CELP displays speech quality similar to GELP across most of the critical bands However, it presents some obvious advantages in the range 3-7, that is 250-900 Hz Unsurprisingly, GSM is substantially better than any other scheme over most of the frequency range However, the perceptual quality of all of the systems is very similar in the top critical bands, approximately 3000 - 4000 Hz Reproducing data at



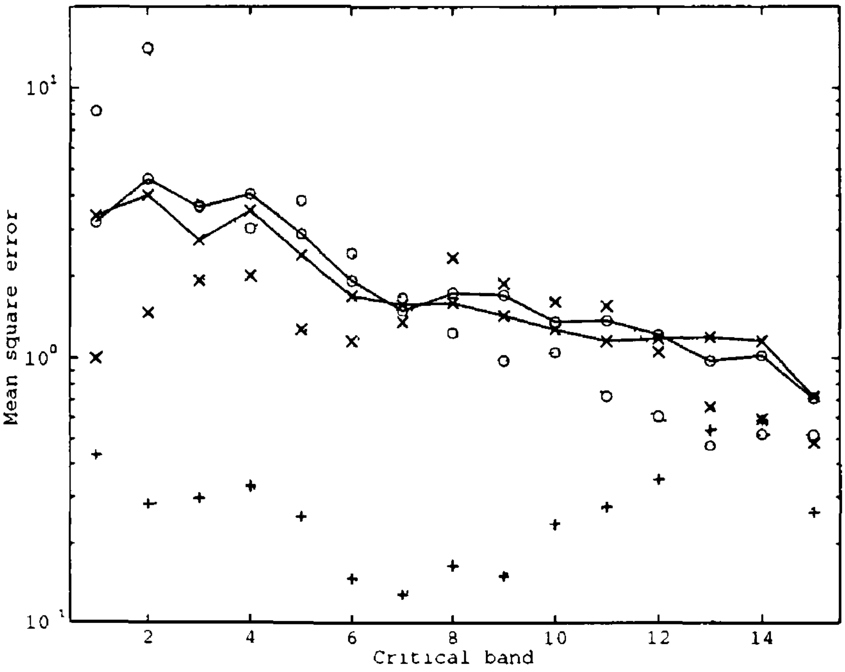
(a)

(b)

Fig 7 15 Variation of mean square Bark spectral error with critical band (a) male subject noiseless recordings (b) female subject noiseless recording (c) male subject office recording, (d) female subject office recording dotted o - LPC-10 dotted x - CELP, dotted + - GSM, solid o - IAIF GELP, solid x - CPIF GELP



(c)



(d)

Fig 7 15 (continued)

these frequencies generally requires a high transmission rate Additionally, there may be some distortion incurred by the presence of unmodelled high order formants Consequently, the spectral distortion incurred by low rate coders tends to be significant at high frequencies

In most of the critical bands, the performance of the two GELP systems is very similar Perhaps, the most noticeable difference is in bands 10-15 (1400-4000 Hz) of the male recordings for which IAIF based coding clearly outperforms the CPIF based system It may be that LP analysis over the closed phase sometimes includes dynamic leakage effects and this prevents accurate determination of the r_a parameter This would, in turn, alter the spectral balance of the re-synthesised speech and cause distortion in the medium to high frequency region

The BSD scores presented concur with the findings of limited informal listening tests

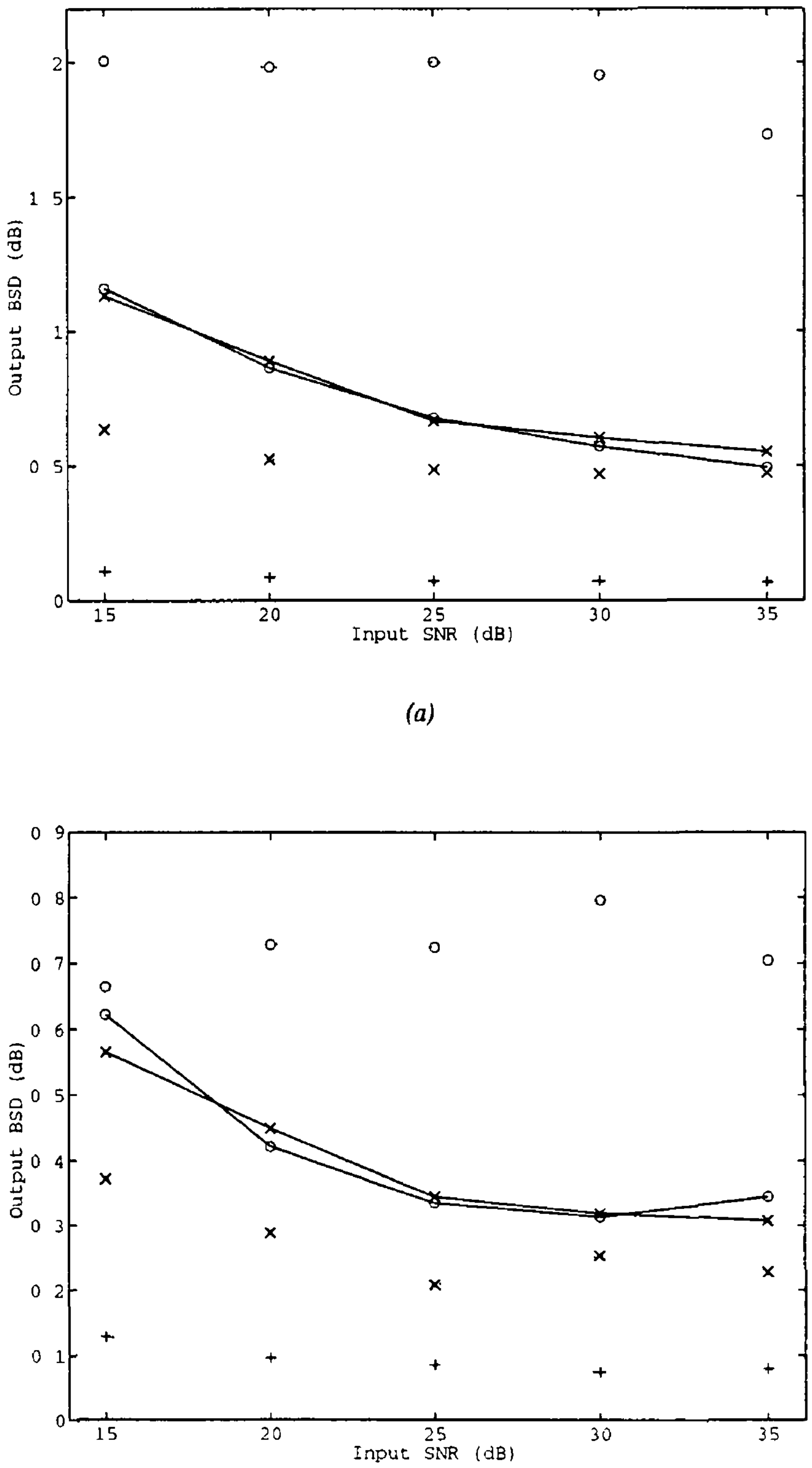
7 4 2 Robustness

The speech quality obtained when coding the noisy and reverberant recordings was also assessed using the BSD measure The results for the five coding systems are shown in Figs 7 16 and 7 17 The bit rates achieved by the GELP coders for these segments are the same as those obtained during the noiseless recordings, see Table 7 9 Note that no band-limitation is performed

Clearly, the GELP coders are less robust than the conventional systems At low distortion levels the speech quality achieved by both GELP systems is comparable to that of CELP, particularly for the male subjects However, as the degree of input distortion increases, the speech quality provided by the GELP coders decreases more rapidly than that of any of the standard systems In the case of the female recordings, the speech quality of the GELP systems approaches that of LPC-10 at the highest distortion levels tested For the male speech, the GELP technique provides significantly better quality than LPC-10 at all distortion levels

Under conditions of significant distortion to the input speech, the choice of inverse filtering procedure makes little difference to the subjective quality of the GELP coded speech In all of the experiments, except perhaps for the female subject at the maximum source-receiver distance, both GELP systems produce similar re-synthesis quality This finding is surprising, given that IAIF has been found to provide more accurate glottal estimation under conditions of noise (see Chapter 6) In the case of speech coding, the coarse quantisation of the LF parameters probably prevents the greater robustness of IAIF from impacting the overall perceptual quality of the re-synthesised speech

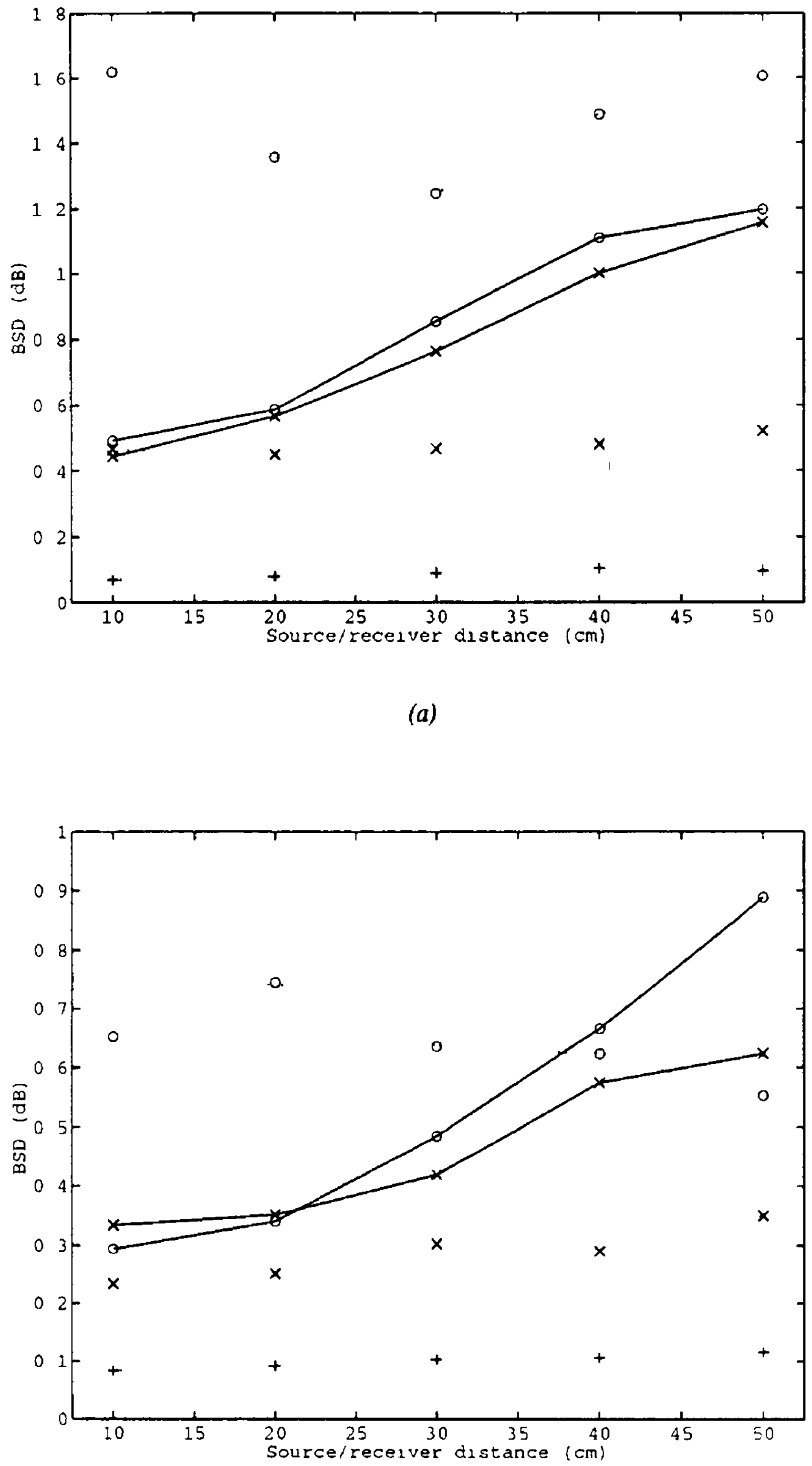
Although the GELP systems are not as robust as the conventional systems, the performance degradation of GELP with increasing distortion is still quite gradual For example, for the male subjects, the quality of GELP still outstrips LPC-10 when the microphone is placed 0 5 m from the subject



(a)

(b)

Fig 7 16 *Variation of BSD with input noise (a) male subject (b) female subject dotted o - LPC-10, dotted x - CELP, dotted + - GSM, solid o - IAIF GELP, solid x - CPIF GELP*



(a)

(b)

***Fig 7 17** Variation of BSD with source-receiver distance (a) male subject, (b) female subject dotted o - LPC-10, dotted x - CELP, dotted + - GSM, solid o - IAIF GELP, solid x - CPIF GELP*

7 5 CONCLUSION

This chapter has described an investigation into Glottal Excited Linear Prediction coding A low bit rate, medium delay GELP coder system has been proposed The system performs GCI detection by Pre-emphasised Maximum Likelihood Epoch (PMLED) detection (Chapter 5) For coding voiced speech, the GELP system uses an LF model excitation applied to an LP filter Two different procedures for glottal waveform estimation, Iterative Adaptive and Closed Phase Inverse Filtering, have been tested in the coding system To facilitate efficient coding, a differential variable rate LF parameter quantisation scheme has been proposed The performance of the GELP systems has been assessed in comparisons with that of the standard speech coders - LPC-10, CELP and GSM

Experiments show that the use of a parameterised glottal model presents advantages, in terms of speech quality, over a fixed glottal pulse excitation, especially in the low frequency range 0-1 kHz Furthermore, it has been shown that allocating bits to the transmission of the LF model parameters, r_g and r_a, can be more effective in terms of improving subjective speech quality than increasing the resolution of the gain parameter The glottal open quotient parameter, o_q, has been shown to have little impact on speech quality and is therefore not used in the proposed GELP system The use of a controllable glottal excitation signal also permits reduction of the LP filter order from 10th to 8th giving further savings in coding rate

Using a perceptual error measure, Bark Spectral Distortion, GELP has been shown to provide better speech quality than LPC-10 for male speech under conditions of telephone bandwidth measurements, input SNRs of up to 15 dB or source-receiver distances up to 50 cm For female speech, GELP has been shown to provide better BSD scores than LPC-10 under conditions of reasonably low distortion For low pitched male subjects, the mean GELP transmission rate is similar to that of LPC-10 while it rises to roughly twice that of LPC-10 for high pitched female speakers

In the case of high quality low pitched male speech material without band-limitation, GELP achieves speech quality in excess of CELP at roughly half the bit rate However, with increasing pitch and distortion, the quality provided by GELP rapidly falls below that of CELP Even for high pitched female subjects, the mean transmission rate provided by GELP is less than that of CELP

Although the test-bed GELP system described in this investigation is computationally expensive, it seems likely that a more efficient system could be developed In particular the LF fitting scheme could be modified to determine the best quantised fit, rather than quantise the best fit Modifications such as this would significantly reduce the computational complexity of GELP encoding - probably to much less than that of CELP encoding

Overall, the GELP approach shows promise for speech coding applications In this study, the advantages of using a parameterised glottal model have been clearly demonstrated However due to its lack of robustness, the system, in its present form, is unsuitable for telecommunication purposes Nevertheless, GELP may be of use in voice messaging applications for which low rate coding of high quality speech is essential It is believed that a further reduction of the bit rate of GELP can be achieved by using adaptive or vector quantisation of the LF parameters Also, as has been demonstrated elsewhere, the robustness of the system can be improved by using a stochastic or multi-band excitation in tandem with the glottal signal

CHAPTER 8

CONCLUSION

8 1 INTRODUCTION

The purpose of this chapter is threefold - to summarise the research described in the thesis, to evaluate the contribution of the work and to suggest directions for further investigation To this end, the remainder of the chapter is split into three sections, each covering one of these elements For further details and references, see the relevant chapters in the body of the thesis

8.2 SUMMARY OF THE THESIS

The aim of the investigation has been to develop and evaluate glottal processing techniques, with particular reference to glottal based speech coding For this purpose, four main studies were undertaken - a study of models for representing the effects of reverberation on speech recordings, development of an improved algorithm for Glottal Closure Instant (GCI) identification, a comparison of two automatic inverse filtering techniques and, finally, the proposal and evaluation of a Glottal Excited Linear Prediction (GELP) coding system for voiced speech

Reverberation has two effects on the recording of speech Firstly, the lip radiation impedance is altered from its free-field value due to interaction between the source and the reverberant field Secondly, the signal picked up at a microphone is distorted, relative to that observed in the free-field, due to the presence of secondary source to receiver transmission paths

The effect of reverberation on the lip radiation impedance was examined by the development of theory for predicting the variation in the radiation impedance at a vibrating piston set in an infinite baffle and operating in a reverberant enclosure The theory was confirmed by comparing the results of Monte Carlo simulations with measurements of the radiation impedance at a loudspeaker placed in a normally reverberant room Following this simulations using typical lip opening areas were conducted The results of these experiments verified that, for speech processing applications in typical enclosures, the variation of the lip radiation impedance due to reverberation is negligible

The simplest way to determine the effects of reverberation on speech processing algorithms is to process a single recording several times using different levels of added reverberation To artificially add reverberation to speech material requires the convolution of anechoic recordings with room impulse responses The Image Method of Berkley and Allen is one method for generating these responses In this thesis, the Image Method was evaluated by comparing narrowband impulse response measurements made in a typical room, to artificial responses generated by the Image Method It was found that the measured and simulated responses displayed a high degree of similarity in terms of decay rate and spectral variation The Image Method was therefore deemed satisfactory for the generation of artificially reverberant test speech material and was used for this purpose throughout the remainder of the investigation

Maximum Likelihood Epoch Detection (MLED) has shown promise as an accurate automatic technique for determining the GCI from voiced speech However in this investigation, the technique was shown to fail for certain speech material The cause of the problem was determined and a new formulation of the technique, Pre-emphasised Maximum Likelihood Epoch Detection (PMLED), was proposed Also, new post-processing techniques were developed to improve the robustness of PMLED by limiting the GCI search based on the periodicity of the speech and PMLED signals In tests, the PMLED technique was shown to be reliable in processing the material for which MLED failed Furthermore, PMLED was found to be accurate in determining the GCIs for vowel sounds, voiced fricatives, voiced plosives and nasals recorded by male and female subjects under conditions of noise and reverberation

The performance of two existing algorithms, Closed Phase (CPIF) and Iterative Adaptive Inverse Filtering (IAIF), for automatic glottal waveform estimation from the speech signal was evaluated in comparative tests In addition, glottal waveform parameterisation by time-domain fitting of the LF model was studied The conventional CPIF algorithm was augmented by the inclusion of a new multiple filter procedure for improving the robustness of the technique Also, LF fitting was expedited by the incorporation of a new polynomial based initialisation procedure In experiments involving natural speech material, both inverse filtering algorithms were found to be effective in terms of formant cancellation and glottal estimation for subjects of both sexes over most phonetic categories However, it was noted that in most instances IAIF provided smoother glottal waveform estimates indicating more effective formant cancellation The only category for which CPIF consistently outperformed IAIF was in the case of vowels with low first formants IAIF was also found to be more robust to distortions of the incoming speech signal, probably due to its longer analysis window The LF model was observed to provide close matches to the estimated glottal excitation in almost all cases, with the exception of voiced fricatives and voiced plosives The statistics of, and correlations between, the extracted LF parameters were presented and discussed with reference to previously published results

A low bit rate, medium delay GELP voiced speech coding system was proposed, developed and tested The system was based on a speech production model consisting of an LF glottal excitation applied to an all-pole vocal tract filter In the GELP encoder, PMLED was employed for GCI identification and glottal estimation was carried out by inverse filtering Both CPIF and IAIF algorithms were tested for this purpose The LF parameters were determined by time-domain fitting to the estimated glottal excitation Based on the quantised and re-synthesised LF input signal, the LP coefficients were calculated by an ARX estimation procedure Subsequently, the glottal gain parameter was optimised by matching the energy of the original and re-synthesised speech signals

Pitch quantisation was based on that of U S Federal Standard CELP, while the quantisers for the LP and LF parameters were derived from the processing of natural speech material In particular, optimum non-linear quantisers were developed for the LF parameters based on a Probability Density Function model matched to the observed parameter distributions The optimum bit allocation scheme for the source parameters was determined by a knockout procedure whereby bit allocation was tested by coding natural speech The experiments clearly demonstrated the importance of the dynamic leakage and glottal frequency parameters for providing good quality re-synthesis The final quantisation system was a variable rate scheme in which the LP coefficients were transmitted once per frame and the LF parameters were refreshed once per pitch period.

The performance of the GELP systems was established in comparison to that of three standard conventional coders - LPC-10, CELP and GSM. The same test data was used for all five systems and the quality of the coded speech was assessed using the Bark Spectral Distortion measure. Under low noise conditions, for all but the highest pitched speaker, IAIF based GELP coding was observed to provide better speech quality than CPIF based coding. At higher distortion levels, the difference in performance between the two GELP systems became less pronounced and less consistent. The speech quality provided by the IAIF GELP system was found to be in excess of that provided by LPC-10, especially in the low frequency range 0-1 kHz. Furthermore, under conditions of low distortion in the incoming speech signal, the speech quality of GELP was shown to be comparable with that of CELP. In experiments involving noisy and reverberant speech material, GELP was observed to be less robust than any of the conventional systems. Nevertheless, the performance degradation of the GELP systems was reasonably graceful. For low pitched male speech, GELP achieved a mean bit rate equivalent to that of LPC-10, while during high pitched female utterances GELP's mean coding rate rose close to that of CELP.

In conclusion, the thesis has proven that the effect of reverberation on the lip radiation impedance is negligible and that the Image Method is effective in generating artificial room impulse responses. An improved method for automatic GCI identification from the speech signal has been proposed and tested. In addition, the IAIF algorithm has been demonstrated to be superior to the CPIF technique, except in circumstances of proximate fundamental and first formant frequencies. The advantages, in terms of speech quality and coding efficiency, of using a controllable glottal waveform excitation in an LP based coder have been clearly demonstrated and GELP coding has been shown to be fairly robust to low levels of noise and reverberation.

8.3 CONTRIBUTION OF THE THESIS

The thesis has made a number of contributions to the body of knowledge regarding glottal based speech processing. This has included the proposal of new processing algorithms and the evaluation of existing techniques.

Inverse filtering algorithms attempt to extract the glottal excitation waveform from voiced speech. In order that the composite glottal signal may be accurately estimated, the speech signal must undergo little or no phase distortion prior to inverse filtering. One possible cause of such distortion is deviation of the lip radiation impedance from its free-field value due to the presence of reverberation. In this thesis, it has been established that, in normal enclosures, the variation of the lip radiation impedance due to reverberation is negligible. This establishes that glottal extraction by inverse filtering is always possible in normal rooms, provided that reverberation in the source-receiver channel is minimised. This can be done either by placing the microphone sufficiently close to the speaker or by applying de-reverberation techniques. More generally, the experiment confirms that any phase sensitive speech processing algorithm can be successfully applied in reverberant enclosures.

The Image Method provides a simple technique for the generation of artificial room impulse responses. This facilitates the production of reverberant test data for evaluating speech processing algorithms (Chapters 5, 6 and 7), conducting auditory experiments [Culling et al., 1994] and developing echo cancellation systems [Mourjopoulos, 1985]. The experiments described in this thesis have demonstrated that the Image Method is effective in generating reasonably accurate artificial room responses. This result confirms the validity of the assumptions underlying the Image Method and supports its use in all of the above application areas.

The pitch micro-melody of voiced speech carries phonetic, speaker and linguistic information to the human listener. Clearly, identification of this quantity is important for speech recognition, while its reproduction is crucial for natural sounding speech synthesis. Conventional pitch detection algorithms operate by determining the long-term periodicity of the speech signal. This approach smoothes out any short-term variation in the pitch and prevents identification of the micro-melody. Determination of GCIs is one method for identifying the pitch micro-melody. In this thesis an improved technique for GCI determination, PMLED, is proposed and demonstrated to be more reliable than a previously used method, MLED. As explained, the new GCI identification technique is of use in speech recognition, synthesis and coding applications. Furthermore, accurate and robust identification of the GCI is essential prior to Closed Phase Inverse Filtering.

Over the years, numerous methods have been proposed for estimation of the glottal waveform from voiced speech. Today, the most common approach is Closed Phase Inverse Filtering [Krishnamurthy and Childers, 1986]. Another technique, Iterative Adaptive Inverse Filtering [Alku, 1992a,b,c], has recently been proposed and has shown promise for automatic glottal estimation. This thesis presents the first comparative evaluation of the two algorithms. The results indicate that both methods perform well over most speech material. However, it was found that, in most cases, IAIF provides slightly more accurate glottal estimation and is more robust to distortion of the incoming speech signal. Additionally, IAIF has the advantage of not requiring accurate *a priori* GCI identification. These findings clearly indicate the superiority of the IAIF procedure and support the contention that it should replace CPIF in most glottal estimation applications. Since the algorithm is fully automatic, IAIF allows the processing of large amounts of speech data and presents a clear advance from the manual systems often cited in the literature.

Two new procedures to aid in glottal extraction were proposed and tested in the thesis. Firstly, a method for expediting LF model fitting based on a polynomial approximation was proposed. This technique provides reasonably accurate initialisation of the LF parameters at low computational cost and is useful in most analysis experiments involving the LF model. Secondly, a multiple inverse filtering procedure was incorporated in the CPIF algorithm. This procedure was demonstrated to improve the robustness of the Close Phase algorithm and is of use in applications where CPIF is the chosen method of glottal estimation.

A spin-off from the study evaluating the inverse filtering algorithms was the extraction of LF parameters from a variety of natural speech material. The statistics of and correlations between the derived LF parameters are published herein and provide a useful database for the study of voice source dynamics in connected speech.

Glottal based speech coding has the potential to provide improved naturalness relative to conventional coding systems because it is based on a more accurate speech production model Moreover, glottal coding techniques promise to provide low bit rate transmission since the glottal waveform parameters follow a smooth slowly time-varying vector trajectory The Glottal Excited Linear Prediction coding system developed in this study clearly demonstrates that allocating transmission bandwidth to the glottal parameters is effective in improving speech quality A full quantisation scheme is developed for the coder, including the specification of differential LF quantisers This is the first time that a GELP coding scheme has been quantitatively compared with conventional systems The findings clearly illustrate the potential of GELP, especially in modelling the low frequency content (0-1 kHz) of the voiced speech signal Also, the low transmission rate required for the LF parameters presents obvious advantages to the use of a conventional stochastic or impulse excitation signal Unfortunately, due to its phase sensitivity, GELP remains unsuitable for certain coding applications However, the approach seems very suitable to voice messaging applications which require high quality speech, inexpensive decoders and low transmission rates To the author's knowledge this was the first investigation of a GELP system with an LF excitation model In addition, this was the first published work to provide a quantitative evaluation of the transmission rate and speech quality achievable by a low rate, medium delay glottal coder in direct comparisons with standard conventional coders

8 4 SUGGESTIONS FOR FURTHER WORK

The work described in this thesis points to several avenues of potentially fruitful further investigation

The Image Method has been shown to be adequate for the generation of room reverberant responses in small rectangular enclosures However, the method lacks precision While newer methods model the room reverberation process in greater detail, they too fail to exactly predict room impulse responses In order to provide precise methods for the prediction of room responses, a better understanding of the reverberation process is required In particular, more detailed analysis must be made of measured reverberant responses and their relationship to the rooms under investigation Certainly, due to the introduction of faster Digital Signal Processors (DSPs), it seems that the time is right for the development of more complex and more precise methods for simulating the reverberation process

In the last few years a number of new algorithms for GCI identification have been developed, the most promising of which appear to be PMLED (Chapter 5), Singular Value Decomposition [Ma et al, 1994] and Murgia's method [Murgia et al, 1994] To aid system developers in algorithm selection and to provide direction for future research, a comparative evaluation of these GCI identification algorithms is required Ideally, the evaluation should cover a wide range of speakers and recording conditions

Deconvolution of the glottal excitation and vocal tract transfer function is a perennial problem in speech science The IAIF algorithm has been demonstrated to provide accurate and reasonably robust glottal waveform estimation under most recording conditions However, the method provides poor glottal estimation when the first formant nears the fundamental frequency Some investigation of this problem is required and improvements may be achieved by the introduction of time or frequency selective LP analysis techniques

The LF model is an accurate and efficient low-dimensional representation for the glottal waveform However, the gradient based minimisation procedure normally used for LF fitting is extremely computationally complex Although polynomial based initialisation of the LF fit alleviates the problem, a much more computationally efficient method must be developed if automatic fitting of the LF model is to be used more extensively Note that some work has already been published on this topic [Qi and Bi, 1994]

Although the GELP coding system proposed in this thesis provides promising results, it is thought that its performance could be enhanced in a number of ways The LF quantisation scheme could be made more efficient by the use of adaptive or vector quantisation techniques The GELP system could be modified for fixed rate transmissions by limiting the refresh rate for the LF parameters The LP filter and pitch quantisation schemes were selected purely arbitrarily - an empirically based enquiry is needed to create more efficient schemes The robustness of the coder could be improved by the inclusion of a multi-band or stochastic excitation operating in tandem with the glottal model [Bergstrom and Hedelin, 1988, 1989] Also, supplementary work is required to extend the coder to provide for the transmission of unvoiced speech Lastly, methods for modelling other phonetic categories, such as voiced plosives or nasals, should be considered

Finally, extensive subjective tests should be carried out to assess the quality of GELP coded speech While the BSD measure provides useful information, the ultimate judge of speech quality must be the human listener

REFERENCES

Alku, P (1990a) "Low bit rate speech coding with glottal linear prediction", *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS) 1990*, vol 3, pp 2149-2152

Alku, P (1990b) "Glottal-LPC based coding of telephone band vowels with simple all-pole excitation", *Proceedings of the International Conference on Spoken Language Processing (ICSLP) 1990*, pp 89-92

Alku, P (1991) "Coding of telephone band speech at 5 kbit/s based on glottal excitation analysis", *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS) 1991*, pp 332-335

Alku, P (1992a) "An automatic method to estimate the time-based parameters of the glottal pulseform', *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1992*, vol 2, pp 29-32

Alku, P (1992b) "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering", *Speech Communication*, vol 11, no 2-3, pp 109-118

Alku, P (1992c) *An automatic inverse filtering method for the analysis of glottal waveforms*, Ph D Thesis, Helsinki University of Technology

Alku, P and Laine, U K (1989a) "A new glottal LPC method of low complexity for speech analysis and coding", *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH) 1989*, vol 2, pp 31-34

Alku, P and Laine, U K (1989b) "A new glottal LPC method for voice coding and inverse filtering", *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS) 1989*, vol 3, pp 1831-1834

Allen, D R and Strong, W J (1985) "A model for the synthesis of natural sounding vowels", *Journal of the Acoustical Society of America*, vol 78, no 1, pp 58-69

Allen, J B and Berkley, D A (1979) "Image method for efficiently simulating small-room acoustics", *Journal of the Acoustical Society of America*, vol 65, no 4, pp 943-950

Almeida, L B and Tribolet, J M (1982) "Harmonic coding a low bit-rate good-quality speech coding technique", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1982* pp 1664-1667

Ananthapadmanabha, T V (1982) "Intelligibility carried by speech-source functions implication for theory of speech perception", *Quarterly Progress and Status Report*, Speech Transmission Laboratory, 1982, pp 49-64

Ananthapadmanabha, T V (1984) "Acoustic analysis of voice source dynamics", *Quarterly Progress and Status Report*, Speech Transmissions Laboratory, 1984, no 2-3, pp 1-24

Ananthapadmanabha, T V and Fant, G (1982) "Calculation of the true glottal flow and its components", *Speech Communication* vol 1 pp 167-184 Also appears in *Quarterly Progress and Status Report* Speech Transmission Laboratory 1982, no 1, pp 1-30

Ananthapadmanabha, T V and Yegnanarayana, B (1975) "Epoch extraction of voiced speech" *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol 23, no 6, pp 562-569

Ananthapadmanabha, T V and Yegnanarayana, B (1979) "Epoch extraction from linear prediction residual for identification of closed glottis interval", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol 27, pp 309-319

André-Obrecht, R (1988) "A new statistical approach for the automatic segmentation of continuous speech signals", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol 36, no 1, pp 29-40

Ansari, R (1987) "IIR discrete-time Hilbert transformers", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol 35, pp 1116-1119

Anthony, D K and Elliott, S J (1991) "A comparison of three methods of measuring the volume velocity of an acoustic source", *Journal of the Audio Engineering Society*, vol 39, no 5, pp 355-366

Astrom, K J and Eykhoff, P (1971) "System identification - a survey", *Automatica*, vol 7, pp 123-162

Atal, B S and Hanauer, S L (1971) "Speech analysis and synthesis by linear prediction of the speech wave", *Journal of the Acoustical Society of America*, vol 50, pp 637-655

Atal, B S and Remde, J R (1982) "A new model of LPC excitation for producing natural-sounding speech at low bit rates", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1982*, vol 1, pp 614-617

Atal, B S and Schroeder, M R (1978) "Linear prediction analysis based on a pole-zero representation", *Journal of the Acoustical Society of America*, vol 64, no 5, pp 1310-1318

Basseville, M and Benveniste, A (1986) "Detection of abrupt changes in signals and dynamical systems", in *Lecture Notes in Control Theory and Information Sciences*, New York, Springer-Verlag

Bendat, J S and Piersol A G (1971) *Random Data Analysis and Measurement Procedures*, New York, John Wiley & Sons

Benitez, M C , Galvez, J A , Rubio, A and Diaz, J (1992) "A codification of error signal by splines functions" in P Laface and R De Mori (eds) *Speech Recognition and Understanding Recent Advances*, NATO ASI Series, vol F75, Berlin, Springer-Verlag

Beranek, L L (1992) *Acoustical Measurements* 2nd edition, New York, American Institute of Physics

Bergstrom, A and Hedelin, P (1988) 'Extended glottal LPC with mixed excitation", *Proceedings of the IEEE EUROCON 1988*, pp 28-31

Bergstrom, A and Hedelin, P (1989) Code-book driven glottal pulse analysis', *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1989*, pp 53-56

Berouti, M G (1976) "Estimation of the glottal volume velocity by the linear predictive inverse filter", Ph D Thesis, University of Florida.

Berouti, M G , Childers, D G and Paige, A (1977) "Glottal area versus glottal volume velocity", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1977*, vol 1 pp 33-36

Bleakley, C J and Scaife, R (1994) "The variation of the lip radiation impedance in a reverberant enclosure", *Proceedings of the European Signal Processing Conference (EUSIPCO) 1994*, vol 3, pp 1689-1692

Bleakley, C J and Scaife, R (1995) "New formulas for predicting the accuracy of acoustical measurements made in noisy environments using the averaged m-sequence correlation technique", *Journal of the Acoustical Society of America*, vol 97, no 2, pp 1329-1332

Blomberg, M (1991) "Adaptation to a speaker's voice in a speech recognition system based on synthetic phoneme references", *Speech Communication*, vol 10, pp 453-461

Blomberg, M (1993) "Synthetic phoneme prototypes and dynamic voice source adaption in speech recognition", *Quarterly Progress and Status Report*, Speech Transmission Laboratory, 1993, no 4, pp 97-140

Borden, G and Harris, K (1980) *Speech Science Primer, Physiology, Acoustics and Perception*, Baltimore, Maryland, Williams & Wilkins

Brandstein, M S , Monta, P A , Hardwick, J C and Lim, J S (1990) "A real-time implementation of the improved MBE speech coder", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1990*, vol 1, pp 5-8

Brookes, D M and Naylor, P A (1988) "Speech production modelling with variable glottal reflection coefficient", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1988*, pp 671-674

Campbell, J P Jr and Tremain, T E (1986) "Voiced/unvoiced classification of speech with applications to the U S government LPC-10e algorithm", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1986*, pp 473-476

Campbell, J P Jr , Tremain, T E and Welch, V C (1991) "The DOD 4 8 KBPS Standard (Proposed Federal Standard 1016)" in B S Atal, V Cuperman and A Gersho (eds) *Advances in Speech Coding*, Norwell, MA , Kluwer, pp 121-133

Carlson, R , Granstrom, B and Karlsson, I (1990) "Experiments with voice modelling in speech synthesis", *Quarterly Progress and Status Report*, Speech Transmission Laboratory, 1990, no 2-3, pp 53-61 Also appears in (1991) *Speech Communication*, vol 10, pp 481-489

Chan, D S F and Brookes, D M (1989) "Variability of excitation parameters derived from robust closed phase glottal inverse filtering", *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH) 1989*, vol 2, pp 199-202

Chen, W T and Chi, C Y (1993) 'New inverse filter criteria for identification and deconvolution of nonminimum-phase systems by single cumulant slice", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1993*, vol 4, pp 192-195

Cheng, Y M and O'Shaughnessy, D (1989) "Automatic and reliable estimation of glottal closure instant and period", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol 37, no 12, pp 1805-1815

Cheng, Y M and O'Shaughnessy, D (1993) "On 450-600 b/s natural sounding speech coding" *IEEE Transactions on Speech and Audio Processing*, vol 1, no 2, pp 207-220

Chiba, T and Kajiyama, M (1941) *The Vowel, Its Nature and Structure*, Tokyo, Tokyo-Kaiseikan Pub Co

Childers, D G and Ahn, C (1995) "Modeling the glottal volume-velocity waveform for three voice types", *Journal of the Acoustical Society of America*, vol 97, no 1, pp 505-519

Childers, D G and Hu, H T (1994) "Speech synthesis by glottal excited linear prediction", *Journal of the Acoustical Society of America*, vol 96, no 4, pp 2026-2036

Childers, D G and Wong, C F (1994) "Measuring and modeling vocal source-tract interaction", *IEEE Transactions on Biomedical Engineering*, vol 41, no 7, pp 663-671

Childers, D G and Wu, K (1990) "Quality of speech produced by analysis-synthesis", *Speech Communication*, vol 9, pp 97-117

Childers, D G and Lee, C K (1991) "Vocal quality factors analysis, synthesis and perception", *Journal of the Acoustical Society of America*, vol 90, no 5, pp 2394-2410

Childers, D G , Hicks, D M , Moore, G P , Eskenazi, L and Lalwani, A L (1990) "Electroglottography and vocal fold physiology", *Journal of Speech and Hearing Research*, vol 33, no 2, pp 245-254

Childers, D G , Naik, J M , Larar, J N and Krishnamurthy, A K (1985) "Electroglottography, speech and ultra-high speech cinematograph", *Proceedings of the Vocal Fold Physiology Conference 1985*, pp 202-220

Childers, D G , Wu, K and Hicks, D M (1987) "Factors in voice quality acoustic features related to gender", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1987*, pp 293-296

Chu, W T (1978) "Comparison of reverberation measurements using Schroeder's impulse method and decay-curve averaging method", *Journal of the Acoustical Society of America*, vol 63, no 5, pp 1444-1450

Cranen, B and Boves, L (1988) "On the measurement of glottal volume velocity", *Journal of the Acoustical Society of America*, vol 84, no 3, pp 888-900

Crosmer, J.R and Barnwell, T P (1985) "A low bit rate segment vocoder based on line spectrum pairs", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1985*, vol 1, pp 240-243

Culling, J F , Summerfield, Q and Marshall, D H (1994) 'Effects of simulated reverberation on the use of binaural cues and fundamental-frequency differences for separating concurrent vowels", *Speech Communication*, vol 14, pp 71-95

Cummings, K E and Clements, M A (1990) 'Analysis of glottal waveforms across stress styles", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1990*, vol 1 pp 369-372

Cummings, K E and Clements, M A (1992) "Improvements to and applications of analysis of stressed speech using glottal waveforms", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1992*, vol 2, pp 25-28

Cummings, K E and Clements, M A (1993) "Application of the analysis of glottal excitation of stressed speech to speaking style modification", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1993*, vol 2, pp 207-210

Czyzewski, Z and Nabelek, A K (1991) "Image method for simulating reverberant room response with directional receiver and source", *Journal of the Acoustical Society of America*, vol 90, no 4, pt 2, pp 2239

Davy, J L (1981) "The relative variance of the transmission function of a reverberation room", *Journal of Sound and Vibration*, vol 77, no 4, pp 455-479

de Mori, R , Laface, P , Makhonine, V A and Mezzalama, M (1977) "A syntactic procedure for the recognition of glottal pulses in continuous speech", *Pattern Recognition*, vol 9, pp 181-189

de Veth, J , van Golstein Brouwers, W and Boves, L (1989) "Robust ARMA analysis for accurate determination of the system parameters of the voice source and vocal tract", *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH) 1989*, vol 2, pp 43-46

Deller, J R Jr (1981) "Some notes on closed phase glottal inverse filtering", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol 29, no 4, pp 917-919

Deller, J R Jr (1983) "On the time domain properties of the two-pole model of the glottal waveform and implications for LPC", *Speech Communication An Interdisciplinary Journal*, vol 2, pp 57-63

Deller, J R Jr , Proakis, J G and Hansen, J H L (1993) *Discrete-time Processing of Speech Signals*, New York, Macmillan Pub

Dologlou, I and Carayannis, G (1989) "Pitch detection based on zero-phase filtering", *Speech Communication*, vol 8, no 4, pp 309-318

Dologlou, I and Carayanms, G (1991) "A reply to 'Some remarks on the halting criterion for iterative low-pass filtering in a recently proposed pitch detection algorithm", *Speech Communication*, vol 10, pp 227-228

Dunn, H K (1950) "The calculation of vowel resonances, and an electrical vocal tract", *Journal of the Acoustical Society of America*, vol 22, pp 740-753

El Mallawany, I (1977) "Area function over the closed glottis interval", *Proceedings of the Articulatory Modeling and Phonetics Symposium 1977*, pp 65-76

ETSI (1989) "Recommendation 06 10 GSM full-rate speech transcoding", European Telecommunications Standards Institute

Fabre, P (1957) 'Un procédé électrique percutané d'inscription de l'accolement glottique au cours de la phonation Glottographie de haute fréquence premiers resultats" *Bull Acad Nat Médition*, pp 66-69

Fant, G (1956) "On the predictability of formant levels and spectrum envelopes from formant frequencies", in *For Roman Jakobson*, The Hague, The Netherlands, Mouton

Fant, G (1970) *Acoustic Theory of Speech Production*, 2nd edition, The Hague, The Netherlands, Mouton

Fant, G (1979a) "Glottal source and excitation analysis", *Quarterly Progress and Status Report* Speech Transmission Laboratory 1979 no 1 pp 85-107

Fant, G (1979b) "Voice source analysis - a progress report" *Quarterlv Progress and Status Report*, Speech Transmission Laboratory, 1979, no 3-4, pp 31-54

Fant, G (1982) "The voice source - acoustic modelling", *Quarterly Progress and Status Report*, Speech Transmission Laboratory, 1982, no 4, pp 28-48

Fant, G (1986) "Glottal flow models and interaction", *Journal of Phonetics*, vol 14, pp 393-399

Fant, G (1993) "Some problems in voice source analysis", *Speech Communication*, vol 13, no 1-2, pp 7-22

Fant, G and Lin, Q (1988) "Frequency domain interpretation and derivation of glottal flow parameters", *Quarterly Progress and Status Report*, Speech Transmission Laboratory, 1988, no 2-3, pp 1-21

Fant, G , Liljencrants, J and Lin, Q (1985) "A four-parameter model of glottal flow", *Quarterly Progress and Status Report*, Speech Transmission Laboratory, 1985, no 4, pp 1-13

Fant, G , Ondráckovà, J , Lindqvist, J and Sonesson, B (1966) "Electrical glottography", *Quarterly Progress and Status Report*, Speech Transmission Laboratory, 1966, no 4, pp 15-21

Faris, W R and Timothy, L K (1974) "Linear predictive coding with zeros and glottal wave", *Proceedings of the National Electronics Conference 1974*, vol 29, pp 409-411

Fenichel, R (1991) "Federal Standard 1016 Telecommunications Analog to digital conversion of radio voice by 4,800 bit/second Code Excited Linear Prediction (CELP)", National Communications System, Office of Technology and Standards, Washington, D C , General Services Administration

Fenichel, R (1992) "Details to assist in implementation of federal standard 1016 CELP", *Technical Information Bulletin 92-1*, National Communications System, Office of Technology and Standards, Washington, D C , General Services Administration

Flanagan, J L (1958) "Some properties of the glottal sound source", *Journal of Speech Hearing Research*, vol 1, pp 99-111

Flanagan, J L (1972) *Speech Analysis, Synthesis and Perception*, 2nd edition, New York, Springer-Verlag

Flanagan, J L , Ishizaka, K and Shipley, K L (1975) "Synthesis of speech from a dynamic model of the vocal cords and vocal tract", *Bell System Technical Journal*, vol 54, no 3, pp 485-506

Flanagan, J L and Landgraf, L L (1968) "Self-oscillating source for vocal-tract synthesis", *IEEE Transactions on Audio and Electroacoustics*, vol 16, pp 57-64

Fourcin, A J (1974) "Laryngographic examinations of vocal fold vibrations", in B Wyke (ed) *Ventilatory and Phonatory Control*, London, Oxford University Press, pp 315-333

Fourcin, A J (1986) "Electrolaryngographic assessment of vocal fold function", *Journal of Phonetics*, vol 14, pp 435-442

Fourcin A J and Abberton, E (1971) "First applications of a new laryngograph", *Medical and Biological Illustration*, vol 21 pp 172-182

Fries, G (1994) 'Hybrid time- and frequency-domain speech synthesis with extended glottal source generation", *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) 1994* vol 1 pp 581-584

Fujisaki H and Ljungqvist, M (1986) "Proposal and evaluation of models for the glottal source waveform , *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1986*, pp 1605-1608

Fujisaki, H and Ljungqvist, M (1987) "Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1987*, vol 2 pp 637-640

Funada, T (1989) "A new method for extracting the glottal closure intervals of voiced speech", *Journal of the Acoustical Society of Japan*, vol 10, no 6 pp 349-355

Gersho, A (1994) "Advances in speech and audio compression", *Proceedings of the IEEE*, vol 82, no 6, pp 900-918

Gobl, C (1988) "Voice source dynamics in connected speech", *Quarterly Progress and Status Report*, Speech Transmission Laboratory, 1988, no 1, pp 123-159

Gobl, C (1989) "A preliminary study of acoustic voice quality correlations", *Quarterly Progress and Status Report*, Speech Transmission Laboratory, 1989, no 4, pp 9-22

Gobl, C and Ni Chasaide, A (1988) "The effects of adjacent voice/voiceless consonants on the vowel voice source a cross language study", *Quarterly Progress and Status Report*, Speech Transmission Laboratory, 1988, no 2-3, pp 23-59

Gobl, C and Ní Chasaide, A (1992) "Acoustic characteristics of voice quality", *Speech Communication*, vol 11, pp 481-490

Guérin, B , Mrayati, M and Carré, R (1976) "A voice source taking account of coupling with supraglottal cavities", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1976*, pp 47-50

Guo, C G and Scherer, R C (1993) "Finite element simulation of glottal flow and pressure", *Journal of the Acoustical Society of America*, vol 94, no 2, pt 1, pp 688-700

Hamlet, S L and Reid, J M (1972) "Transmission of ultrasound through the larynx as a measure of determining vocal fold activity", *IEEE Transactions on Biomedical Engineering*, vol 19, pp 34-37

Hanson, D G , Gerratt, B.R and Berke, G S (1990) "Frequency, intensity and target matching effects on photoglottographic measures of open quotient and speed quotient", *Journal of Speech and Hearing Research*, vol 33, pp 45-50

Harris, J D and Nelson, D (1993) "Glottal pulse alignment in voiced speech for pitch determination *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1993*, vol 2, pp 519-522

Hedelin, P (1981) "A tone-oriented voice-excited vocoder", *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) 1981* pp 205-208

Hedelin, P (1984) "A glottal LPC-vocoder", *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) 1984*, pp 1 6 1-1 6 4

Hedelin, P (1986) "High quality glottal LPC-vocoding", *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) 1986* pp 465-468

Hedelin, P (1988) "Phase compensation in all-pole speech analysis", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1988*, pp 339-342

Heinz, R (1993) "Binaural room simulation based on an image source model with addition of statistical methods to include the diffuse sound scattering of walls and to predict the reverberant tail", *Applied Acoustics*, vol 38, pp 145-159

Helstrom, C W (1960) *Statistical Theory of Signal Detection*, New York, Pergamon Press

Hertegård, S and Gauffin, J (1995) "Glottal area and vibratory patterns studies with simultaneous stroboscopy, flow glottography and electroglottography", *Journal of Speech and Hearing Research*, vol 38, pp 85-100

Hess, W (1983) *Pitch Determination of Speech Signals Algorithms and Devices*, New York, Springer-Verlag

Hess, W and Indefrey, H (1987) "Accurate time-domain pitch determination of speech signals by means of a laryngograph", *Speech Communication*, vol 6, no 1, pp 55-68

Holmes, J N (1962) "An investigation of the volume velocity waveform at the larynx during speech by means of an inverse filter" in Fant, G (ed) *Proceedings of the Speech Communication Seminar 1962*, Speech Transmission Laboratory, Royal Institute of Technology, vol 1, paper B4 Also appears in *Congress Report 4th International Congress on Acoustics*, pp 1-4

Holmes, J N (1973) "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer", *IEEE Transactions on Audio and ElectroAcoustics*, vol 21, pp 298-305

Holmes, J N (1975) "Low-frequency phase distortion of speech recordings", *Journal of the Acoustical Society of America*, vol 58, no 3, pp 747-749

Holmes, J N (1976) "Formant excitation before and after glottal closure", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1976*, pp 39-42

Hult, G (1991) "Some remarks on a halting criterion for iterative low-pass filtering in a recently proposed pitch detection algorithm", *Speech Communication* vol 10, no 3, pp 223-226

Hunt, M J , Bridle, J S and Holmes, J N (1978) "Interactive digital inverse filtering and its relation to linear prediction methods", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1978*, vol 1 pp 15-19

IEEE (1969) "IEEE recommended practice for speech quality measurements", *IEEE Transactions on Audio and Electroacoustics*, Sept 1969, pp 227-246

IEEE (1979) *Programs for Digital Signal Processing*, New York, John Wiley & Sons

Iijima H , Miki, N and Nagai, N (1992) "Glottal impedance based on a finite element analysis of two-dimensional unsteady viscous flow in a static glottis", *IEEE Transactions on Signal Processing* Vol 40, no 9, pp 2125-2135

Isaksson A and Millnert, M (1989) "Inverse glottal filtering using a parameterized input model" *Signal Processing*, vol 18, no 4, pp 435-445

Ishizaka K and Flanagan, J L (1972) "Synthesis of voiced sounds from a two-mass model of the vocal cords" *Bell System Technical Journal* vol 51 no 6 pp 1233-1268

ITU (1993) Special Issue on the CCITT Recommendation G 728 LD-CELP , *Speech Communication* vol 13, no 2 pp 97-204

Jansen, J , Cranen, B and Boves, L (1991) "Modelling of source characteristics of speech sounds by means of the LF-model", *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH) 1991*, vol 1, pp 259-262

Jayant, N S (1974) "Digital coding of speech waveforms PCM, DPCM and DM quantizers", *Proceedings of the IEEE*, vol 62, pp 611-632

Jayant, N S (1990) "High-quality coding of telephone speech and wideband audio", *IEEE Communications Magazine*, Jan 1990, pp 10-20

Jetzt, J J (1979) "Critical distance measurement of rooms from the sound energy spectral response", *Journal of the Acoustical Society of America*, vol 65, no 5, pp 1204-1211

Karlsson, I (1988) "Glottal waveform parameters for different speaker types", *7th FASE Symposium, Proceedings of Speech*, vol 1, pp 225-231 Also appears in *Quarterly Progress and Status Report*, Speech Transmission Laboratory, 1988, no 2-3, pp 61-69

Karlsson, I (1990) "Voice source dynamics for female speakers", *Proceedings of the International Conference on Spoken Language Processing 1990*, vol 1, pp 69-72

Karlsson, I (1991) "Female voices in speech synthesis", *Journal of Phonetics*, vol 19, pp 111-120

Karlsson, I (1992) "Modelling voice variations in female speech synthesis", *Speech Communication*, vol 11, pp 491-495

Kelly, J L and Lochbaum, C C (1962) "Speech synthesis", *Proceedings of the 4th International Congress on Acoustics*, vol G42, pp 1-4

Kemp, D P , Collura, J S and Tremain, T E (1991) "Multi-frame coding of LPC parameters at 600-800 bps", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1991*, pp 609-612

Kinsler, L E , Frey, A R , Coppens, A B and Sanders, J V (1982) *Fundamentals of Acoustics*, 3rd edition, New York, John Wiley & Sons

Klatt, D H and Klatt, L C (1990) "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *Journal of the Acoustical Society of America*, vol 87, no 2, pp 820-857

Kleijn, W B (1991) "Continuous representations in linear predictive coding", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1991*, vol 1, pp 201-204

Kleijn, W B and Granzow, W (1991) "Methods for waveform interpolation in speech coding", *Digital Signal Processing*, vol 1, no 4, pp 215-230

Kompis, M and Dillier, N (1993) 'Simulating transfer functions in a reverberant room including source directivity and head-shadow effects", *Journal of the Acoustical Society of America*, vol 93 no 5, pp 2779-2787

Krishnamurthy, A K (1990) "Glottal source models for speech coding and synthesis' *Proceedings of the 32nd Midwest Symposium on Circuits and Systems*, vol 1, pp 93-96

Krishnamurthy, A K (1992) Glottal source estimation using a sum-of-exponentials model , *IEEE Transactions on Signal Processing*, vol 40, pp 682-686

Krishnamurthy, A.K and Childers, D.G. (1981) "Vocal fold vibratory patterns: comparison of film and inverse filtering", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1981*, vol. 1, pp. 133-136.

Krishnamurthy, A.K. and Childers, D.G. (1986) "Two channel speech analysis", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 730-743.

Kuttruff, H. (1991) *Room Acoustics*, 3rd edition, New York, Elsevier Applied Science.

Ladefoged, P.A. (1975) *A Course in Phonetics*, New York, Harcourt Brace Jovanovich College Pub.

Laine, U.K. (1982) "Modeling of lip radiation impedance in two-domain", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1982*, pp. 1992-1995.

Lalwani, A.L. and Childers, D.G. (1991) "Modelling vocal disorders via formant synthesis", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal (ICASSP) 1991*, vol. 1, pp. 505-508.

Larar, J.N., Alsaka, Y.A. and Childers, D.G (1985) "Variability in closed phase analysis of speech", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1985*, pp. 1089-1092.

Larsson, B. (1977) "Pitch tracking in music signals", *Quarterly Progress and Status Report*, Speech Transmission Laboratory, 1977, no. 4, pp. 1-8.

Lecluse, F.L.E., Brocaar, M.P. and Verschure, J. (1975) "The electroglottography and its relation to glottal activity", *Folia Phoniatr.*, vol. 27, pp. 215-224.

Leung, S.H., Peng, L.F., Wong, O.Y., Chan, C.F. and Luk, A. (1990) "On an efficient decomposition of LPC excitation for producing natural sounding speech", *IEEE Region 10 Conference on Computer and Communication Systems 1990*, pp.329-333.

Lewers, T. (1993) "A combined beam tracing and radiant exchange computer model of room acoustics", *Applied Acoustics*, vol. 38, pp. 161-178.

Liljencrants, J. (1991) "Numerical simulations of glottal flow", *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH) 1991*, vol. 1, pp. 255-258.

Lindbolm, B.E.F. and Sundberg, J.E.F. (1971) "Acoustic consequences of lip, tongue, jaw, and larynx movement", *Journal of the Acoustical Society of America*, vol. 50, pp. 1166-1179.

Lindqvist, J. (1964) "Inverse filtering. Instrumentation and techniques", *Quarterly Progress and Status Report*, Speech Transmission Laboratory, 1964, no. 4, pp. 1-4.

Lindqvist, J. (1965) "Studies of the voice source by means of inverse filtering technique", *5th International Congress Acoustics*, vol. 1, paper A35. Also appears in *Quarterly Progress and Status Report*, Speech Transmission Laboratory, 1965, no. 2, pp. 8-13.

Liu, Y.J. (1989) "A high quality speech coder at 400 bps", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1989*, pp. 204-206.

Liu, Y.J. (1990) "A high quality speech coder at 600 bps", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1990*, pp. 645-648.

Liu, Y.J. (1991) "A robust 400-bps speech coder against background noise", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1991*, pp. 601-604.

Lobo, A P and Ainsworth, W A (1988) "Variation of glottal pulse shape with fundamental frequency", *7th FASE Symposium, Proceedings of Speech*, vol 1, pp 217-224

Lobo, A P and Ainsworth, W A (1992) "Evaluation of a glottal ARMA model of speech production", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1992*, vol 2, pp 13-16

Lu, J , Murakami, H and Kasuya, H (1990) "Estimation of vocal tract transfer functions using multi-closure intervals linear prediction", *Transactions of the Institute of Electronic, Information and Computer Engineers*, vol J73A, no 5, pp 1011-1014

Lucero, J C (1993) "Dynamics of the two-mass model of the vocal folds equilibria, bifurcations and oscillation region", *Journal of the Acoustical Society of America*, vol 94, no 6, pp 3104-3111

Ma, C , Kamp, Y and Willems, L F (1994) "A Frobenius norm approach to glottal closure detection from the speech signal", *IEEE Transactions on Speech and Audio Processing*, vol 2, no 2, pp 258-265

Makhoul, J (1975) "Linear prediction a tutorial review", *Proceedings of the IEEE*, vol 63, pp 561-580

Markel, J D and Gray, A H (1976) *Linear Prediction of Speech*, New York, Springer-Verlag

Markel, J D and Wong, W Y (1976) "Considerations in the estimation of glottal volume velocity waveforms", *Journal of the Acoustical Society of America*, vol 59, supl 1, pp S96-S97 (Paper RR6, 91st Meeting ASA)

Marques, J S , Almeida, L B and Tribolet, J M (1990) "Harmonic coding at 4 8 kb/s", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1990*, vol 1, pp 17-20

Mataušek, M R and Batalov, V S (1980) "A new approach to the determination of the glottal waveform", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol 28, no 6, pp 616-622

Mathews, M V , Miller, J E and David, E E Jr (1961) "An accurate estimate of the glottal waveshape', *Journal of the Acoustical Society of America*, vol 33, pp 843 (Paper J8, 61st Meeting ASA)

Max, J (1960) 'Quantizing for minimum distortion", *IEEE Transactions on Information Theory*, vol 6, pp 7-12

McAulay, R J and Quatieri, T F (1986) Speech analysis-synthesis based on a sinusoidal representation", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol 34, pp 744-754

McAulay, R J and Quatieri, T F (1992) Low-rate speech coding based on the sinusoidal model' in M Sondhi and S Furui (eds) *Advances in Acoustics and Speech Processing*, New York, Marcel Deckker, pp 165-207

McCree, A V and Barnwell, T P (1992) "Improving the performance of a mixed-excitation LPC vocoder in acoustic noise ', *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) 1992*, vol 2, pp 137-140

McCree, A V and Barnwell, T P (1993) "Implementation and evaluation of a 2400 bps mixed excitation LPC vocoder", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1993*, vol 2, pp 159-162

Miki, N , Motoki, K and Nagai, N (1987) "A lattice filter model with accurate lip impedance for dynamic articulatory movement", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1987*, pp 952-955

Milenkovic, P (1986) "Glottal inverse filtering by joint estimation of an AR system with a linear input model", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol 34, no 1, pp 28-42

Milenkovic, P (1993) "Voice source model for continuous control of pitch period", *Journal of the Acoustical Society of America*, vol 93, no 2, 1087-1096

Miller, J A , Pereira, J C and Thomas, D W (1988) "Fluid flow through the larynx channel", *Journal of Sound Vibration*, vol 121, pp 277-290

Miller, J E and Mathews, M V (1963) "Investigation of the glottal waveshape by automatic inverse filtering", *Journal of the Acoustical Society of America*, vol 35, pp 1876 (Paper B3, 66th Meeting ASA)

Miller, R L (1959) "Nature of the vocal cord wave", *Journal of the Acoustical Society of America*, vol 31, no 6, pp 667-677

Mizuno, H and Abe, M (1994) "Voice conversion based on piecewise linear conversion rules of formant frequency and spectrum tilt", *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) 1994*, vol 1, pp 469-472

Morse, P M and Ingard, K U (1968) *Theoretical Acoustics*, New York, McGraw-Hill

Moulines, E and Di Francesco, R (1990) "Detection of the glottal closure by jumps in the statistical properties of the speech signal", *Speech Communication*, vol 9, no 5-6, pp 401-418

Mourjopoulos, J (1985) "On the variation and invertibility of room impulse response functions", *Journal of Sound and Vibration*, vol 102, no 2, pp 217-228

Murgia, C , Mann, I and Feng, G (1994) "An algorithm for the estimation of glottal closure instants using the sequential detection of abrupt changes in speech signals", *Proceedings of the European Signal Processing Conference (EUSIPCO) 1994*, pp 1685-1688

Nakagawa, K , Miyajima, T and Tahara, Y (1993) "An improved geometrical sound field analysis in rooms using scattered sound and an audible room acoustic simulator", *Applied Acoustics* vol 38, pp 115-129

Ni, J and Alipour, F (1993) "Animation and numerical simulations of laryngeal airflow", *Proceedings of the 3rd International Conference on CAD and Computer Graphics*, vol 2, pp 849-854

Nishiguchi M , Matsumoto, J , Wakatsuki, R and Ono, S (1993) "Vector quantized MBE with simplified V/UV division at 3 0 kbps", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1993*, vol 2, pp 151-154

O Shaughnessy, D (1987) *Speech Communication Human and Machine*, Redding, Mass , Addison-Wesley

Orlikoff, R F (1991) "Assessment of the dynamics of vocal fold contact from the electroglottogram data from normal male subjects", *Journal of Speech and Hearing Research*, vol 34, pp 1066-1072

Parthasarathy, S and Tufts, D W (1987) "Excitation-synchronous modelling of voiced speech", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol 35, no 9, pp 1241-1249

Peterson, P M (1986) "Simulating the response of multiple microphones to a single acoustic source in a reverberant room", *Journal of the Acoustical Society of America*, vol 80, no 5, pp 1527-1529

Pierrehumbert, J B (1989) "A preliminary study of the consequences of intonation for the voice source , *Quarterly Progress and Status Report*, Speech Transmission Laboratory 1989, no 4, pp 23-36

Pinto, N B , Childers, D G and Lalwani, A J (1989) "Formant speech synthesis improving production quality", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol 37, no 12, pp 1870-1887

Press, Flannery, Teukolsky and Vutterling (1986) *Numerical Recipes - The Art of Scientific Computation*, New York, Cambridge University Press

Price, P J (1989) "Male and female voice source characteristics inverse filtering results", *Speech Communication*, vol 8, pp 261-277

Qi, Y and Bi, N (1994) "A simplified approximation of the four-parameter LF model of voice source", *Journal of the Acoustical Society of America*, vol 96, no 2, pt 1, pp 1182-1185

Quackenbush, S R , Barnwell, T P and Clements, M A (1988) *Objective Measures of Speech Quality*, Englewood Cliffs, New Jersey, Prentice Hall

Rabiner, L R (1994) "Applications of voice processing to telecommunications", *Proceedings of the IEEE*, vol 82, no 2, pp 197-228

Rabiner, L R and Schafer, R W (1978) *Digital Processing of Speech Signals*, Englewood Cliffs, New Jersey, Prentice Hall

Riegelsberger, E L and Krishnamurthy, A K (1993) "Glottal source estimation methods of applying the LF-model to inverse filtering", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1993*, vol 2, pp 542-545

Rosenberg, A E (1971) "Effects of glottal pulse shape on the quality of natural vowels", *Journal of the Acoustical Society of America*, vol 49, no 2, pt 2, pp 583-590

Ross, M J , Shaffer, H L , Cohen, A , Freudberg, R and Manley, H J (1974) "Average magnitude difference function pitch extractor" *IEEE Transactions on Acoustics Speech and Signal Processing*, vol 22, pp 353-361

Rothenberg, M (1970) "New inverse filtering technique for deriving the glottal air flow waveform during voicing", *Journal of the Acoustical Society of America*, vol 48, no 1, pt 1 pp 130 (Paper I12, 79th Meeting ASA)

Rothenberg, M (1973) "A new inverse-filtering technique for deriving the glottal air flow during voicing", *Journal of the Acoustical Society of America*, vol 53, no 6, pp 1632-1645

Rothenberg, M (1981) "An interactive model for the voice source', *Quarterly Progress and Status Report*, Speech Transmission Laboratory, 1981, no 4, pp 1-17

Rothenburg, M and Zahorian S (1977) "Nonlinear inverse filtering technique for estimating the glottal-area waveform", *Journal of the Acoustical Society of America*, vol 61 no 4, pp 1063-1071

Salava T (1988) "Acoustic load and transfer functions in rooms at low frequencies" *Journal of the Audio Engineering Society*, vol 36, no 10, pp 763-775

Schoentgen, J (1988) "Non-linear modelling of the glottal waveform", *7th FASE Symposium, Proceedings of Speech 1988*, vol 1, pp 211-216

Schoentgen, J (1989) "The spectral dynamics of a non-linear model of the glottal waveform", *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH) 1989*, vol 2, pp 481-484

Schoentgen, J (1990) "Non-linear signal representation and its application to the modelling of the glottal waveform", *Speech Communication*, vol 9, no 3, pp 189-201

Schoentgen, J (1992a) "Glottal waveform synthesis with Volterra shaping functions", *Speech Communication*, vol 11, pp 499-512

Schoentgen, J (1992b) "Nonlinear synthesis of the glottal waveform", *Proceedings of the European Signal Processing Conference (EUSIPCO) 1992*, vol 1, pp 335-338

Schroeder, M R (1954) "Statistical parameters of the frequency response curves of large rooms", *Acustica*, vol 4, pp 594-600 Also appears in (1987) *Journal of the Audio Engineering Society*, vol 35, no 5, pp 299-305

Schroeder, M R (1962) "Frequency-correlation functions of frequency responses in rooms", *Journal of the Acoustical Society of America*, vol 34, pp 1819-1823

Schroeder, M R (1965) "New method for measuring reverberation time", *Journal of the Acoustical Society of America*, vol 37, pp 409-412

Schroeder, M R (1979) "Integrated-impulse method measuring sound decay without using impulses", *Journal of the Acoustical Society of America*, vol 66, pp 497-500

Schroeder, M R and Atal, B S (1985) "Code-Excited Linear Prediction (CELP) High-quality speech at very low bit rates", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1985*, pp 937-940

Schroeder, M R and Kuttruff, K H (1962) "On frequency response curves in rooms Comparison of experimental, theoretical, and Monte Carlo results for the average frequency spacing between maxima', *Journal of the Acoustical Society of America*, vol 36, no 1, pp 76-80

Schroeter, J , Larar, J N and Sondhi, M M (1987) "Speech parameter estimation using a vocal tract/cord model', *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) 1987*, pp 308-311

Scordilis, M and Gowdy, J N (1990) "Effects of the vocal tract shape on the spectral tilt of the glottal pulse waveform", *Proceedings of the IEEE Southeastcon 1990*, pp 86-89

Shoham, Y (1993a) 'High-quality speech coding at 2 4 to 4 0 kbps based on time-frequency interpolation' *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) 1993*, vol 2, pp 167-170

Shoham, Y (1993b) "High-quality speech coding at 2 4 kbps based on time-frequency interpolation *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH) 1993*, vol 2, pp 741-744

Smith C P (1954) "Device for extracting the excitation function from speech signals" U S Patent no 2691137, issued Oct 1954, filed June 1952, reissued 1956

Sobakin, A N (1972) "Digital computer determination of the formant parameters of the vocal tract from a speech signal", *Soviet Physics - Acoustics*, vol 18, pp 84-90

Sondhi, M M (1974) "Model for wave propagation in a lossy vocal tract", *Journal of the Acoustical Society of America*, vol 55, pp 1070-1075

Sondhi, M M (1975) "Measurement of the glottal waveform", *Journal of the Acoustical Society of America*, vol 57, no 1 pp 228-232

Stanton, B J Jamieson, L H and Allen, G D (1989) "Robust recognition of loud and Lombard speech in the fighter cockpit environment", *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) 1989*, pp 675-678

Steiglitz, K and Dickinson, B (1977) "The use of time-domain selection for improved linear prediction", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol 25, no 1, pp 34-39

Stevens, K N and House, A S (1955) "Development of a quantitative description of vowel articulation" *Journal of the Acoustical Society of America*, vol 27, pp 484-493

Strik, H and Boves, L (1992) "On the relation between voice source parameters and prosodic features in connected speech", *Speech Communication*, vol 11, pp 167-174

Strube, H W (1974) "Determination of the instant of glottal closure from the speech wave", *Journal of the Acoustical Society of America*, vol 56, no 5, pp 1625-1629

Strube, H W (1980) "Comments on 'Least-squares glottal inverse filtering from the acoustic speech waveform' ", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol 28, no 3, pp 343

Sugamura, N and Itakura, F (1981) "Speech data compression by LSP analysis-synthesis technique", *Transactions of the Institute of Electronics, Information and Computer Engineers*, vol J64-A, pp 599-606

Takasugi, T (1971) "'Analysis-by-synthesis' method utilizing spectral features of voice source, and measurement of glottal waveform parameters", *Journal of the Radio Research Laboratory*, vol 18, no 97, pp 209-220

Teagar, H M and Teagar, S M (1990) "Evidence for nonlinear production mechanisms in the vocal tract" in *Proceedings of the NATO Advanced Study Institute*, Norwell, Mass , Klumer

Thomson, M M (1992) "A new method for determining the vocal tract transfer function and its excitation from voiced speech", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1992*, vol 2, pp 37-40

Titze, I.R (1984) "Parameterisation of the glottal area, glottal flow and vocal fold contact areas", *Journal of the Acoustical Society of America*, vol 75, pp 570-580

Titze, I R (1989) "A four-parameter model of the glottis and vocal fold contact area" *Speech Communication*, vol 8, pp 191-201

Tremain, T E (1982) "The government standard linear predictive coding algorithm LPC-10", *Speech Technology*, vol 1, no 2, pp 40-49

van den Berg, J (1958) "Myoelastic-aerodynamic theory of voice production", *Journal of Speech and Hearing Research*, vol 1, pp 227-244 van den Berg, J , Zantema, J T and Doornenbal, P (1957) "On the air resistance and the Bernoulli effect of the human larynx", *Journal of the Acoustical Society of America*, vol 29, pp 626-631

Veeneman, D E and BeMent, S L (1984) "Automatic glottal inverse filtering", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1984*, pp 36 5 1-36 5 4

Veeneman, D E and BeMent, S L (1985) "Automatic glottal inverse filtering of speech from speech and electroglottographic signals", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol 33, no 2, pp 369-377

Wang, S , Sekey, A and Gersho, A (1992) "An objective measure for predicting subjective quality of speech coders", *IEEE Journal on Selected Areas in Communications*, vol 10, no 5, pp 819-829

Wattel, E , Pomp, R , van Rietschote, H F and Steeneken, H J M (1981) "Predicting speech intelligibility in rooms from the modulation transfer function III Mirror image computer model applied to pyramidal rooms", *Acustica*, vol 48, pp 320-324

Wong, D Y , Markel, J D and Gray, A H (1979) "Least squares glottal inverse filtering from the acoustic speech waveform", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol 27, no 4, pp 350-355

Wong D Y and Markel, J D (1978) "An excitation function for LPC synthesis which retains the human glottal phase characteristics", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1978*, pp 171-174

Yaggi, L A (1962) "Full duplex digital vocoder Scientific Report no 1", Texas Instruments Inc , Dallas TX, rept no SP14-A62

Yegnanarayana, B (1981) "Speech analysis by pole-zero decomposition of short-time spectra", *Signal Processing*, vol 3, pp 5-17

Young, T Y (1965) "Epoch detection - a method for resolving overlapping signals", *Bell System Technical Journal*, vol 44, pp 401-426

APPENDIX A

THE VARIATION OF THE LIP RADIATION IMPEDANCE IN A REVERBERANT ENCLOSURE

by
C.J. Bleakley and R. Scaife

Originally published in the
Proceedings of the European Signal Processing Conference
(EUSIPCO) 1994, vol. 3, pp. 1689-1692.

The Variation of the Lip Radiation Impedance in a Reverberant Enclosure

Chris J. BLEAKLEY and Ronan SCAIFE

School of Electronic Engineering, Dublin City University, Glasnevin, Dublin 9, Ireland,
Tel: +353-1-7045434, Fax: +353-1-7045508, E-Mail: bleakc@vax1.dcu.ie, scaifer@eeng.dcu.ie

Abstract. Extraction of the glottal waveform from voiced speech shows promise in improving speech coding and recognition systems. However, under normally reverberant conditions the glottal signal becomes randomised. This paper describes an experimental investigation into the effects of reverberation on estimation of the glottal waveform. The radiation and transmission impedance of a speaker were measured under free field and reverberant conditions. Their variation was then estimated and applied to a model of the human speech production system. The degree of randomisation of the glottal waveform due to reverberation was then determined by computer simulation. The simulations show that the error energy to signal energy ratio for glottal waveform extraction falls from almost free field values to 15-20dB less as the source-receiver separation is increased from zero to the reverberation distance. This suggests that glottal waveform extraction is always possible proved that a microphone can be placed arbitrarily close of the lips.

1. Introduction

In experiments on non-reverberant speech, it has been shown that extraction of the glottal waveform can improve the performance of low bit rate speech coders [1] and speech recognition systems [2]. Thus, one of the goals of current research is to find methods for glottal waveform extraction that are effective in both reverberant and non-reverberant conditions.

The major obstacle for glottal waveform extraction under reverberant conditions is the complex nature of the acoustic load seen looking outwards from the lips into the enclosure. There are two related effects - firstly, a non-smooth transfer function between the lip flow signal and the pressure at a microphone and, secondly, an unknown perturbation of the vocal tract resonances.

This paper describes experiments which are designed to determine the conditions under which extraction of the glottal waveform is possible without the use of adaptive filtering or echo cancellation. A realistic speech production model for speech generation and an inverse filter for glottal waveform extraction were implemented in Matlab on a PC. Actual reverberant impedance functions were then measured and applied to the model in order to assess the effect of room reflections on the accuracy of glottal waveform extraction by inverse filtering.

2. Room Acoustic Measurements

2.1 Impedance Functions

The performance of loudspeakers is generally characterised by their acoustic impedance. It is defined as the ratio of the pressure generated by the speaker to the volume velocity of the speaker. The acoustic impedance measurement method of Salava [4] was chosen for these experiments since it is both cost effective and accurate. Flow is measured with an inverted loudspeaker which is acoustically coupled to an identical driver unit, see Figure 1. The driver unit is driven by a pseudo random sequence and the inverted, or passive, cone moves in sympathy. This generates an e.m.f. in its speaker coil which is directly proportional to the volume velocity of air displaced. In the case of radiation impedance measurements, pressure is measured by a microphone fixed close to the back of the passive speaker. When measuring transmission impedance, the microphone is fixed some distance from the speaker.

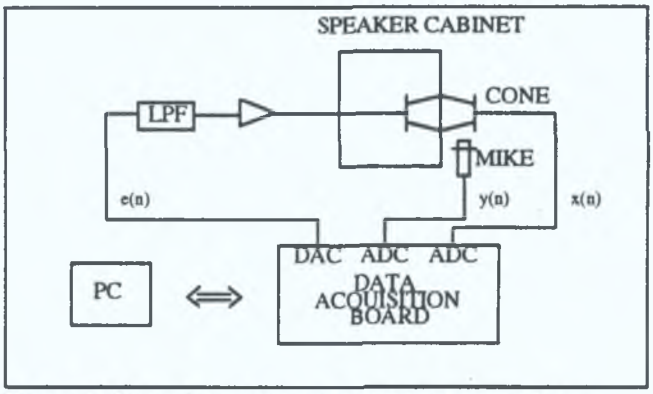


Figure 1. Room impedance measurement apparatus.

The experiment was controlled by a PC using an Loughborough Sound Images development card with a Texas Instruments TMS320C30 and on-board ADC/DAC. The driver speaker was excited with a maximal length sequence (length 32767) at a sampling frequency of 16kHz. This excitation signal was anti-aliased using a passive 5kHz low pass filter and amplified with a JVC AX-11 amplifier. The speakers, Radionics 8ohm 6.5in. bass/mid-range units, were installed in a 30cm by 20cm by 13cm wooden speaker cabinet which was lined with sound absorbing foam. The acoustic coupling between the speakers was stiffened by partially filling the air gap between the cones. The pressure signal was measured using a B&K microphone (4006) with diffusion cap and, like the passive cone signal, was amplified using a Alice Soundtek pre-amplifier. For each measurement 50 records of the pressure and flow signals were obtained. The impedance functions were calculated using the cross-spectral technique [5].

The free field impedance of the speaker was estimated by measuring the impedance at certain fixed on-axis source-receiver distances (0cm, 3cm, 7cm, 13cm, 32cm, 62cm) at four different locations in a hemi-aneochic chamber. The spatially averaged impedance, which tends to the free field value [6], was calculated for each source-receiver distance. The chamber is well damped, with the walls covered in acoustic wadding and mesh behind heavy curtains. The chamber measures 3.0m by 3.0m by 2.7m.

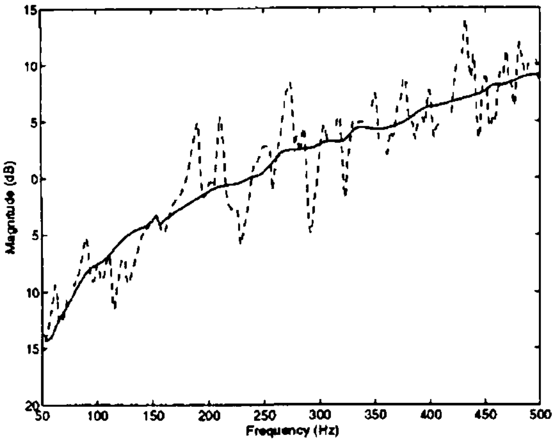


Figure 2 Measured transmission impedance for source-receiver distance of 32cm (solid line - spatially averaged hemi-anechoic, dashed line - in room speaker position 1)

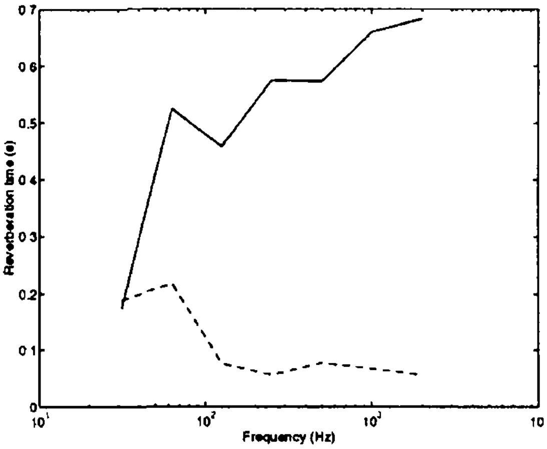


Figure 3 Measured reverberation times in third octave bands (solid line - room, dashed line - hemi-anechoic chamber)

The reverberant measurements were performed in a similar manner to the free field measurements but without spatial averaging The room used for the reverberant measurements was 3 4m by 2 6m by 2 7m with smooth plastered walls a concrete floor and no windows

A example of the measured free field and reverberant transmission impedances can be seen in Figure 2

2 2 Reverberation Measurement

In order to relate the impedance fluctuation measurements to the acoustics of the rooms under investigation the reverberation times of the rooms were determined using the integrated impulse response method of Schroeder [3] The reverberation time of a room is defined as the time interval during which the reverberant sound field drops by 60dB Schroeders method involves exciting the room with a maximal length sequence measuring the response in the reverberant field and calculating the impulse response by circular cross-correlation The average sound decay can then be calculated as the integral of the time reversed impulse response squared The reverberation time was estimated by fitting a straight line to a manually selected portion of the early sound decay

Again the measurements were taken using a PC and LSI board The excitation signal (length 32767) was emitted by a Fostex 6301B loudspeaker and recorded using the same microphone and pre-amplifier The reverberation time was measured at six receiver positions for each of four source locations and averaged The impulse responses were split into third octave bands using digital third-order Butterworth filters The results are shown in Figure 3

An associated reverberation measure is the reverberation distance [3] This is defined as the source receiver distance at which the direct energy density is equal to the reverberant energy density The reverberation distances have been calculated as 0 9m in the hemi-anechoic chamber and as 0 4m in the room

3 Reverberant Speech Modelling

3 1 Speech Production Model

The speech production model of Kelly and Lochbaum [7], sampled at a frequency of 16kHz, was used to synthesise speech This lossless tube model was augmented with glottal termination lip termination and wall loss models The glottal termination model was based on that of Badin and Fant [8] The wall loss model followed the proposal of Liljencrants [9] For simplicity, these models were lumped together at the glottis and represented by fitting a 2 pole-2 zero filter [10] For the lip termination the Laine digital approximation to the piston-in-a-spherical-baffle model was selected [11] The parameters of the model were determined by table look-up and interpolation from the lip opening area [10]

The glottal flow waveform was modelled using the LF time-domain model [12] This models the differentiated glottal waveform as a sinusoid with an exponential trailing edge The glottal flow itself is then calculated by integration

The combined model was tested using realistic LF-model and area function parameters The vowel sounds and speech spectra generated by the model were found to be consistent with real speech

3 2 Adding Reverberation

Since the free field radiation impedance at the lips and the speaker are very different, the acoustic impedance measurements must be scaled for application to the speech production model Two functions must be calculated - the variation in the radiation impedance which will effect the resonances in the vocal tract and thus the flow at the lips and the variation in the transmission impedance which will effect the pressure signal actually received by a microphone some distance from the lips Both functions can be represented by the ratio between the measured free field impedance and the reverberant impedance

$$T_{rad} = \frac{Z_{reverb\ rad}}{Z_{ff\ rad}} \qquad T_{trans} = \frac{Z_{reverb\ trans}}{Z_{ff\ trans}} \qquad (1)$$

where T_{rad} and T_{trans} are the radiation and transmission ratios $Z_{ff\ rad}$ and $Z_{ff\ trans}$ are the free field radiation and transmission impedances and $Z_{reverb\ rad}$ and $Z_{reverb\ trans}$ are the reverberant radiation and transmission impedances respectively

For application to the speech production model the impedance ratios must be converted into z-domain models This was done by calculating the inverse Fourier Transforms of the impedance ratios and truncating the resulting impulse responses to obtain FIR filters It was found that for the measurements under investigation an FIR filter of 8192 taps (0 5s) was adequate for modelling the measured impedance variations

For a given source location, the radiation ratio filter was applied to Laine's free field lip radiation impedance model to obtain the reverberant lip radiation impedance. The reflection coefficient at the lips was recalculated using the new lip impedance function. The flow from the lips was then calculated from the glottal flow signal and area function.

$$Z_{\text{lip reverb rad}} = T_{\text{reverb rad}} Z_{\text{lips ff rad}} \qquad (2)$$

where $Z_{\text{lip reverb rad}}$ and $Z_{\text{lip ff rad}}$ are the reverberant and free field lip radiation impedances respectively.

Similarly, the transmission ratio filter was applied to Laine's free field lip radiation impedance model to obtain the reverberant transmission impedance. The reverberant pressure signal at the microphone was then obtained by applying this filter to the flow signal at the lips.

$$Z_{\text{lip reverb trans}} = T_{\text{reverb trans}} Z_{\text{lips ff rad}} \qquad (3)$$

where $Z_{\text{lip reverb trans}}$ is the reverberant transmission impedance.

It must be noted that the calculation of the transmission impedance ratio assumes that the frequency dependent nature of the radiation pattern in front of the speaker assembly is the same at that in front of the lips. Obviously, this is not the case. However, for the wavelengths in question, which are large compared with source size, both sources can be considered as approximating spherical sources.

3.3 Inverse Filtering

In order to determine the effects of the impedance variation on glottal waveform extraction, inverse filtering was performed on the reverberant pressure signal. An inverse filter was constructed for the free field speech production model and was applied to the reverberant pressure signal to estimate the glottal flow. The differentiated glottal flow was then calculated as the difference signal of the recovered glottal flow. This estimated differentiated flow was then band limited using a thirty point FIR filter and decimated to 4kHz, so as to exclude frequencies at which the impedance measurements were unreliable. This signal was then scaled and compared to the similarly band limited and decimated LF signal, see Figure 4. The accuracy of the inverse filtering was quantified by calculating the ratio of the energy of the LF signal to the energy of the error between it and the estimated differentiated glottal flow (signal to error ratio, SER).

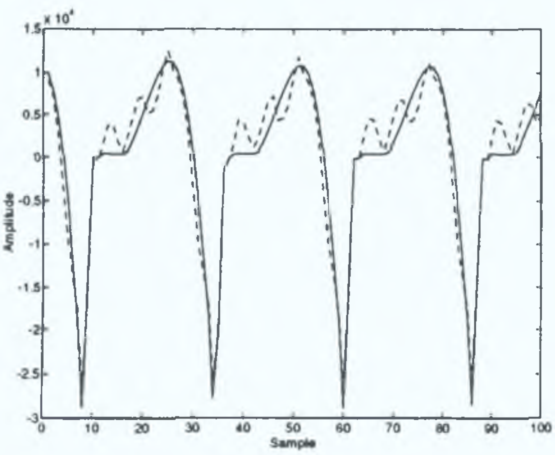


Figure 4. True and extracted differentiated glottal waveforms (solid line - LF signal; dashed line - extracted differentiated glottal flow for vowel /a/ after simulation in room position 1)

4. Results

The measured room acoustic impedance variations were applied to synthetic speech and inverse filtered as described above. Figure 5 shows the variation of the extracted glottal waveform SER with source-receiver distance for different vowels in two room locations. Clearly, extraction of the glottal waveform becomes less accurate with increasing source-receiver distance.

Comparison of Figure 5 (a), (b) with (c), (d) indicates that the accuracy is similar for different vowels generated in the same location. The randomisation effect appears largely independent of the vocal tract configuration. This is confirmed by the high accuracy of inverse filtering applied to the pressure signal at the lips, see Table 1. Reverberation has little effect on the resonances in the vocal tract itself.

Vowel	Time of estimation (s)		
	0.03	0.23	0.45
/a/	42.27	41.97	41.60
/i/	47.42	47.53	47.35

Table 1. Free field SER (dB) for vowels /a/ and /i/ measured after different durations of phonation.

Contrasting Figure 5 (a), (c) with (b), (d) shows a large discrepancy in the SER for the same vowel recorded at the same source-receiver distance but at different points in the same room. It is instructive to note that position 1 is close to the centre of the room, while position 2 is close to a wall. It would seem that the larger number of modes excited in position 2 produces a greater reverberant sound field which interferes more strongly with direct sound transmission and so reduces the accuracy of inverse filtering.

Examining the results obtained for the same vowel in the same position but extracted after different durations of speech production indicates the build up of the interfering reverberant field with time. As the reverberant sound field increases in intensity, so the accuracy of glottal waveform extraction reduces, resulting in lower SERs for the waveforms extracted after longer durations of phonation.

5. Conclusions

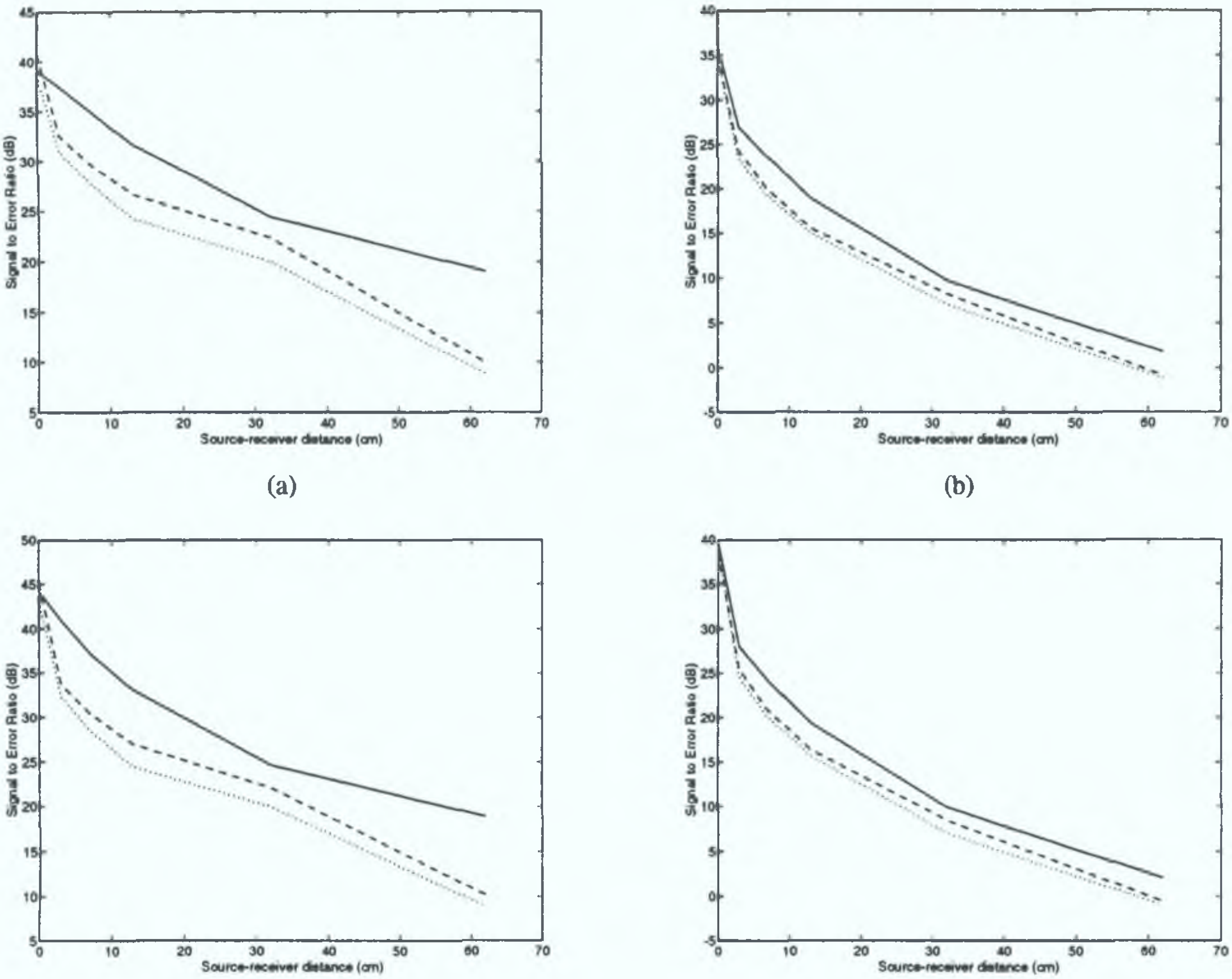
These results suggest that retrieval of the glottal waveform is always possible, provided that the microphone can be placed arbitrarily close to the speaker. However, in a hands-free or distant microphone situations it would appear that glottal waveform recovery is impossible without the use of some reverberation compensation algorithm.

It should be noted that the inverse filtering accuracies obtained in this investigation are probably unobtainable for normally recorded speech. Under normal conditions the accuracy of the technique is severely reduced by inaccurate inverse filter identification and by the addition of noise.

The procedures outlined above provide a test-bed for the analysis of in-room glottal waveform extraction methods. Work is continuing to investigate reverberant effects and to assess the effectiveness of dereverberation algorithms in determining the glottal waveform.

Acknowledgements

One of the authors (Chris J. Bleakley) is in receipt of a PhD studentship from Forbairt, the Irish Science and Technology Agency under the Programme of Advanced Technology.



(a)

(b)

(c)

(d)

Figure 5. Variation of estimated differentiated glottal flow SER (dB) with source-receiver distance for: (a) vowel /a/ at position 1; (b) vowel /a/ at position 2; (c) vowel /i/ at position 1; (d) vowel /i/ at position 2. (solid line - measured after 0.03s; dashed line - measured after 0.23s; dotted line - measured after 0.45s)

References

[1] **A. Bergstrom and P. Hedelin (1989):** Code-book driven glottal pulse analysis, *Proc. Int. Conf. ICASSP-89, 53-56.*

[2] **M. Blomberg (1991):** Adaptation to a speaker's voice in a speech recognition system based on synthetic phoneme references, *Speech Comm. 10(5-6),* 453-456.

[3] **H. Kuttruff (1991):** Room Acoustics, *Elsevier Applied Science.*

[4] **D.K. Anthony and S.J. Elliott (1991):** A comparison of three methods of measuring the volume velocity of an acoustic source, *J. Audio Eng. Soc. 39(5),* 355-366.

[5] **J.S. Bendat and A.G. Piersol (1971):** Random Data: Analysis and Measurement Procedures, *John Wiley & Sons.*

[6] **J.L. Davy (1981):** The relative variance of the transmission function of a reverberation room, *J. Sound and Vibration 77*(4), 455-479.

[7] **J.L. Kelly, Jr. and C. Lochbaum (1962):** Speech synthesis, *Proc. Stockholm Speech Comm. Seminar,* R.I.T., Stockholm, Sweden, Sept. 1962.

[8] **P. Badin and G. Fant (1984):** Notes of vocal tract computation, *STL-QPSR '84(2-3),* 53-109.

[9] **J. Liljencrants (1985):** Speech Synthesis with a Reflection-type Line Analog, *PhD Thesis,* R.I.T., Stockholm, Sweden.

[10] **R. Scaife (1989):** Vocal tract estimation - Extending the Wakita inverse filter, *Proc. Int. Conf. EUROSPEECH-89, vol II, 648-651.*

[11] **U.K. Laine (1982):** Modelling of lip radiation impedance, *Proc. Int. Conf. ICASSP-82, 1992-1995.*

[12] **G. Fant, J. Liljencrants, and Q. Lin (1985):** A four-parameter model of glottal flow, *STL-QPSR '85(4),* 1-13.

APPENDIX B

NEW FORMULAS FOR PREDICTING THE ACCURACY OF ACOUSTICAL MEASUREMENTS MADE IN NOISY ENVIRONMENTS USING THE AVERAGED *M*-SEQUENCE CORRELATION TECHNIQUE

by
C.J. Bleakley and R. Scaife

Originally published in the
Journal of the Acoustical Society of America,
vol. 97, no. 2, pp. 1329-1332.

New formulas for predicting the accuracy of acoustical measurements made in noisy environments using the averaged *m*-sequence correlation technique

Chris Bleakley and Ronan Scaife
School of Electronic Engineering, Dublin City University, Glasnevin, Dublin 9, Ireland

(Received 4 July 1994; accepted for publication 23 September 1994)

The averaged *m*-sequence correlation technique has been established as a means of measuring linear system responses under high noise conditions. This Letter examines the theoretical basis for the technique and derives an analytical formula for the expected error in the estimated system response under noisy conditions at the system input and/or output. The formula shows close agreement with previously published experimental results [W. Zuomin and W. T. Chu, J. Acoust. Soc. Am. **94**, 1409–1414 (1993)] and with newly derived simulation results.

PACS numbers: 43.55.Mc, 43.60.Qv

INTRODUCTION

Originally proposed for architectural acoustic measurements by Schroeder,[1] the *m*-sequence technique has proved useful in determining system responses in noisy environments. The method involves exciting the system with an *m*-sequence and measuring the output. The system's impulse response can then be calculated as the circular cross correlation of the input and output sequences. The accuracy of the technique can be improved by averaging the measurement over several cycles of the *m*-sequence.

In order to determine the improvement in accuracy provided by the averaging process, Zuomin and Chu[2] tested the technique under varying degrees of measurement noise and with various numbers of cycles. After analyzing their experimental results, Zuomin and Chu fitted an empirical formula to the data. This formula predicted the accuracy of the impulse response estimate based on the signal-to noise (S/N) ratio of an individual measurement and the number of cycles over which the results were averaged. This formula provides a means of choosing the number of cycles required to obtain a given accuracy in estimating the impulse response.

In this letter the theoretical aspects of the *m*-sequence method are analyzed and an analytic formula, giving the accuracy of the impulse response estimate based on the S/N ratio and the number of cycles, is derived. The new formula very closely matches the experimental results of Zuomin and Chu but differs from their empirical formula. In addition, the theory covers the general case of input noise and/or measurement noise occurring during the measurements. Simulation results, which confirm the improved accuracy of the new formula, are provided.

I. NOISELESS CASE

Consider the discrete time system depicted in Fig. 1. The sampled impulse response $h(k)$ of a linear time invariant system is to be determined by the *m*-sequence method. The *m* sequence $x(n)$ is used to excite the system and the output $y(n)$ is recorded. The system is assumed to be causal and the impulse response $h(k)$ is required to be shorter than the excitation sequence $x(n)$, where the length N of the *m*-sequence equals 2^m-1. Under noiseless conditions [$a_1(n)=0$ and $a_2(n)=0$] an estimate $h'(k)$ of the impulse response of the system can be obtained by calculating the circular cross correlation of the input and output sequences (based on Ref. 1)

$$h'(k)=\frac{1}{N+1}\sum_{n=0}^{N-1} y(n)x(\overline{\overline{n-k}}),\quad \text{where } \bar{\bar{q}}\equiv q \bmod N \tag{1}$$

It can be shown that the circular autocorrelation of an *m*-sequence is given by[3]

$$\sum_{n=0}^{N-1} x(n)x(\overline{\overline{n-k}})=\begin{cases} N, & \text{if } (k \bmod N)=0,\\ -1, & \text{otherwise} \end{cases} \tag{2}$$

Since the output of the system is equal to the convolution of the impulse response and the input sequence, we can rearrange and substitute to obtain

$$h'(k)=\frac{N+1}{N+1}\,h(k)-\frac{1}{N+1}\sum_{n=0}^{N-1} h(n)\approx h(k) \tag{3}$$

That is, under noiseless conditions the estimated impulse response is approximately equal to the true impulse response.

II. NOISY CASE

Now consider the general case with additive input and measurement noise. We assume that the noise sequences are stationary, zero-mean, and Gaussian.

The addition of the input noise $a_1(n)$ and the measurement noise $a_2(n)$ leads to a new output sequence $y'(n)$. If we attempt to estimate the impulse response of the system using the circular cross-correlation method

$$h'(k)=\frac{1}{N+1}\sum_{n=0}^{N-1} y'(n)x(\overline{\overline{n-k}}) \tag{4}$$

The new output sequence can then be expressed as measurement noise added to the convolution of the true impulse re

0001-4966/95/97(2)/1329/4/$6.00

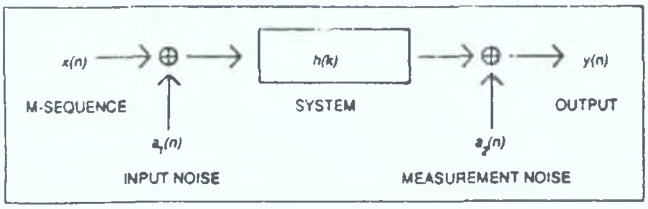


FIG. 1. Measurement system.

sponse with the input sequence plus input noise. Replacing and rearranging we obtain:

$$h'(k)=\frac{1}{N+1}\sum_{n=0}^{N-1}\sum_{l=0}^{N-1} h(l)x(\overline{n-l})x(\overline{n-k}) \tag{5a}$$

$$+\frac{1}{N+1}\sum_{n=0}^{N-1}\sum_{l=0}^{N-1} h(l)a_1(\overline{n-l})x(\overline{n-k}) \tag{5b}$$

$$+\frac{1}{N+1}\sum_{n=0}^{N-1} a_2(n)x(\overline{n-k}). \tag{5c}$$

Examining each term in turn:

(a) As in the noiseless case, this term is equal to the true impulse response $h(k)$.

(b) Consider:

$$b_1(k)=\frac{1}{N+1}\sum_{l=0}^{N-1} h(l)\left[\sum_{n=0}^{N-1} a_1(\overline{n-l})x(\overline{n-k})\right]. \tag{6}$$

If we examine the bracketed expression we can see that it is the sum of N Gaussian variables, $(N-1)/2$ of which have been multiplied by $+1$ and $(N+1)/2$ of which have been multiplied by -1. It can be shown that the weighted sum of any number of independent Gaussian variables is a Gaussian variable itself.[4] The mean of the sum is equal to the weighted sum of the means of the constituent variables. The variance of the sum is equal to the sum of the variances of the constituent variables each multiplied by its weight squared. Thus the bracketed expression in Eq. (6) reduces to a zero-mean, Gaussian sequence with

$$\operatorname{var}\left(\left[\sum_{n=0}^{N-1} a_1(\overline{n-l})x(\overline{n-k})\right]\right)=N\operatorname{var}(a_1). \tag{7}$$

Note that, for clarity in variance expressions, the function arguments have been omitted.

Again, applying the linear combination property of Gaussian variables, we find that $b_1(k)$ itself is a zero-mean, Gaussian sequence with variance:

$$\operatorname{var}(b_1)\approx\operatorname{var}(h)\operatorname{var}(a_1). \tag{8}$$

Note that the variance of the impulse response is calculated over the length of the m-sequence.

(c) Consider:

$$b_2(k)=\frac{1}{N+1}\sum_{n=0}^{N-1} a_2(n)x(\overline{n-k}). \tag{9}$$

As before, the sequence $b_2(k)$ is zero mean and Gaussian:

$$\operatorname{var}(b_2)\approx(1/N)\operatorname{var}(a_2). \tag{10}$$

Overall, we have

$$h'(k)\approx h(k)+b_1(k)+b_2(k), \tag{11}$$

where $b_1(k)$ and $b_2(k)$ are independent, zero-mean, Gaussian random sequences and

$$E(h'(k))\approx h(k), \tag{12}$$

$$\operatorname{var}(h'(k)-h(k))\approx\operatorname{var}(h)\operatorname{var}(a_1)+(1/N)\operatorname{var}(a_2). \tag{13}$$

Thus the estimation procedure provides, for large N, an approximately unbiased estimate of the true impulse response. In addition, as $N\rightarrow\infty$, so $\operatorname{var}(h'(k)-h(k))\rightarrow 0$ and the estimate is consistent.

III. AVERAGING

In order to improve the accuracy of the estimate $h'(k)$ it is desirable to reduce the variances of $b_1(k)$ and $b_2(k)$. This may be done by repeating the measurement and averaging the results over M cycles. Since circular cross correlation is a linear operation, this is equivalent to averaging the estimated impulse response:

$$\langle h'(k)\rangle\approx h(k)+\frac{1}{M}\sum_{i=1}^{M} b_{1,i}(k)+\frac{1}{M}\sum_{i=1}^{M} b_{2,i}(k). \tag{14}$$

Assuming the noise sequences are independent:

$$E(\langle h'(k)\rangle)\approx h(k), \tag{15}$$

$$\operatorname{var}(\langle h'(k)\rangle-h(k))\approx\frac{1}{M}\operatorname{var}(h)\operatorname{var}(a_1)+\frac{1}{MN}\operatorname{var}(a_2). \tag{16}$$

So the averaging procedure provides an approximately unbiased estimate of the true impulse response. Also, as $M\rightarrow\infty$ then $\operatorname{var}(\langle h'(k)\rangle-h(k))\rightarrow 0$, and the averaged estimation procedure is also consistent.

IV. SIGNAL TO NOISE RATIO

We can assess the effects of the m-sequence averaging technique by examining its influence on the signal-to-noise ratio of the estimated impulse response. We shall make the assumption that the input noise is negligible in comparison to the measurement noise, a circumstance which is often true in practical situations.

The signal-to-noise ratio of the single impulse response method, in dB, is given by

$$\xi_I=10\log_{10}(\operatorname{var}(h)/\operatorname{var}(a_2)). \tag{17}$$

From Eq. (16), assuming $\operatorname{var}(a_1)\approx 0$, the signal-to-noise ratio of the estimate calculated by the averaged m-sequence method is given by

$$\xi_E=10\log_{10}(\mathrm{var}(h)/\mathrm{var}(\langle h'\rangle-h)) \quad (18a)$$

$$=10\log_{10}(\mathrm{var}(h)/\mathrm{var}(a_2))+10\log_{10}(MN). \quad (18b)$$

The improvement in signal-to-noise ratio can be easily seen by comparing Eqs. (17) and (18b).

V. COMPARISON WITH ZUOMIN AND CHU

Zuomin and Chu analyzed the success of the averaged m-sequence technique in measuring room impulse responses under noisy conditions.[2] They compared the acoustic impulse response measured under noiseless conditions to that estimated under varying intensities of additive pink noise. They then assessed the accuracy of the estimate by calculating the difference in sound pressure level (SPL) between the estimated and "true" impulse responses. Since the logarithmic operation is approximately linear in the region where the argument is close to 1, and since the expected value of the error is zero, we can calculate the expected value of the difference in SPL as:

$$E(L_W-L)=10\log_{10}\left[1+E\left(\sum_{k=0}^{N-1}(\langle h'(k)\rangle-h(k))^2\right)\Bigg/\sum_{k=0}^{N-1}h(k)^2\right]. \quad (19)$$

The expectation term follows a chi-squared distribution with the mean equal to the number of terms in the summation multiplied by the variance of the terms:[4]

$$E(L_W-L)=10\log_{10}(1+\mathrm{var}(\langle h'\rangle-h)/\mathrm{var}(h)) \quad (20a)$$

$$=10\log_{10}\left(1+\frac{1}{M}\mathrm{var}(a_1)+\frac{1}{MN}\frac{\mathrm{var}(a_2)}{\mathrm{var}(h)}\right). \quad (20b)$$

Zuomin and Chu relate these measurements to the signal-to-noise ratio measured at the receiving microphone ξ_l in Eq. (17) (see Zuomin and Chu, Fig. 4). Now, $y(n)$ is equal to the convolution of the m-sequence with the true impulse response. If we rearrange the terms in the convolution and assume that the impulse response samples have a Gaussian distribution then each sample of $y(n)$ can be viewed as the sum of N Gaussian variables, each multiplied by $+1$ or -1. Thus

$$\mathrm{var}(y)=N\,\mathrm{var}(h). \quad (21)$$

Substituting into Eq. (20b) and rearranging, we obtain the new formula:

$$E(L_W-L)=10\log_{10}\left(1+\frac{1}{M}\frac{1}{10^{\xi_l/10}}\right). \quad (22)$$

This equation is analogous to the empirical formula given in Zuomin and Chu:

$$E(L_W-L)=(3.12e^{-0.18\xi_l})M^{-1/1.23}. \quad (23)$$

The predictions of the new formula are plotted along with Zuomin and Chu's experimental results in Fig. 2. The accuracy of the new formula can be clearly seen.

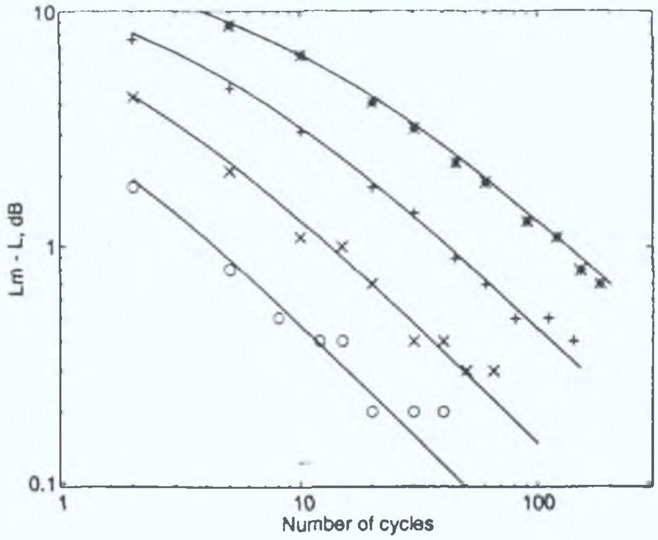


FIG. 2. Comparison of the new formula Eq. (22) (solid line) with Zuomin and Chu's experimental results (points). The figure shows the difference between the recovered and the "true" 1/3-octave-band SPL versus the number of cycles. The system has zero input noise and a measurement S/N ratio of (○) −0.5 dB; (×) −5.4 dB; (+) −10.4 dB; (∗) −15.4 dB.

The percentage difference between the predictions of Zuomin and Chu's formula Eq. (23) and the new formula Eq. (22) is shown in Fig. 3. For values of the S/N ratio and number of cycles close to those used in Zuomin and Chu's experiment the difference between the predictions is negligible. However, for values of the S/N ratio above and below this range the predictions of Zuomin and Chu's formula are much larger than those of the new formula.

In order to determine which formula is the more accurate of the two, the system depicted in Fig. 1 was simulated in MATLAB on an IBM PC. The simulation calculated the true system output by convolving an m-sequence (length 255) with an arbitrary bandpass FIR filter (length 200). Gaussian measurement noise was then added to the true output and the impulse response of the filter estimated by the cross-correlation technique of Eq. (1). The accuracy of each

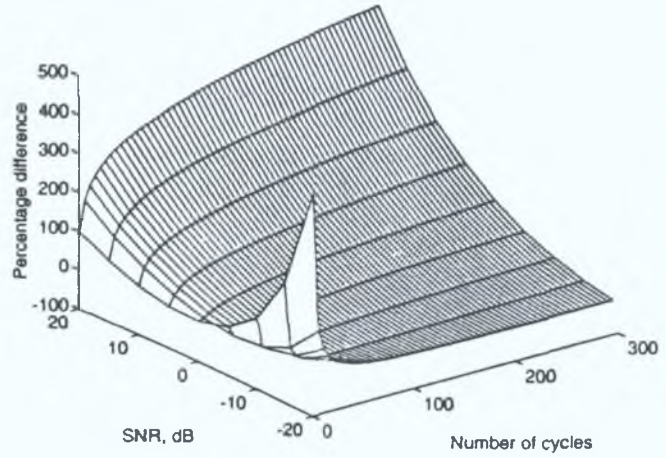


FIG. 3. Percentage difference between predictions of Zuomin and Chu's formula Eq. (23) and the new formula Eq. (22). The figure shows the difference between the recovered and the "true" SPL versus the number of cycles and the measurement S/N ratio, assuming zero input noise.

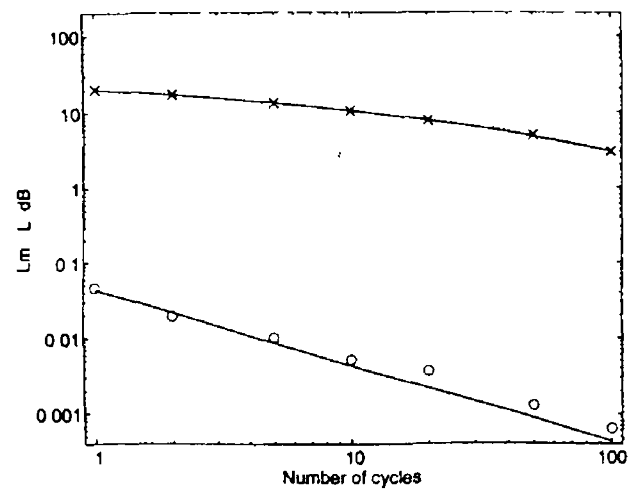


FIG 4 Comparison of simulation results (points) Zuomin s and Chu s formula Eq (23) (dotted line) and the new formula Eq (22) (solid line) The figure shows the difference between the recovered and the true SPL versus the number of cycles The system simulated had zero input noise and a measurement S/N ratio of (O) +20 dB (X) −20 dB The results were averaged over 200 and 20 trials, respectively

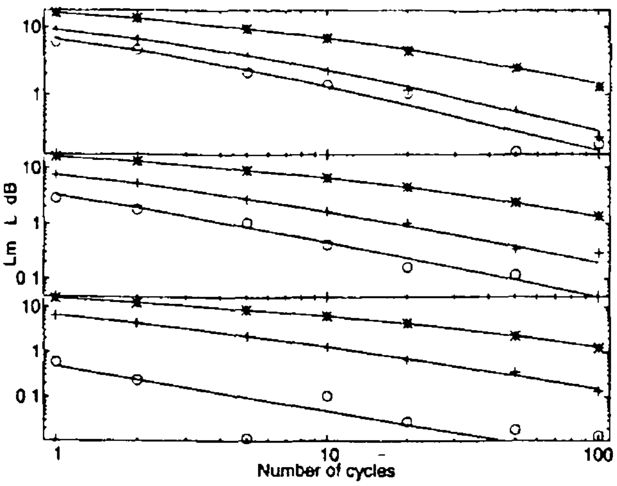


FIG 5 Comparison of simulation results (points) and the new formula Eq (16) (solid line) The figure shows the difference between the recovered and the 'true' SPL versus the number of cycles The system simulated had input noise of (bottom) −0 5 dB, (middle) −10 4 dB, (top) −15 4 dB and a measurement S/N ratio of (O) infinity (+) −5 4 dB, (•) −15 4 dB The results were averaged over 20 trials

estimate was then determined as the difference in SPL between the estimated and "true" impulse response The results were averaged over a number of trials for several combinations of the S/N ratio and number of cycles The simulation results are plotted in Fig 4 along with the predictions of Eqs (22) and (23) Clearly, the new formula Eq (22) is more accurate for these more extreme parameter values In addition, the new formula has the desired property of predicting a zero error for impulse response extraction under noiseless conditions

VI INPUT AND MEASUREMENT NOISE

In order to test the predictions of Eq (16) for both input and measurement noise, the simulations described in Sec V were extended to include input noise as well as measurement noise The results obtained were then plotted in Fig 5, together with the theoretical predictions of Eq (16) This evidence further supports the accuracy of the new formula

VII CONCLUSIONS

An improved formula for predicting the accuracy of the averaged m-sequence method has been derived The accuracy of the formula has been confirmed by the experimental results of Zuomin and Chu and by simulations carried out by the authors The formula is both more general and more accurate than that proposed by Zuomin and Chu This work will provide for more accurate error prediction when using the averaged m-sequence method

ACKNOWLEDGMENTS

One of the authors (CB) is in receipt of a Ph D studentship from the Programme in Advanced Technology in Telecommunications of Forbairt, the Irish Science and Technology agency

[1] M R Schroeder Integrated impulse method measuring sound decay without using impulses, J Acoust Soc Am 66 497–500 (1979)

[2] W Zuomin and W T Chu, Ensemble average requirement for acoustical measurements in noisy environment using m sequence correlation technique, J Acoust Soc Am 94 1409–1414 (1993)

[3] M R Schroeder, *Number Theory in Science and Communication* (Springer Verlag, Berlin, 1986), 2nd ed , Chap 26 pp 274–276

[4] E Lloyd, *Handbook of Applicable Mathematics—Volume II—Probability*, edited by W Ledermann (Wiley, New York, 1980) Chap 11 pp 208–209 and 216

[5] H Alrutz and M R Schroeder A fast Hadamard Transform method for the evaluation of measurements using pseudorandom test signals 11th Int Conf Acoust pp 235–238 (1983)

APPENDIX C

TEST DATA

C 1 INTRODUCTION

Determining the performance of the glottal extraction techniques described in this thesis requires the use of test speech data This appendix describes the generation of that data.

The test data is split into four categories - noiseless, noisy, reverberant and office recordings The noiseless data was recorded in a hemi-anechoic studio Care was taken to ensure that these recordings were of the highest possible fidelity To obtain the noisy data, white noise signals, of varying intensity, were added to the noiseless recordings The reverberant data was generated by filtering the noiseless recordings with simulated room reverberant impulse responses The noisy and reverberant data were generated in this way so that the performance of the glottal extraction algorithms could be tested under controlled levels of distortion The office test data was captured by recording natural speech in a typical office environment The office speech data contained both noise and reverberation at levels typically encountered in speech coding applications In each category, the speech data consisted of a single all-voiced sentence read aloud by a male and a female subject. All of the recordings were made with phase linear recording equipment

The appendix is divided into five sections Section two describes the recording techniques used to capture the noiseless test data Section three details generation of the noisy test data Section four describes the generation of the reverberant data Finally, section five details how the office test data was obtained and section 6 concludes the appendix

C.2 NOISELESS SPEECH DATA

This section describes how noiseless speech data was recorded The sentence "We were away a year ago" was read aloud by a male and a female subject. Both subjects are speakers of British English with Irish dialects The sound was captured using a Bruel and Kjær microphone (model 4006) with an Alice Soundtek pre-amplifier The signal was anti-aliased and sampled at 48 kHz to a 16 bit resolution using a Loughborough Sound Images TMS320C30 development card in an IBM PC The recording equipment was found to have negligible phase and amplitude distortion over the range 20 Hz to 20 kHz The recordings were digitally decimated to a sampling frequency of 8 kHz, using a 8th order FIR low pass filter with cut-off at 3 2 kHz [IEEE, 1979] Additionally, a FIR high pass filter with cut-off at 20 Hz was applied to remove low frequency noise Both filters were passed forwards and backwards across the signal to ensure that they introduced no phase distortion

Two further recordings were made in the same manner using a different male subject. These recordings were designed to capture three types of voicing not included in the original sentence The utterances chosen were "Eva" for the voiced fricative [v] and "who's been" for the high back vowel [u] and the voiced plosive [b]

The recordings were conducted in a studio which was well sound-proofed and acoustically deadened The studio walls were covered in acoustic wadding and heavy curtains, the floor was carpeted and acoustic tiles were fixed to the ceiling The room measured 3 m by 3 m by 2 7 m A lip-microphone distance of approximately 10 cm was chosen, providing comfort for the speakers and low noise in the recordings The signal to noise ratio (SNR) of the recordings was found to be approximately 60 dB and 58 dB for the male and female data, respectively The speech waveform for the first word of the test sentence is shown in the top panels of Fig C 1

C.3 NOISY SPEECH DATA

In order to generate samples of noisy speech, white noise was added to the noiseless recordings Test data with SNRs of 35, 30, 25, 20 and 15 dB were generated Note that the SNRs were determined by calculating the ratio of the energy of the speech signal in the non-silent portions of the recordings to the energy of the added noise Thus, at voicing onsets and offsets the SNR is significantly less than the nominal values given above Fig C 1 panels 2 and 3 show the first part of the test data with added noise giving SNRs of 25 and 15 dB, respectively

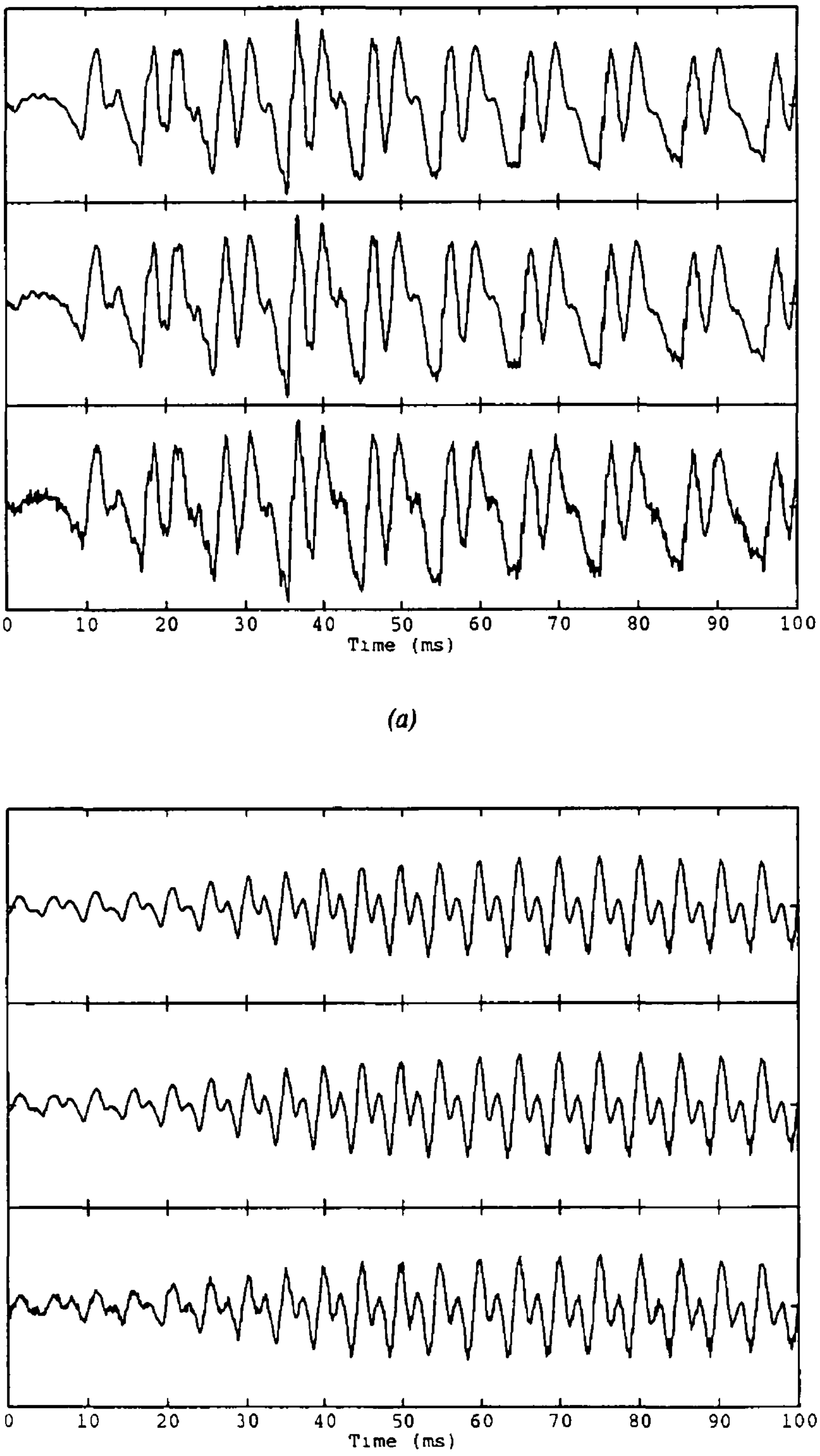
C 4 REVERBERANT SPEECH DATA

The reverberant speech data was generated by convolving the noiseless recordings with simulated room impulse responses The artificial responses were generated using the Image Method [Allen and Berkley, 1979] The accuracy of this procedure was established beforehand, see Chapter 4

Five artificial room impulse responses were generated and applied to the speech data The impulse responses were designed to represent typical transfer functions occurring between the lips and a microphone placed in a normal room The dimensions of the room were selected as 2 5 m by 3 0 m by 2 7 m Reflection coefficients of 0 9, 0 7 and 0 7 were chosen for the walls, floor and ceiling, respectively These parameters led to a reverberation time of 0 25 s for the enclosure This is typical for an office environment with some soft fabric wall and floor coverings The source was placed near the centre of the room, at coordinates (0 9 1 9, 0 7) m and source-receiver distances of 10, 20, 30, 40 and 50 cm along the y-axis were used In this way, the test data spans the range of lip-microphone distances normally encountered in conventional speech processing applications The first part of the test data generated for source-receiver distances of 30 and 50 cm is shown in Fig C 2

C.5 OFFICE SPEECH DATA

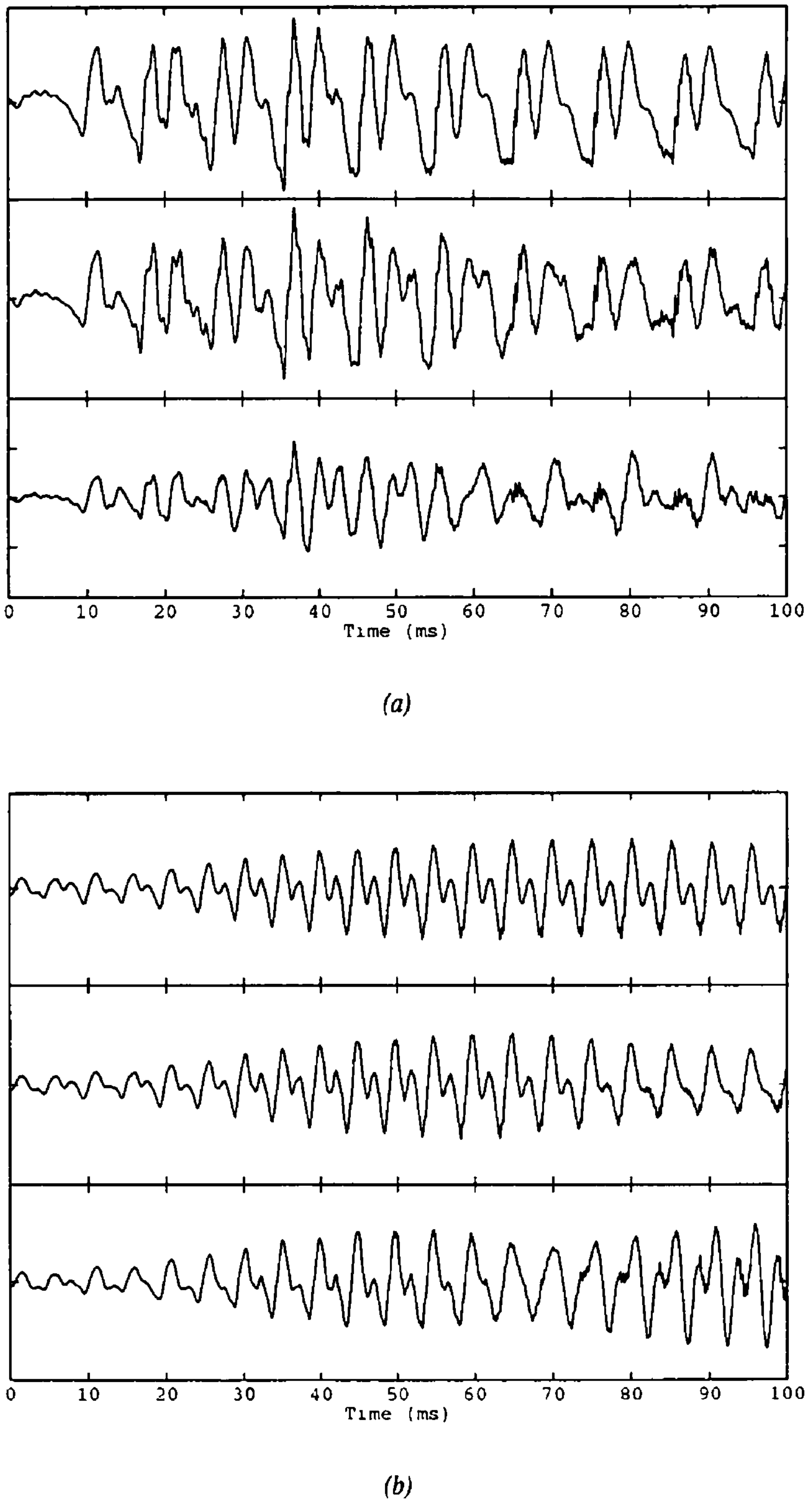
To obtain samples of speech under normal ambient conditions, the recording process was carried out in a typical office environment in the presence of background noise and reverberation Two subjects were used one male and one female Both subjects were different from those used in the noiseless experiments but, again, they were Irish dialect speakers of British English The subjects were requested to read the sentence "Early one morning, a man and a woman ambled along a one mile lane" The



(a)

(b)

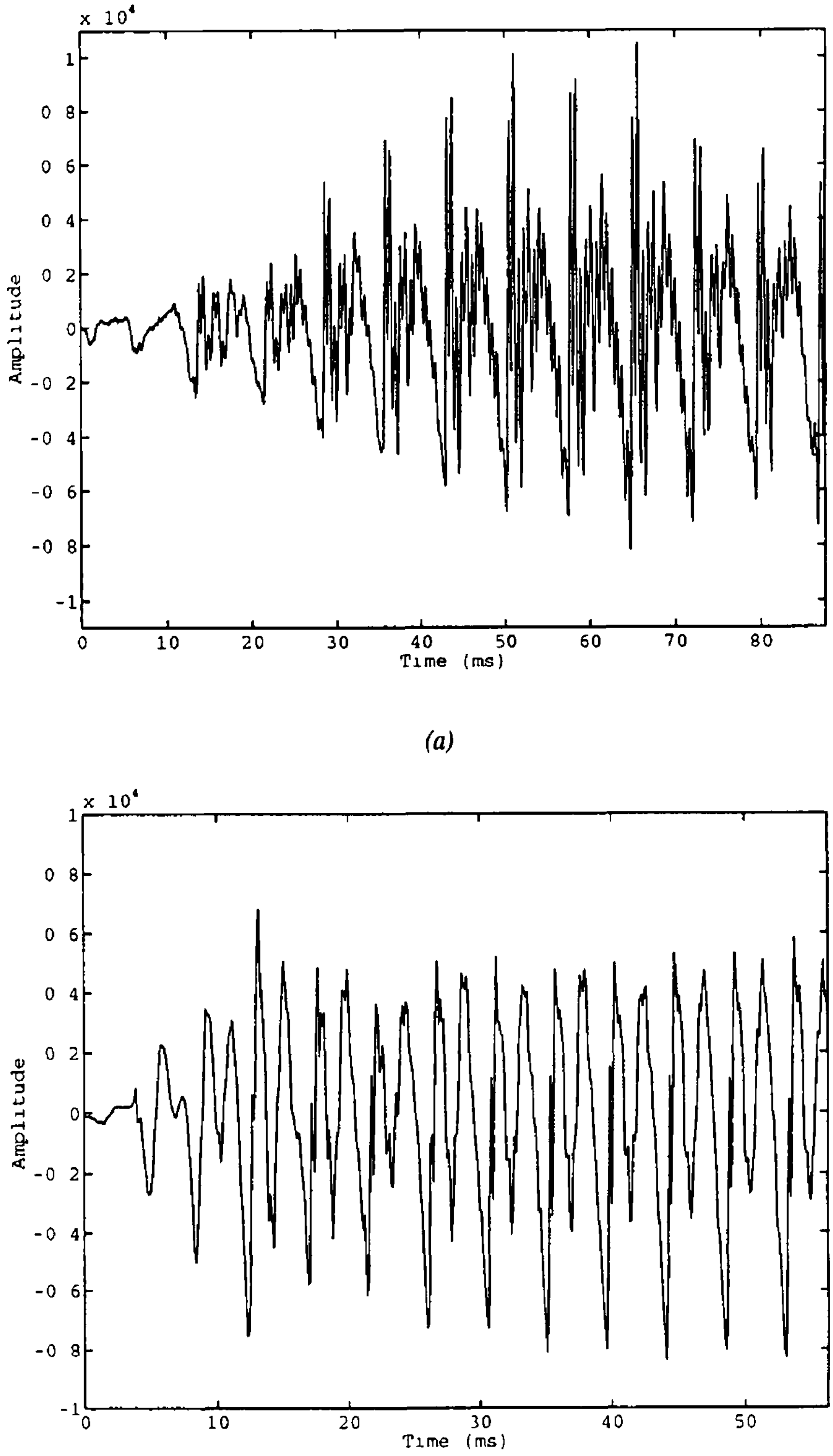
Fig C 1 Segments of noiseless and noisy speech test data (a) male subject, (b) female subject from top to bottom noiseless recording, 25 dB SNR, 15 dB SNR



(a)

(b)

***Fig C 2** Segments of noiseless and reverberant speech test data (a) male subject, (b) female subject from top to bottom. noiseless recording, 30 cm source-receiver distance 50 cm source-receiver distance*



(a)

(b)

Fig C 3 Segments of the office speech test data (a) male subject, (b) female subject

recording room was L-shaped with rough brick walls, a smooth concrete floor and acoustic ceiling tiles The dimensions of the room were as follows room annex 3 m by 3 m, main room 7 m by 6 m, height 2 5 m The recordings were made with the subjects standing about 1 m from the nearest wall in the room annex A lip-microphone distance of roughly 20 cm was used in both cases Furthermore, the recording equipment and procedure were the same as that employed for the noiseless recordings The SNRs of the male and female recordings were approximately 38 dB and 41 dB, respectively Fig C 3 shows part of the speech waveform from the first word recorded by the two subjects

C 6 CONCLUSION

This appendix describes the production of test speech data for use in the experiments detailed in this thesis Four types of test data were generated - noiseless, noisy, reverberant and office The noiseless data was recorded under near anechoic conditions in a sound-proofed studio The noisy data was produced by adding white noise to the noiseless recordings The reverberant data was obtained by convolving simulated room impulses with the noiseless speech recordings The office data was captured by recording natural speech in a typical office environment under conditions of noise and reverberation The speech data consisted of all-voiced sentences read aloud by male and female subjects Different subjects were used for the noiseless and office recordings Care was taken to ensure that the recordings underwent no phase distortion