JOINT ESTIMATION OF VOCAL TRACT AND SOURCE PARAMETERS OF A SPEECH PRODUCTION MODEL

A thesis for the degree of Ph.D.

Presented to

DUBLIN CITY UNIVERSITY

by

KANAGARATNAM JEEVAKUMAR, B.E.

SCHOOL OF ELECTRONIC ENGINEERING

DUBLIN CITY UNIVERSITY

RESEARCH SUPERVISOR

DR. RONAN SCAIFE

July, 1993

IN

MEMORY OF MY PARENTS

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: K Jeewakumar.

Date: July, 1993

ACKNOWLEDGEMENTS

I am greatly indebted to Ronan Scaife, my research supervisor. Ronan has constantly guided my work and encouraged my development in the field of speech technology and communication systems. Without his enthusiasm and continual drive to increase our understanding of speech technology, we would not have acquired such an in-depth knowledge in modelling of speech production systems.

I am grateful to Professor Charles McCorkell, Head of the School of Engineering, for obtaining financial support and providing encouragement. I would also like to thank the staff of the School of Engineering for their help during the course of this project

I am indebted to Dr. Ailbhe Ni Chaside of Trinity College Dublin for discussions on speech topics and for providing some of the financial support The first year of this research was carried out in University College Galway (UCG), after my undergraduate studies there I am thankful to my teachers and friends at UCG for their contribution to the development of my career.

I am extremely grateful to my colleagues and friends in the speech laboratory. Chris and Billy gave freely of their time to read my thesis. Chris has also made critical comments on the format of the thesis. Thanks to Albert for helping with some of the diagrams.

Thanks to Lotus, Gulyani, Motti, Joice, David, Michael, Suhumar and other friends who have made my social life enjoyable in Dublin. I am also grateful to my friends Sebastian and his family for their love and care

Finally and above all I acknowledge the love and understanding of my brother, BIL Jeya and sisters Acca, Mala, Pama and Kala. I also acknowledge the kind support of Mathini and Chandran Hugs and kisses to my nieces and nephews

JOINT ESTIMATION OF VOCAL TRACT AND SOURCE PARAMETERS OF A SPEECH PRODUCTION MODEL

KANAGARATNAM JEEVAKUMAR

ABSTRACT

This thesis describes algorithms developed to jointly estimate vocal tract shapes and source signals from real speech The methodology was developed and evaluated using simple articulatory models of the vocal tract, coupled with lumped parametric models of the loss mechanisms in the tract

The vocal tract is modelled by a five parameter area function model [Lin, 1990] Energy losses due to wall vibration and glottal resistance are modelled as a pole- zero filter placed at the glottis A model described in [Laine, 1982] is used to approximate the lip radiation characteristic.

An articulatory-to-acoustic "linked codebook" of approximately 1600 shapes is generated and exhaustively searched to estimate the vocal tract parameters

Glottal waveforms (input signals) are obtained by inverse filtering real speech using the estimated vocal tract parameters. The inverse filter is constructed using the estimated area function A new method is proposed to fit the Liljencrants - Fant glottal flow model [Fant, Liljencrants and Lin, 1985] to the inverse filtered signals Estimates of the parameters are found from both the inverse filtered signal and its derivative

The described model successfully estimates articulatory parameters for artificial speech waveforms Tests on recorded vowels suggest that the technique is applicable to real speech

The technique has applications in the development of natural sounding speech synthesis, the treatment of speech disorders and the reduction of data bit rates in speech coding

CONTENTS

List of symbols

c Velocity of sound (= 353 m s^{-1})

ρ Density of air (= 1.14 kg m^{-3})

μ Viscosity coefficient of air (= $1.86*10^{-6}$ Pa s)

f Frequency (Hz)

ω Angular frequency (rad s^{-1}), $\omega=2\pi f$

L Inductance

C Capacitance

R Resistance

G Conductance

Z_t Input impedance of the vocal tract

Z_w Wall impedance, $Z_w = R_w + j\omega L_w$

Z_g Glottal impedance, $Z_g = R_g + j\omega L_g$

Z_l Lip impedance, $Z_l = R_l + j\ \omega L_1$

r_G Glottal reflection coefficient, $r_G = \dfrac{Z_g - \frac{\rho c}{A_1}}{Z_g + \frac{\rho c}{A_1}}$

r_L Lip radiation reflection coefficient, $r_L = \dfrac{\frac{\rho c}{A_N} - Z_l}{\frac{\rho c}{A_N} + Z_l}$

r_k Reflection coefficient between two tubes $r_k = \dfrac{A_{k+1} - A_k}{A_{k+1} + A_k}$

A_k Cross sectional area of k^{th} section of the vocal tract area function (cm^2)

F_i i^{th} formant frequency (Hz)

B_i i^{th} formant bandwidth (Hz)

U_G Volume flow at the glottis, (cm^3 s^{-1})

U_L Volume flow at the lips, (cm^3 s^{-1})

A_c Minimum area of the main tongue body constriction in the vocal tract

X_c Axial location of the tongue body constriction along the vocal tract

A_0 Lip opening area

L_0 Lip protrusion length

List of Acronyms and keywords

CELP - Code Excited Linear Predictive coder

DFT - Discrete Fourier Transform

DIF - Differentiated Inverse Filtered signal

EGG - Electro Glotto Graphy

HMM - Hidden Markov Model

IF - Inverse Filtered signal

LPC - Linear Predictive Coding

RELP - Residual Excited Linear Predictive coder

SPM - Speech Production Model

VQ - Vector Quantisation

List of keywords

Articulatory model - Described by the positions of articulators - jaw, tongue body, tongue tip, velum, hyoid etc.

Linked codebook - List of ordered vector pairs. One vector of the pair describes the parameters of the SPM, the other describes the acoustic parameters of the corresponding SPM parameters.

Formant frequency - Resonance frequency of the vocal tract

Cepstrum - Inverse Fourier transform of the logarithm of the power spectrum of a signal

Lifter - Filter used in the cepstral domain.

CHAPTER 1

INTRODUCTION

1.0 INTRODUCTION

PHYSIOLOGICAL DATA on the human vocal tract apparatus during the production of speech has a number applications in speech research. The central theme of this thesis is the extraction, from voiced speech, of information about the physiology of the human speech production system, such as vocal tract shapes and glottal flow signals A simple Speech Production Model (SPM), based on an area function model has been employed for this purpose This SPM is an enhancement and integration of previous designs [Lin, 1990; Scaife, 1989, Liljencrants, 1985, Laine, 1982]

1.2 FORMAT OF THE THESIS

This thesis is divided into four chapters

Chapter 2: Various models of speech analysis, based on area function and articulatory models are discussed Human speech is produced by the muscular control of articulators (jaw, lip, tongue and velum movements) However, the relationship between the speech signals and the articulatory movements is complex This investigation attempts to extract the variation of cross sectional area with distance along the vocal tract, i.e. the vocal tract area function The area function can be determined from articulatory [Mermelstein, 1973; Coker,1976; Shirai and Honda, 1980, etc] and geometrical vocal tract models [Stevens and House, 1955, Fant, 1960; Flanagan et al, 1980, etc] In the current research, a slightly modified version of the model described in [Lin, 1990] is used. It is based on the following parameters - the minimum tongue body constriction area (A_c), its axial location (X_c), the lip opening area (A_0), the lip protrusion length beyond the teeth (L_0) and the number of small cylindrical tubes needed to represent the total vocal tract length (N). A computationally efficient SPM is designed for the estimation of area function, based on a previous design [Scaife, 1989] This non-interactive (i.e. coupling of the vocal tract to the trachea is ignored in the modelling) SPM is an integration of four separate models - a glottal model, a vocal tract area function model, a wall vibration model and a lip radiation model. The vocal tract, represented by a lattice filter, is related to the area function model [Lin, 1990] The glottal end of the vocal apparatus is terminated by a composition of a wall vibration and a glottal model The lip radiation effects are described by Laine's pole - zero digital model [Laine, 1982]. The equivalent digital network of the SPM is shown in fig. 2.2

This network is similar to the network used by [Scaife, 1989] and hence the same wave propagation equations hold for the current SPM A detailed description of this model is given in section 2.3 and the related articulatory and mathematical models of the vocal tract are discussed in chapter 2

Chapter 3: This chapter describes estimation of the vocal tract parameters of the SPM for artificial speech and for real vowels A 'linked codebook' is employed in the estimation of the model parameters. The linked codebook contains a set of acoustic to vocal tract area function mappings generated by the SPM. The acoustic signal is represented by cepstral coefficients The creation, merits and application of the codebook are discussed This estimation technique reliably chooses the correct vocal tract length, and so eliminates vocal tract normalisation problems Estimation of the parameters based on the 'simplex method' of optimisation is also discussed in this chapter

Chapter 4: The goal of the project is to extract slowly varying vocal tract shapes and glottal flow signals of the vocal apparatus from the rapidly changing speech signals The estimation of the air flow through the vibrating vocal folds is described in this chapter, using the area function estimated in chapter 3. Inverse filtering techniques have been employed previously, using LPC type models and other models [Veeneman and BeMent, 1985; Wong et al ,1979, Price, 1989] However, the use of an inverse filter related to a vocal tract shape model is not a common feature in those techniques This technique and aspects of parameterisation of inverse filtered speech are discussed The parametrisation of the glottal flow is performed using the Differentiated Inverse Filtered (DIF) speech and Inverse Filtered (IF) speech

Chapter 5: Discusses the possibilities of estimating model parameters with a clustered codebook A large codebook may be generated and the size of the codebook can be reduced by clustering e g. k-means algorithm, techniques.

1.2 WHY AREA FUNCTION BASED MODELS?

Synthesising natural sounding speech, developing low bit rate coders and deriving robust algorithms for speech recognition can be said to be the goals to which speech engineering research is aimed. The general opinion among the speech community is that developing a knowledge-based Speech Production Model (SPM), including

comprehensive models of the human speech apparatus, would lead to progress in the fields outlined above [Fant, 1991]. It is thought that one of the most economical descriptions of the speech waveform is in terms of articulatory parameters - parameters that specify the geometrical properties of the vocal and nasal tract, the mechanical properties of the tract walls and the properties of the vocal cord oscillator [Sondhi and Schroeter, 1986]. However, speech devices based on other models e.g. LPC, formant frequency based models, waveform coders, etc , outperform the currently available articulatory based SPMs [Holmes, 1988]. So why expend more research effort in the articulatory modelling (i e. area function based models) of speech? The answers are summarised in the following paragraphs (see also fig 1.1).

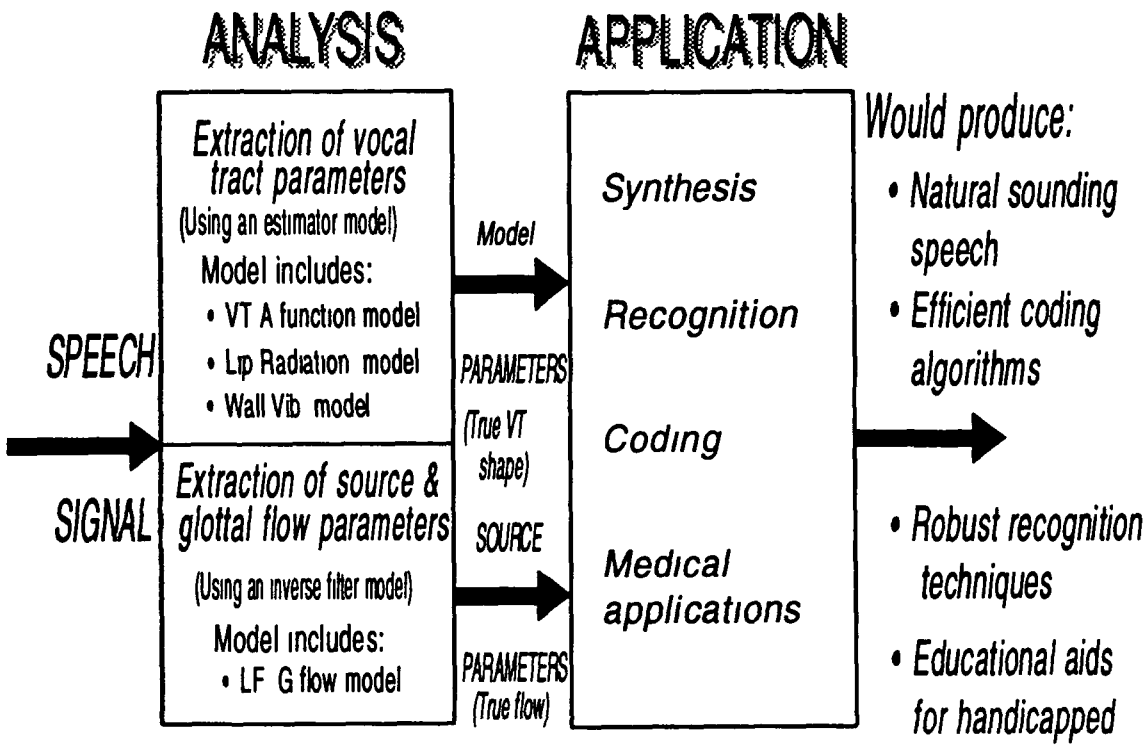


Fig. 1.1. Diagram showing the applications and benefits of using the physiological data of the human speech production system in speech analysis

Speech signals are the output of the articulatorily controlled vocal tract and vocal cord systems. The rapidly changing speech signal can be easily recorded with the aid

of a microphone. However, there is no easy way of monitoring the slowly varying vocal tract shapes generated by the coordination and the regulation of the articulators. It is believed, though, that information about the comparatively slowly varying vocal tract shape (at a frame rate of 100 Hz [Schroeter and Sondhi, 1989]) will lead to better schemes for speech analysis and will provide better educational aids in the medical area. The reasons are as follows:

a) Speech signals can be described by a fewer parameters in articulatory systems than in waveform coders or for LPC coders. These parameters are likely to interpolate over long time intervals. These attributes when applied in speech coding algorithms will reduce the data rate required by speech coders.

b) It has been demonstrated that good quality speech can be produced through phoneme - level control of an articulatory model [Flanagan, Ishizaka and Shipley, 1975].

c) Variations in vocal tract length and the articulatory organs in different speakers make speaker independent recognition difficult. Ideally, the estimation of the articulatory model does not depend on the speaker be they for example male, female or child. Therefore, articulatory parameters can be used to develop speaker independent recognition systems [Shirai and Kobayashi, 1986].

d) In speech therapy: Where the vocal tract shapes and the air flow through the glottis (excitation signal) captured by the articulatory model can be used in speech therapy for the teaching deaf.

As an introduction to this area of speech research, we will briefly discuss speech synthesis, coding and recognition techniques.

1.3 SPEECH SYNTHESIS

Artificial speech synthesis is the process of generating speech from mathematical models and machines using control parameters. The applications of speech synthesis include information services, reading machines for the blind (text to speech systems) and communication aids for people with speech disorders.

Historically, the first speech synthesiser was produced by Kratzenstein in 1779. By using a mechanical model for the human speech production system, Kratzenstein synthesised five English vowels. In 1791, Von Kempelen produced an elaborate model, similar to Kratzenstein's model which synthesised connected utterances [Flanagan,

1972]. Since then many mechanical, electrical and computer models have been developed for speech synthesis The basic principles used in these devices are similar The main difference between the devices is that the continuous improvement of digital signal processors, accompanied by their reduced cost allows current synthesisers to use more complex algorithms and better models One of the most recent speech synthesisers is Klattalk (1987) Klattalk led to the development of the commercial synthesiser Dectalk produced by Digital Equipment Cooperation, currently thought to be the best speech synthesiser available in the market [Edwards,1991] Speech synthesis can be classified into three groups

a) Synthesis based on waveform coding Speech signals are synthesised from stored human speech.

b) Synthesis based on analysis - synthesis. Parameters of a Speech Production Model (SPM) e.g. Channel vocoders, formant vocoders, are estimated from a recorded speech signal and stored. The speech is then synthesised from these stored parameters

c) Synthesis by rule Speech is produced based on phonetic and linguistic rules from stored letter sequences or sequences of phoneme symbols and prosodic features

These three types of synthesis methods are described below.

1.3.1 Speech synthesis based on waveform coding

Speech sounds can be synthesised by a variety of methods which construct the signal waveforms from a number of basic sounds. Early synthesisers used a large amount of stored words to produce messages The first application of this technique was a speaking clock introduced into the U.K telephone system in 1936 [Holmes,1988] Good quality speech can be produced by this technique for a limited number of sentences

High quality speech can be generated using large elements of stored speech sounds such as phrases and sentences. However, this restricts the number of sentences which can be generated With small elements e g. phonemes and syllables, a large number of sentences can be produced However the quality is degraded The reason is that with smaller elements, connecting neighbouring elements with acoustic continuity becomes more difficult Therefore, extra care must be taken concerning pitch, timing and intonation to achieve fluent and natural sounding speech.

Currently available systems generate speech from stored words and phrases

[Furui, 1989] The storage requirement is reduced by storing waveform coded speech (e.g Adaptive Pulse Code Modulation (ADPCM)') rather than simply storing analog or digital (Pulse Code Modulation i.e. PCM) speech signals. However, the problems of connecting the neighbouring elements, coupled with the large storage requirement for a large number of words, make this synthesis method less elegant in modern speech synthesis.

1.3.2 Speech synthesis using concatenation of vocoded speech (analysis - synthesis method)

In this method, speech sounds are analysed using SPMs and these SPM parameters are then stored and used in the synthesis of speech The number of parameters required to describe the SPM in the analysis synthesis method is very small compared to the number of parameters required to directly represent the speech signals. Therefore the space required for storage is smaller

Two methods are used to simulate the human vocal apparatus - a) terminal analog method b) vocal tract analog method The terminal analog method simulates the frequency spectrum structure of the speech sounds to produce speech e.g Formant speech synthesisers, channel vocoders [Furui, 1989] The vocal tract analog method simulates the acoustic wave propagation in the vocal tract e g articulatory speech synthesis.

As mentioned, articulatory modelling of the speech production system is the topic of this thesis. Articulatory speech synthesis, though not widely used in current systems, promises to produce high quality speech at a low bit rate (see section 1 2)

The basic principles involved in synthesising speech are the same. However, the efficiency of the algorithms can be improved in a number of ways. New technologies, due to the availability of fast signal processors, improve the speech quality by using complicated matching and estimation techniques

1.3.2.1 Speech synthesis using LPC vocoded speech

The first single chip (TMC0280) speech synthesizer was introduced by the Texas Instruments in 1978. The chip, which was produced by combining DSP techniques with VLSI technology, was available at low cost

The TMC0280 chip was implemented using LPC techniques of speech analysis LPC techniques, which have been a major research tool in speech processing, vastly reduce the storage requirement without seriously degrading the intelligibility of speech The number of LPC coefficients required in LPC analysis vary with the sampling frequency. This is described in other literature [Atal and Hanauer, 1971]

In the implementation of this chip, 10 LPC coefficients were used for representing the spectra together with 2 further parameters describing the energy and the pitch. These 12 parameters are updated every 20 msec The 10 LPC coefficients were converted to reflection coefficients before coding. This is because reflection coefficients can be quantised more efficiently than the LPC coefficients. In addition the values of reflection coefficients indicate the stability of the LPC filter. Further, it was noticed that the lower order reflection (or LPC) coefficients are more important to intelligibility than the higher ones, thus, lower order coefficients were coded with a larger number of bits.

LPC based synthesisers in the modern devices can be improved in a number of ways, for example

a) A variable number of LPC coefficients can be used to more accurately represent the formants of different sounds.

b) A more realistic glottal flow signal, rather than a train of impulses or white noise, can be used to excite synthesiser

c) Better matching algorithms can be used.

1.3.2.2 Formant speech synthesiser

As the name implies, the formant speech synthesiser is defined in terms of the first few formant frequencies (i.e resonance frequencies of the vocal tract) and the amplitudes associated with the resonance modes of the vocal tract Formant synthesisers are different from the LPC synthesisers in that the formants of the filter are closely related to the shape of the vocal tract. Like LPC synthesisers formant speech synthesisers are also available in a single chip One such example is the NEC μPD7752 device, while another is the MEA8000 from Phillips [Quarmby and Holmes, 1984]

One advantage of the formant synthesiser is that variation of the first formant bandwidth due to time varying glottal impedance can be easily provided [Witten, 1982,

Holmes, 1988] In addition, as with other synthesisers, different voice source models can be incorporated into the system. These features have been included in the KLSYN88 formant synthesiser [Klatt and Klatt, 1990].

Two types of formant synthesisers exist, cascaded and parallel. For voiced sounds, cascade synthesisers offer a theoretical representation of the un-branched acoustic tube of the real system Consonants and nasal sounds can be better represented by a parallel synthesiser. Formant peaks and the boundaries between the formants are not well defined in the nasal and consonant sounds Therefore, controlling each formant separately by parallel filters provides a better representation for "non-voiced" sounds Also, since each filter is independent of all others, the amplitude and frequency of the formants and the excitation signal of the filter can be individually controlled to produce perceptually good speech.

A typical formant synthesiser incorporates both cascade and parallel filters to produce continuous utterances (e g KLSYN88 synthesiser [Klatt and Klatt, 1990]). The *computalker* [Witten, 1982], developed as a cascade synthesiser, produces intelligible synthetic speech.

1.3.2.3 Speech synthesis using channel vocoded speech

Speech, in this case, is represented by the response of a bank of contiguous variable-gain band pass filters. This is a terminal analog method, that is, the articulation is simulated by the frequency structure of the speech signals Very briefly, individual filters are obtained by matching the input spectrum to the sum of the frequency responses of all the individual filters. A large number of filters (channels) are required to produce high quality speech However, intelligible speech can be produced with 15-20 filters.

1.3.2.4 Articulatory based system

This type of synthesis is based on the vocal tract analog method, whereby speech is produced by simulating the propagation of acoustic waves in the vocal tract Synthesis of speech by the control of articulatory parameters is called articulatory speech synthesis It is believed that articulatory speech synthesis has several advantages over other methods The advantages of this system and different articulatory models of

speech are discussed in chapter 2. Briefly, the method has the potential to produce natural sounding speech at a low bit rate.

However, obtaining the articulatory parameters is not a trivial task. Therefore, much research effort is being expended in this area to improve the speech quality. Systems based on articulatory models [Gupta and Schroeter, 1991; Parthasarathy and Coker, 1992; Schroeter, Larar and Sondhi, 1987] are available for speech synthesis. However, these models have not yet produced good quality speech, in comparison to LPC vocoded and waveform concatenated speech systems.

1.3.3 Speech synthesis by rule

Speech synthesis by rule is a method for producing any words or sentences based on phonetic/syllabic symbols or letters. The rules are formed to connect the feature parameters of speech elements (phonemes, syllables, etc.), prosodic parameters (amplitude and pitch), intonation and accent from the stored material. High quality speech is difficult to obtain using the small speech elements due to the complexity of the connecting rules. However, a good phoneme based system for synthesis was produced by [Klatt, 1987].

Large elements e.g. CV syllables (Consonant Vowel syllable), CVC syllables and VCV syllables, when used in the synthesis systems reduce the complexity of the connecting rules [Furui, 1989].

The synthesis by rule technique is used in Text - to - speech systems. In these systems, the input text is first converted to a phonetic transcription. The phonetic transcription is then used to generate the speech waveform. Presently available Text - to - Speech systems are Votrax, MITalk and DECtalk which is an advanced version of the earlier MITalk system.

1.4 SPEECH CODING

Efficient coding of the speech signal is necessary for transmitting speech over existing communication channels with reduced cost. As mentioned previously, the topics of coding, synthesis and recognition are highly related and therefore efficient coding of speech would also provide good quality synthesisers and recogniser. Very high quality speech can be coded at 300 kbits/sec with wide bandwidth (20 Hz - 20 kHz) and high

Signal to Noise Ratio (SNR) However, good telephone quality speech can be produced at 40 kbits/sec using a narrower bandwidth (20 Hz - 3 3 kHz) and a lower Signal to Noise Ratio (SNR) [Furui, 1989]

Several different coders have been produced over the years (see Table 1.1) Earlier systems coded the speech directly and consequently used a high bit rate As technology and methods have improved, speech is now coded at low bit rates Modern algorithms achieve this reduction by eliminating redundancies in the speech signal These redundancies are due to the physiology of the human speech production and auditory system

Two classes of coders exist: waveform coders and voice coders (vocoders) The waveform coders, in general, produce high quality speech at high bit rates The vocoders on the other hand, produce intelligible speech at low to medium bit rates using knowledge based speech models

1.4.1 Basic Coders

The properties of waveform coders and analysis-synthesis coders are summarised in Table 1.1

Waveform coders are well established and are widely used for telephone transmission. References for these coders can be found in several speech text books [Rabiner and Schafer, 1978; Furui, 1989; etc.,]

In contrast to the waveform coders, vocoders exploit the special properties of speech. Though, this requires complex mathematical algorithms and mathematical modelling of the human vocal apparatus, it reduces the bit rate required for the transmission of speech

1.5 SPEECH RECOGNITION

Attempting to teach computers to recognise spoken words is termed speech recognition Speech recognition can be simply done, at least in theory, by storing the words as acoustic signals (templates) and subsequently comparing the input word with the templates. However, it is not a trivial task as a word cannot be repeated in exactly the same way even by the same person. Therefore simple comparisons of time domain waveforms are, in general, not very successful Recognition can be better performed

	Waveform coder (Time domain)	Vocoder (Spectral domain)
Information coded	Waveform	Spectral envelope and source signals(pitch,amplitude, voiced/unvoiced)
Coding bit rate	9.6-64 kbps (medium or wide band)	2.4-4.8 kbps (narrow band)
Coding objective	Any sound	Speech sounds
Coding performance measure	Signal to Noise Ratio SNR	Spectral distortion
Disadvantages	Difficult to reduce the bit rate	Vulnerable to noise and transmission error, Processing is complex
Examples	Time domain coding: PCM,ADPCM,ΔM,AΔM ,APC. Frequency domain: SBC,ATC Combination: APC-AB	Channel vocoder, formant vocoder, phase vocoder, LPC vocoders

Table 1.1 Properties of waveform and vocoders (after [Furui, 1989])

in the spectral domain.

As with the speech coding and synthesis systems, articulatory parameters could help to produce better speech/speaker recognition systems. In [Shirai and Kobayashi,

1986], a feature vector based on an articulatory model, was successfully used to recognise the vowel sounds We believe that models based on articulatory parameters could be used in the recognition algorithm. As mentioned previously, articulatory parameters vary slowly with time and therefore can produce good results in recognition systems. As shown in fig 1.2, speech recognition involves a few vital steps Extraction of suitable feature vectors by the analysis of speech, application of an optimum distortion measure and recognition decision.

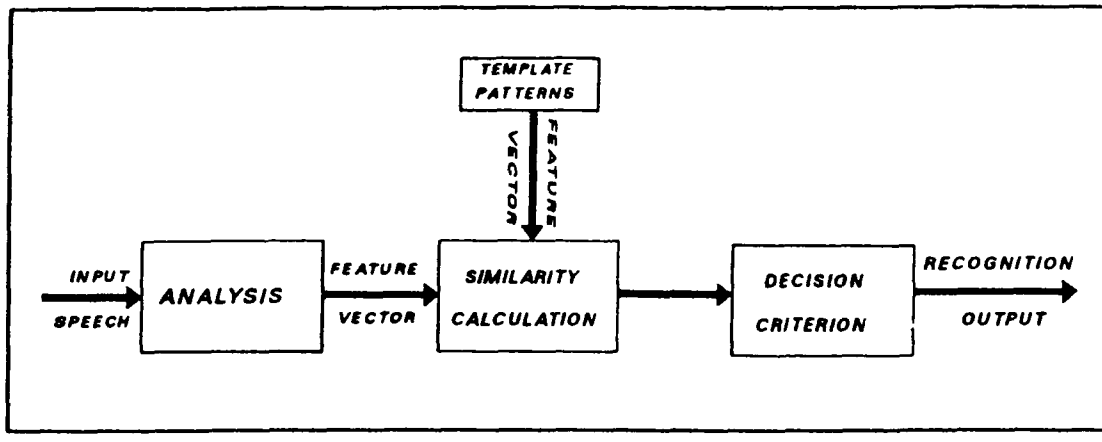


Fig. 1.2 Block Diagram of a Speech Recogniser

As with vocal tract estimation and source estimation techniques, a good distortion measure is essential for successful recognition systems A good distortion measure should be sensitive to important characteristics (e.g. formants , bandwidth, etc.) of the input and the template patterns, and insensitive to irrelevant components (eg. high frequency components) of the input and template patterns.

A variety of different spectral distortion measures are available. Spectral parameters such as LPC spectral envelope, cepstral coefficients, autocorrelation coefficients and FFT spectrum are often used for such spectral comparisons

Recently, Hidden Markov Model (HMM) has been widely used in speech recognition. HMMs consist of a non-observable or 'hidden' process and an observation process which links the acoustic vector extracted from the speech signal, to the states of the hidden process. This model has been widely exploited using different processing methods to recognise speech [Merhav and Epraim, 1991, Ghitza and Sondhi,1993]

1.6 SUMMARY

We discussed the possible applications of area function/articulatory models in the area of speech synthesis, coding and speech/speaker recognition A few examples of existing systems and generally used methods were discussed The benefits of the use of physiological models were also discussed Different types of speech sounds and different phoneme categories are given in appendix A1 and A2.

CHAPTER 2

SPEECH PRODUCTION THEORY AND MATHEMATICAL MODELLING OF THE SYSTEM

2.0 INTRODUCTION

Section 2.3 gives a detailed description of work done on simple Speech Production Models (SPM) for voiced sounds.

To introduce the reader to the articulatory modelling of speech, a short discussion on the theory of speech production and a brief review of articulatory models are given in sections 2.1 and 2.2.

The human vocal apparatus includes the lungs, larynx, vocal tract and nasal tract. The vocal tract, the principal path of the sound transmission, particularly for vowels, begins at the glottis and terminates at the lips The nasal tract is an ancillary path of sound transmission which begins at the velum and ends at the nostrils In the production of nasalised sounds, the nasal tract and vocal tract are coupled at the velum and the sound pressure waveform is radiated at the nostrils and lips

For non-nasalised sounds, the human vocal apparatus can be divided into three sections (see Fig 2 1), namely the subglottal system, the vocal tract system and the lip radiation system. These three sections interact with each other in the production of speech

The subglottal system generates a steady flow of air from the lungs which powers the speech production system The larynx transforms this airflow into a series of puffs of air by opening and closing the vocal cords The transformation of this steady airflow into the puffs of air is called phonation The primary function of the larynx/vocal cords is to modulate airflow in such a way to generate acoustic energy at audible frequencies.

The vocal tract system acts as a variable acoustic filter for the puffs of air obtained at the larynx. This filter lets differing amounts of acoustic energy through at different frequencies by varying the vocal tract shape. This relationship between the vocal tract and frequency spectra implies that the vocal tract shapes attained during the production of speech sounds can be characterised by the speech spectra

Finally, the volume flow of air through the larynx and the vocal tract is radiated as a pressure signal at the lips and/or at the nose

Much research effort has been expended in the last few decades to more accurately model this system As a result, a variety of different mathematical models, based on articulatory parameters and area function parameters have been produced [Stevens, Kasowski and Fant 1953; Stevens and House, 1955; Fant, 1960, Mermelstein,

1973, Flanagan, Ishizaka and Shipley, 1975, Coker, 1976, Atal et al, 1978, Maeda, 1979, Shirai and Honda, 1977; Lin, 1990] A simple SPM is proposed in line with current speech research. This model, described in section 2 3, includes wall vibration, lip radiation effects together with the glottal losses of the real system.

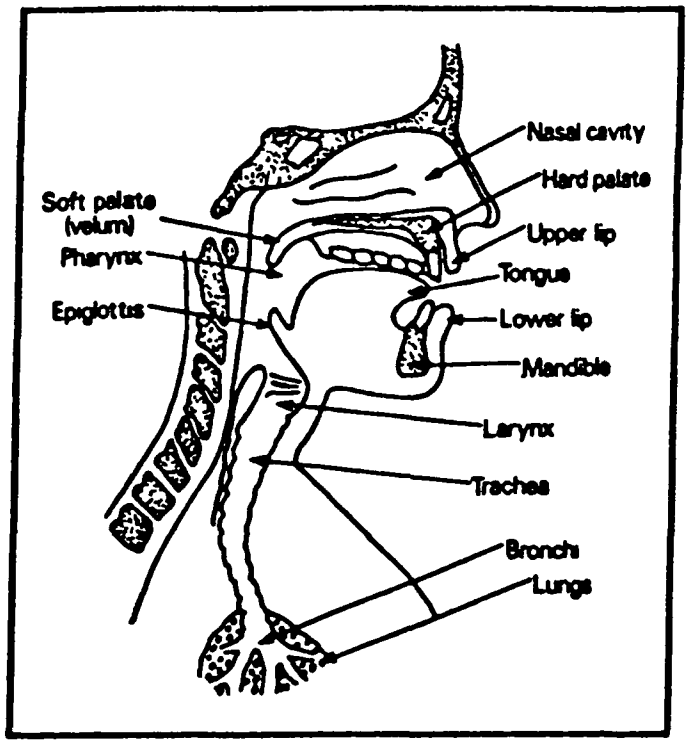


Fig. 2.1 Human vocal mechanism, showing the vocal organs [Holmes,1988]

2.0.1 Sources of speech data for speech research

The speech pressure waveforms we discussed above are referred to as acoustic waveforms or acoustic parameters. Whereas, vocal cord positions (open/closed) , vocal tract shapes (depending on the jaw, velar, tongue movements) and the lip size (rounding/opening) are referred to as the articulatory parameters

Acoustic data which can be easily recorded with the aid of a microphone is usually the basis of any speech research. However, acoustic data alone is not enough to fully understand the dynamics of speech production Additional data on the physiological system, for example the position of the constriction point of the vocal tract, the constriction area and the vocal tract length should help to increase our understanding of the speech production system

There are several techniques available to accumulate articulatory data during

phonation with reasonable accuracy However the techniques used, e.g. X - radiography, high speed photography, etc. [e.g. Holmes, 1988, Schroeter, 1967, Lieberman, 1977], for the acquisition of articulatory parameters are expensive and not easily available Furthermore, these measuring techniques involve physical contact of some instruments with the speech production system and hence they inevitably change the natural speaking conditions. For these reasons, direct measurements are not very popular in speech research. On the other hand, obtaining the articulatory data from the easily available acoustic speech signal is a more attractive method However the transformation of acoustic to articulatory parameters is not straightforward Since speech signals contain a variety of different sounds (voiced, unvoiced, silent) with different characteristics, forming a unique scheme covering all of the speech sounds for the acoustic to articulatory transformation is not feasible Therefore, separate schemes should be formed using appropriate articulatory models for different sounds In chapter 3, a set of acoustic - to - articulatory mappings (linked codebook) for vowel type sounds is generated usmg a modified version of the area function model described in [Lin, 1990]. This codebook is then used to transform the acoustic parameters to articulatory parameters.

To be able to form mathematical models for the speech production system, one has to understand the basic principles involved in the generation and propagation of speech The aim of this chapter is threefold Firstly, it summarises the essential theories and models used in speech engineering Secondly, it discusses various mathematical models. Finally, a simple non-interactive SPM is developed, using the models of various parts of the vocal apparatus, to analyze and synthesize vowel type sounds

2.1 ACOUSTIC THEORY OF SPEECH PRODUCTION

In order to design and implement mathematical models of the speech production system, it is essential to understand the fundamentals of the speech production Knowledge of the physical system is of paramount importance. This includes knowledge of the losses in the system, softness of the vocal tract walls, position of the articulators and their effects on tract shapes. The propagation of the sound pressure waveform along the vocal tract for different frequencies and tract shapes must also be understood

2.1.1 Waveform propagation along the vocal tract

The airflow generated by the lungs, propagates along the vocal tract from glottis to the lips. The propagation of the waveform is described by fundamental laws of physics In particular, the conservation of energy, the conservation of momentum and the conservation of mass along with the laws of thermodynamics and fluid mechanics [Rabiner and Schafer, 1978]

Using the above laws and including some simple assumptions about the losses and vocal tract shapes, a set of partial differential equations can be obtained With the additional assumption of plane wave propagation (which holds for frequencies less than 4 kHz [Rabiner and Schafer, 1978]), eqn (2 1) was derived for sound waves in a lossless tube [Portnoff, 1973].

$$-\frac{\delta p}{\delta x} = \rho \frac{\delta (u/A)}{\delta t} \qquad 2.1a$$

$$-\frac{\delta u}{\delta x} = \frac{1}{\rho c^2} \frac{\delta (pA)}{\delta t} + \frac{\delta A}{\delta t} \qquad 2.1b$$

where

$p = p(x,t)$. the variation in sound pressure in the tube at position x and time t

$u = u(x,t)$: the variation in volume velocity flow at position x and time t

ρ the density of air in the tube

c · the velocity of sound

$A = A(x,t)$ the cross sectional area of the tube at position x and time t

The solutions to eqn. (2.1) are not easily obtainable for sound wave propagation in the human vocal tract. Area function data $A(x,t)$ is required, as are the boundary conditions at the glottis and lip However, for a small tube of constant area A, eqn (2 1) has the solution of the form·

$$u(x,t) = [u^{+}(t-x/c) - u^{-}(t+x/c)] \qquad 2.2a$$

$$p(x,t) = \frac{\rho c}{A}[u^{+}(t-x/c) + u^{-}(t+x/c)] \qquad 2.2b$$

where $u^{+}(t-x/c)$ and $u^{-}(t+x/c)$ are positive and negative going travelling waves in the tube respectively (see fig B1b in appendix B1)

2.1.2 Transfer function of the speech production system

By considering the vocal tract as a concatenation of small cylindrical tubes (fig B1a in appendix B1) the transfer function relating the pressure at the mouth to the flow at the glottis is obtained (eqn (2 3)). The continuity principle that the pressure and volume velocity cannot change instantaneously, is used at the tube junctions to derive the transfer function. A derivation of eqn (2 3) is given in appendix B1

$$V(z) = \frac{U_L(z)}{U_G(z)} = \frac{0.5(1+r_G)(1+r_L)z^{-N/2}\prod_{k=1}^{N-1}(1+r_k)}{D(z)} \qquad 2.3$$

where

$$D(z) = [1, -r_G]\begin{bmatrix} 1 & -r_1 \\ -r_1 z^{-1} & z-1 \end{bmatrix} \cdots \begin{bmatrix} 1 & -r_{N-1} \\ -r_{N-1} z^{-1} & z-1 \end{bmatrix}\begin{bmatrix} 1 \\ -r_L \end{bmatrix}$$

r_k - refection coefficients at the junctions of small cylindrical vocal tract tubes

r_L,- frequency dependent reflection coefficient at the lips

r_G,- frequency dependent reflection coefficient at the glottis

The transfer function of the above type $V(z)$, is used to implement the SPM of the current research.

2.2 MATHEMATICAL MODELLING OF THE SPEECH PRODUCTION SYSTEM

The transfer function of the vocal apparatus discussed in section 2 1 is estimated differently according to the application

Statistical modelling of speech is a well established area and produces good quality speech by analysing the speech signals Different techniques, e g Hidden Markov processes, different type of LPC models and very recently "neural net works", have been used. Although, the statistical models perform excellently in the speech analysis, synthesis and recognition area, the drawback is that these models do not explain/extract the knowledge (physiological parameters of the vocal apparatus) of the system which produced the speech [Fant, 1990] Formant vocoders (see section 1.1.3.2), for example, characterise the vocal tract by the first few formants i e resonance frequencies of the vocal tract of the system LPC analysis represents the vocal tract section by an all pole filter The poles of this filter are estimated by optimising the transfer function of the filter with respect to the acoustic properties of the speech signal The parameters of the models have, in general, no relation to the true physical shapes of the system attained during speech production

In contrast to statistical modelling, the articulatory modelling of speech tries to depict the physiological shapes of the system obtained during speech production A variety of different vocal tract shapes are produced during the production of speech signals Therefore the goal in articulatory modelling is to construct a model whose solutions match those of the vocal tract Several models have been proposed in terms of the articulators (jaw, lip, tongue, velum movements) that control the mechanism A few of the parametric models widely used to represent the instantaneous articulatory state of the vocal tract are.

a) Mermelstein's model (defined in terms of articulators [Mermelstein, 1973])
b) Coker's model (defined in terms of articulators [Coker, 1976])
c) Fant's three parameter model (defined m terms of area function parameters [Fant, 1960])

2.2.1 Review of some of the articulatory models of the vocal tract

A vocal tract model represented by six variables, specifying the position of the jaw,

tongue body, tongue tip, lip, velum and hyoid, is described in [Mermelstein, 1973]. This model uses both fixed and movable coordinate systems to represent the articulators. The position of the jaw and tongue in this model are expressed in a fixed coordinate system. Lip and tongue body positions are expressed with respect to the moving jaw. Tongue tip position is specified relative to the tongue body. The function of each variable and the conversion of this articulatory model to the area function are discussed in [Mermelstein, 1973]. Shapes for both the vowel and the consonants can be expressed using this model.

A model with a smaller number of parameters is introduced in [Coker and Fujimura, 1966]. This model represents the vocal tract configuration in terms of tongue body, tongue tip, lip and velum positions. When appropriately controlled the model can generate all the English phonemes and has been used to generate contextually complex sentences. This model was later modified in [Coker, 1976]. In this modified model the vocal tract is represented by nine variables. Three variables define the tongue coordinates, two define lip rounding and lip closure and two others define the raising and curling back of the tongue tip. Of the remaining two variables one variable controls the general-purpose cross section transformation and the other represents the position of the velum. Some of the variables in combination have major roles in producing different sounds. For example only two of the three tongue body variables are important in producing vowel sounds and the third variable is important in generating */g/* type sounds. Similarly, variables which control the constriction are important in producing */s/*, */d/* and */l/* consonants. A detailed discussion of the model can be found in [Coker, 1976].

The Mermelstein and Coker models when properly controlled, can generate all the English phonemes. These two models have been successfully used in articulatory speech synthesizers and in articulatory controlled text-to-speech systems [Parthasarathy and Coker, 1992].

2.2.2 Vocal tract cross sectional area described by area function parameters

It was shown in [Stevens and House, 1955; Fant, 1960] that the vocal tract system can be approximately described by three area function parameters. These three variables, called the *place of articulation* (constriction point along the vocal tract) , the *degree of opening* (constriction area) and the *amount of lip rounding* (lip area/lip protrusion

length) have been used differently in different models

A parabolic vocal tract model using the above three variables was given in [Stevens and House, 1955]. Vocal tract cross sectional area, in this case, is defined in terms of cross sectional radii as.

$$r_b(x) = r_f(x) = r_c + 0.025(1.2 - r_c)x^2 \qquad 2.4$$

where the radius of the constriction area $r_c = \sqrt{(A_c/\pi)}$, and the cross sectional area at point x $A(x) = \pi r^2(x)$ $r_b(x)$ & $r_f(x)$ are the radii of the back and the front vocal tract cavities from the constriction

Fant illustrated the relevance of the above three parameters by three-tube and four-tube resonator models [Fant, 1960] By varying the parameter values Fant showed that a variety of different formants can be produced.

Fant also used a horned-shaped tongue section to produce a more suitable model for the vocal tract using the three parameters. This model produces more natural shapes for the vocal tract and generates more useful formant frequencies with the variation of parameters [Fant, 1960].

A similar model (fig 2a, next page) has been designed by Ishizaka using more parameters in [Flanagan, Ishizaka and Shipley, 1980]. In this model an asymmetrical tongue section has been allowed In addition, the area of the back cavity and the front cavity have been allowed to vary, unlike the constant area cavities used in the Fant model The Ishizaka model is given by.

$$A_b(x) = \frac{(A_b + A_c)}{2} - \frac{(A_b - A_c)}{2}\cos\left[\frac{\pi x}{l_b}\right] \qquad 2.5a$$

$$A_f(x) = \frac{(A_f + A_c)}{2} - \frac{(A_f - A_c)}{2}\cos\left[\pi\left[0.4 + 0.6\frac{x}{l_f}\right]\frac{x}{l_f}\right] \qquad 2.5b$$

where A_f and A_b are the maximum area in front of and behind the constriction, l_f and

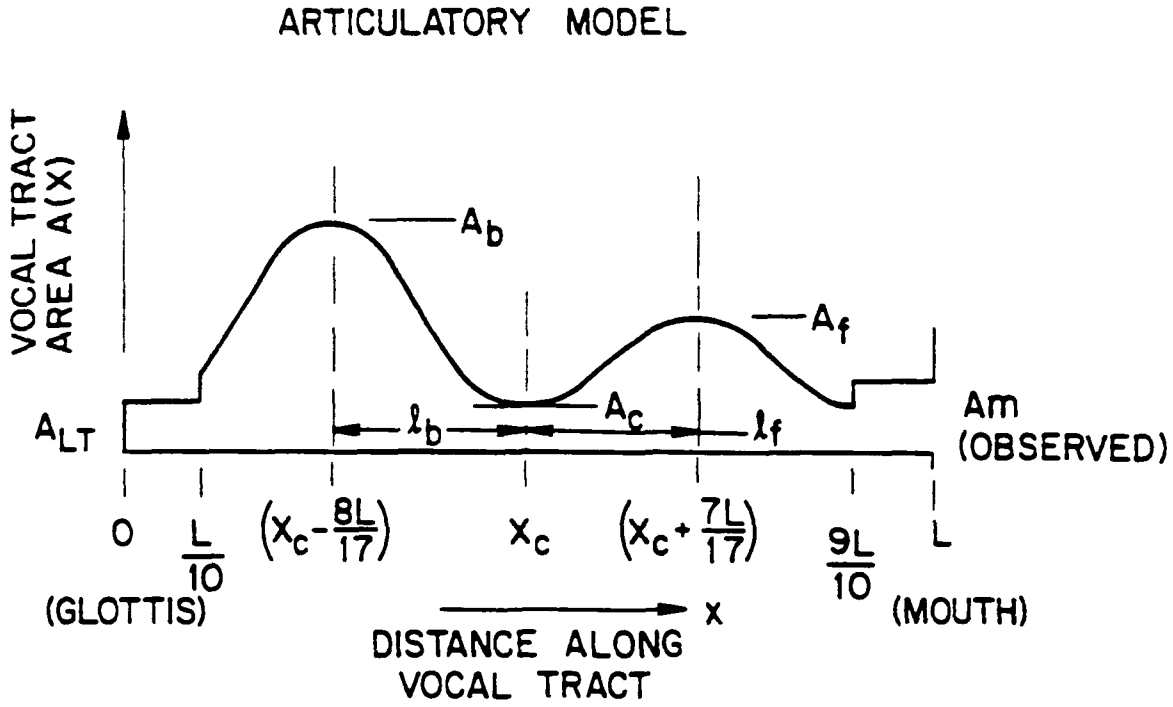


Fig. 2a **Ishizaka articulatory model defined by the eqn. (2.5). l_b and l_f are fixed at 8L/17 and 7L/17 respectively, where L is the vocal tract length (after [Flanagan et al., 1980])**

l_b are distances of A_f and A_b from the constriction.

A four parameter vocal tract model, developed on the basis of original three parameter model [Fant, 1960], is described in [Lin, 1990]. The models which exclude the nasal tract are in general more suited for vowel - like sounds than for consonants. However, to a first order of approximation, these models can also be used for consonants without apical modifications or a secondary posterior articulation [Lin, 1990].

2.3 A SIMPLE CONSTRAINED MATHEMATICAL MODEL FOR THE SPEECH PRODUCTION SYSTEM

A simple SPM has been proposed considering three principal parts of the human vocal apparatus the vocal tract, the glottis and the lip radiation sections. This SPM is used in the present work to estimate the vocal tract shapes and glottal flow parameters from the speech signals.

A modified version of the area function model described in [Lin, 1990] is used to represent the vocal tract cross sectional area of the speech production system. This constrained area function model was introduced in the SPM for the following reasons:

a) The use of a constrained area function model in the SPM produces/estimates only humanly attainable vocal tract shapes. This eliminates the problems faced in deriving the unique area functions from the LPC type models [Sondhi, 1979].
b) Vocal tract is described by a smaller number of parameters.

An equivalent digital filter realisation of the SPM is given in fig. 2.2, by converting the vocal tract area function into a chain of cylindrical tubes of equal length.

The glottis section of the system is modelled by a 2 - pole 2 - zero filter. This filter, obtained by lumping the losses due to wall vibration and source impedance gives realistic formant bandwidth at low frequencies (see also [Scaife, 1989]).

The radiation characteristics at the lips are modelled by a one pole-one zero impedance after [Laine, 1982].

Fig. 2.2a shows the equivalent lattice filter (Kelly - Lochbaum filter type) model of the system. Fig. 2.2b shows the inverse filter network obtained for the proposed reference model. U_G is the volume velocity at the glottis and P_L is the resulting speech

pressure at the lips. r_G and r_L are frequency dependent reflection coefficients at the glottis and lip end respectively.

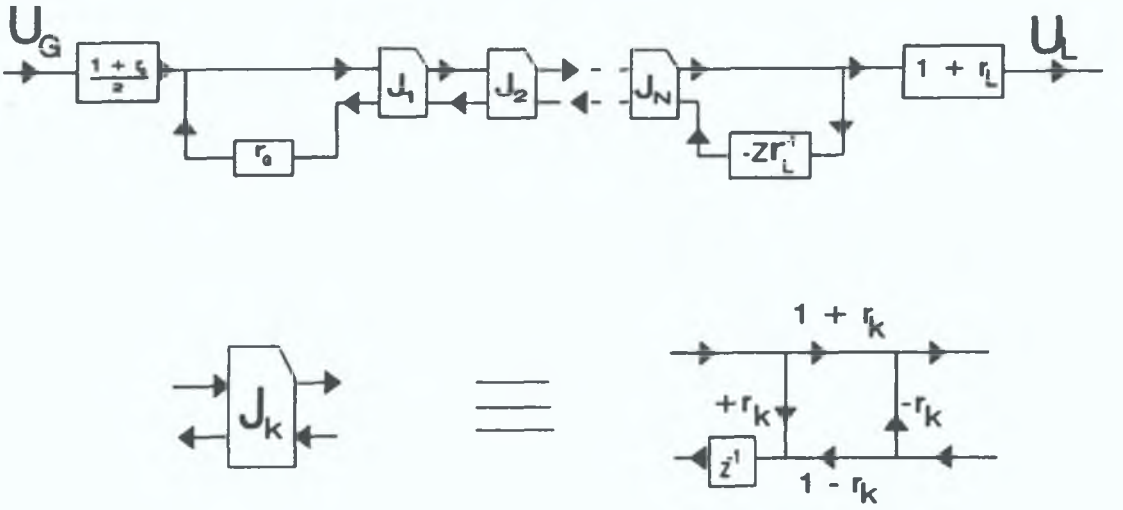


Fig. 2.2a Kelly-Lochbaum type synthesis filter model (after [Scaife,1989])

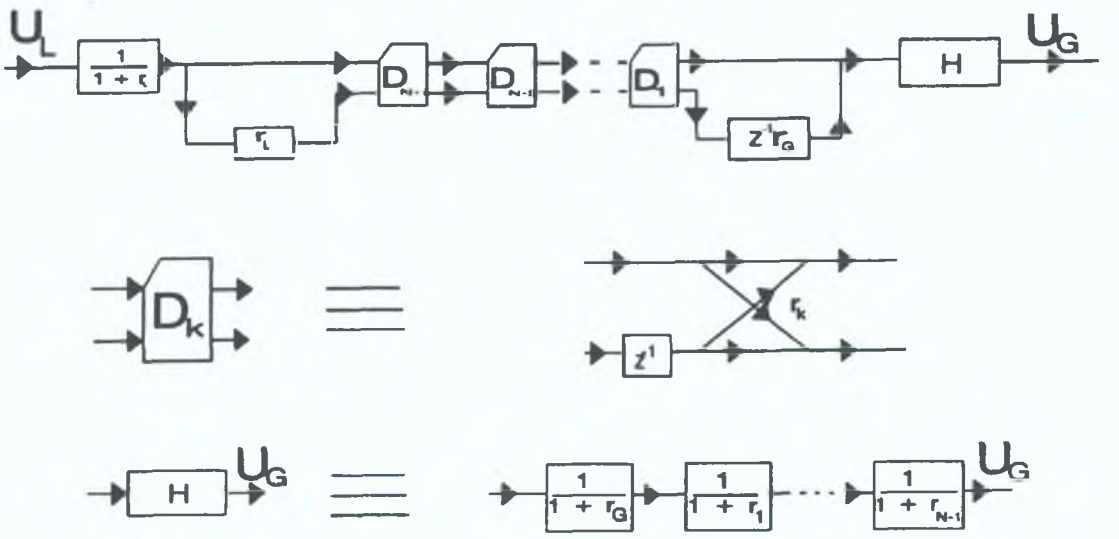


Fig. 2.2b Inverse filter model (after [Scaife,1989])

When r_G and r_L assume constant values the SPM reduces to an all pole network. When r_L takes the value of one (a complete reflection), the model reduces to the simple all pole LPC model [Wakita, 1973]. While this model is not structurally or computationally more complex than the LPC models, it does alleviate some of the

shortcomings of lossless all-pole models The shortcomings of LPC models are well known and discussed elsewhere in this thesis (see section 3.4.1) It is worth reiterating the fact that the proposed model is similar to the LPC model in the sense that the vocal tract is represented as a lossless stepped tube of equal length sections However, the model improves the formant bandwidth and closely matches the real vocal apparatus as some of the major loss mechanisms are included.

The results (formants and bandwidth) obtained here for six Fant vowels, compare well (Table 2 1) with those computed with more comprehensive loss SPMs [Badin and Fant, 1984] In these comprehensive loss models, all the main losses in the system (the radiation load, the viscous and thermal losses, wall vibration losses both distributed and lumped models, glottal and subglottal impedances and the constriction resistances) are considered and included

2.3.1 Description of the vocal tract section of the SPM

In [Lin, 1990], the vocal tract model has been made more flexible by the addition of variables to the original 3-parameter model [Fant, 1960]. The added features mainly control the symmetry of the tongue hump, the effective length of the constriction, the size of the larynx tube, piriformis cavities and the total vocal tract length. This model uses two different area functions depending on the type of vowels (back/front) being modelled. We have used a slightly modified version of this model. A linear combination of the two area functions is used during the transition from back vowel (X_c > 9 5 cm) to front vowel (X_c < 7.5 cm) In agreement with the original 3-parameter model [Fant, 1960], Lin also represents the lip condition by the ratio of the lip opening area to the lip protrusion length i.e; (A_0/L_0). However (A_0/L_0) values are estimated differently in [Lin, 1990] to avoid unrealistic A_0 values. In the present work, A_0 and L_0 are allowed to vary independently. Therefore the model used here, is constructed by five independent parameters X_c, A_c, A_0, L_0 and L_{tot} (see eqn (2.3) for the definition of the variables). In this model, the cross sectional area of the tongue hump $A(x)$ is given by:

$$A(x) = A_c + \frac{(A_f - A_c)}{H_f}\left[1 - \cos\frac{R_f \pi x}{X_f}\right] \qquad if\ X_c > X_2 \quad 2.6a$$

$$A(x) = A_c + \frac{(X_2 - X_c)}{(X_2 - X_1)} \cdot \frac{(A_f - A_c)}{H_f} \left[1 - \cos \frac{R_f \pi x}{X_f} \right]$$

$$+ \frac{(X_2 - X_c)}{(X_2 - X_1)} \cdot \frac{(A_f - A_c)}{H_f} \left[1 - \cos \frac{R_f \pi x}{X_f} \right]^2 \qquad \text{if } X_1 < X_c < X_2 \qquad \text{2.6b}$$

$$A(x) = A_c + \frac{(A_f - A_c)}{H_f} \left[1 - \cos \frac{R_f \pi x}{X_f} \right]^2 \qquad \text{if } X_c < X_1 \qquad \text{2.6c}$$

where the variables are defined as follows:

x : Displacement from X_c (along vocal tract centre line);

A_c : Constriction area of the tongue hump;

X_c : Distance of tongue hump constriction from the lip end;

and the auxiliary parameters A_f, H_f, X_f and R_f control the general shape of the tongue. Also the constants X_1, X_2 are defined as $X_1 = 7.5$ cm and $X_2 = 9.5$ cm for the current work.

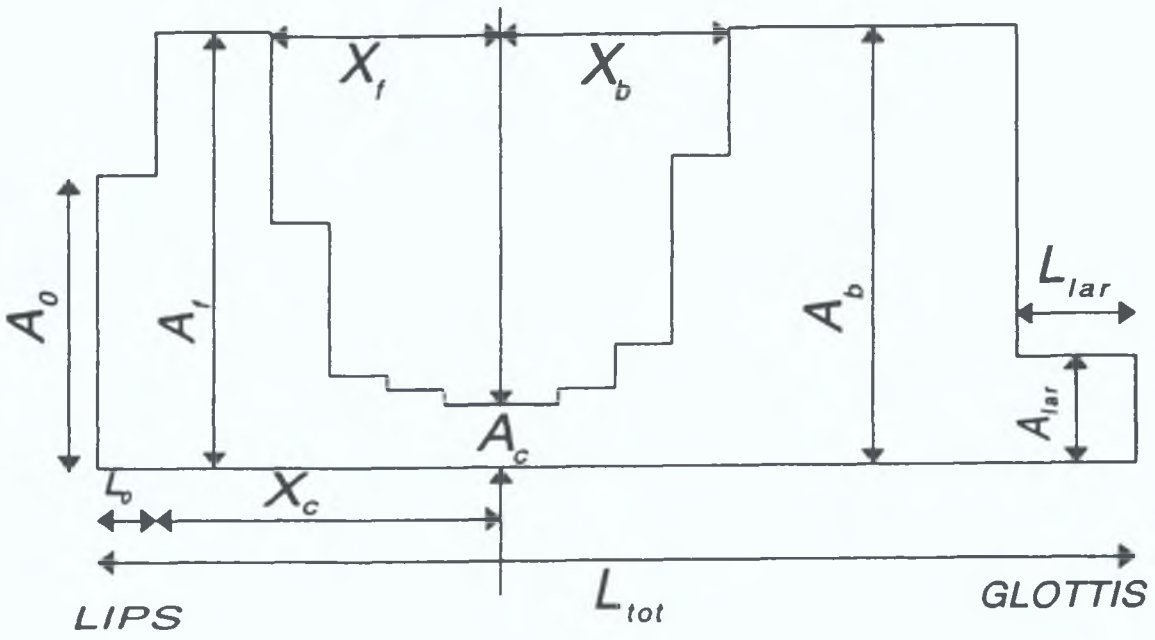


Fig. 2.3 A typical vocal tract cross sectional area generated by the eqn. (2.6). A_f and A_b are fixed at 8 cm^2. The values of the subsidiary parameters and the independent parameters A_c, X_c, A_0, L_0 and N are given in page 26.

The area function of the vocal tract from the glottis to the lips is completed by adding lip protrusion to the lip end and larynx cavity to the glottis end The larynx cavity in the model is represented by a constant tube of length 2.2 cm (i e, two small tubes) with a cross sectional area of 2 cm^2. Therefore, in addition to the parameters mentioned above, the parameters A_0, L_0, A_{lar}, L_{lar}, L_{tot} and N are added to the model The parameters are defined as follows: A_0 . Area of lip opening; L_0 . Length of lip protrusion beyond the teeth; A_{lar} . Area of the larynx tube; L_{lar} : Length of the larynx tube and L_{tot} : Total length of the vocal tract given by N (the number of small vocal tract tubes) multiplied by the length of the small cylindrical tube

The modified area function model has been implemented as a *"C"* program The flow chart of the program is given in appendix B2. The analog area function is quantised to a chain of tubes of equal length 11 mm (given by $c/2*f_s$) at 16 kHz sampling rate. The quantised area function is then incorporated into the SPM

The modified model has five *independent* parameters· X_c, A_c, L_0, A_0 and L_{tot} R_f , A_f and H_f control the general shape of the tongue hump With the substitution of $\{A_b, X_b, R_b, H_b\}$ for $\{A_f, X_f, R_f, H_f\}$ the same expressions in eqn (2 6) yield the shape behind the tongue hump. It is important to note that R_f, A_f and H_f are not independent parameters The values for these subsidiary parameters are fixed at . $A_f = A_b = 8\ cm^2$, $H_f = H_b = 1$, $X_f = X_b = 4\ cm$, $R_f = R_b = 0.4$ *and* $X_1 = 7.5\ cm$, $X_2 = 9.5\ cm$ An N section tube quantised area function generated with $A_c = 0.96$, $X_c = 6.45$, $A_0 = 5.6$, $L_0 = 1\ 2$, *and* $N = 18$ is shown in Fig. 2.3. The reflection coefficients that will be obtained using the cylindrical tube area function are discussed in appendix B1.

2.3.2 Description of the glottis section of the SPM

The glottis is the orifice between the vocal cords in the subglottal system. As we mentioned, the steady air flow from the lungs is modulated at the vocal cords. This modulation mechanism i.e; the vocal cords and its associative organs can be more quantitatively described in terms of an equivalent electrical network Equivalent acoustic impedance values for the electrical circuit are estimated from the knowledge of data available on subglottal pressure, glottal dimension and glottal area function

Schematic diagrams of the human subglottal system and its equivalent circuit are shown in Fig 2 4 The operation of the system is clearly explained in [Flanagan, 1972] Very briefly, because of the mass and the elastic properties, the vocal cords

vibrate during the pressure variation in the glottis This vocal cord section (glottis) can be represented as a series impedance ($R_g + j\omega L_g$) in the electrical equivalent circuit As can be seen from Fig 2 4 the pressure P_s at just before the glottis area is converted to glottal volume flow U_g. This time varying glottal impedance which is very much related to the glottal area characteristics is discussed in the next section

2.3.2.1 Glottal impedance

For steady flow conditions through the glottis, the acoustic impedance of the glottis is given by·

$$Z_g = R_g + j\omega L_g \qquad 2\ 7$$

Where the acoustic resistance of the orifice (glottis) R_g is calculated as the ratio of pressure drop across the orifice to the volume flow through the orifice By considering the effects of viscosity and nonuniform velocity distribution of the air, the resistance R_g is expressed as

$$R_g = R_v + k\left[\frac{\rho U}{2A^2}\right] = R_v + kR_k \qquad 2.8$$

where R_v is viscous resistance and k is a real constant. R_k is the kinetic resistance of the glottis with the assumption of uniform velocity distribution of air at the glottis orifice.

For steady laminar flow, using Van Den Berg's data [Flanagan, 1972], the value of R_g in eqn (2 8) can be expressed as:

$$R_g = \frac{P_s}{U} = \frac{12\mu d}{lw^2} + \frac{0\ 875\rho U}{2(lw)^2} \qquad 2\ 9$$

where l - length of the glottis, w - width, P_s - Pressure difference across the glottis; $U = u.A$ - volume velocity

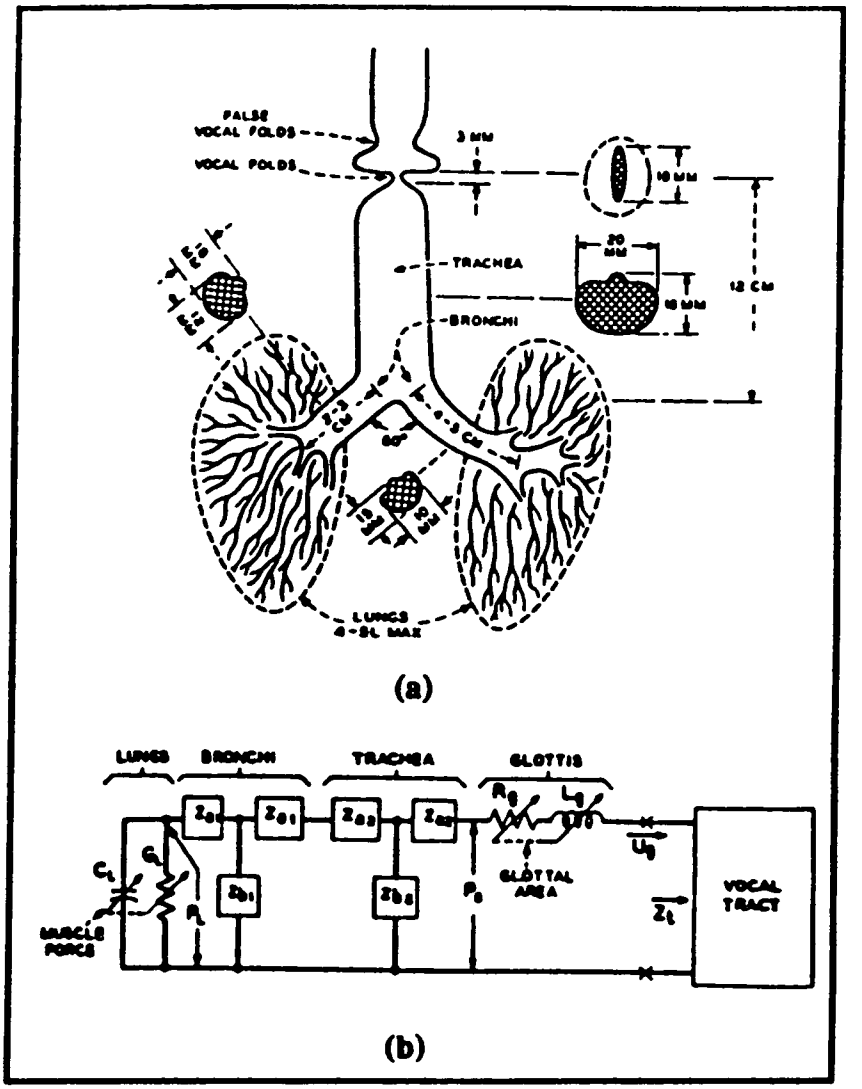


Fig. 2.4. a): Human sub glottal system

b): An equivalent circuit for the sub glottal system (after [Flanagan, 1972])

The glottal resistance R_g is either dominated by the kinetic term R_k or by the viscous term R_v during each cycle (eqn. (2.8)), depending on the values of the area and the subglottal pressure They become equal when $(\rho P_s)^{1/2}A^2 = 19.3\mu dl^2$

For typical values of P_s, it is found that for most of the open cycle of the vocal cord (for area > 1/5 of maximum glottal opening area) glottal resistance is determined by the kinetic term R_k

The inductance L_g due to the elastic properties of the glottis is given by.

$$L_g = \frac{\rho d}{A_g(t)} \qquad 2.10$$

Therefore the glottal impedance Z_g is given by:

$$Z_g = R_g + j\omega L_g \qquad 2.11$$

In the non-interactive modelling of speech production system, it is assumed that the glottal impedance Z_g is large compared to the vocal tract input impedance. This assumption allows the airflow through the glottis to be estimated (without considering the vocal tract loading) using the impedance Z_g alone by short circuiting the vocal tract i.e. $Z_t = 0$ in fig. 2.4b. From fig. 2.4b, we can write

$$U_g(t)R_g(t) + \frac{d(L_g(t).U_g(t)}{dt} = P_s \qquad 2.12$$

Using the above equation together with reasonable data (e.g. 4 cm $H_2o < P_s <$ 16 cm H_2o, $\rho = 0.00114$ g/cm^3, etc.) for the human vocal apparatus, it can be shown [Flanagan, 1972, pp. 43-49] that the time constant of the above equation is negligible compared to the fundamental vocal cord period. Using this fact it can be shown [Flanagan, 1972] that to a first order approximation, the glottal flow can be estimated from P_s and $A_g(t)$ using the eqn. (2.9).

However, in reality the vocal tract load has significant effect on the glottis at formant frequencies and at high frequencies. As a result, glottal flow is skewed at the closing phase and hence is not same as the glottal area variation. The skewing depends on the input inductance of the load. The effect of this skewing in the frequency domain is to uniformly increase the level of all formant frequencies. This vocal tract source interaction has been discussed and to a certain extent eliminated, by the interactive source-filter modelling of speech [Nord, Ananthapadmanabha, and Fant, 1984].

From the above discussions we conclude that the glottis can be represented by a resistor in series with an inductor circuit for non-interactive modelling and the flow can be calculated from $A_g(t)$ and P_s alone.

2.3.2.2 Model at glottal end termination for the SPM

The parallel combination of a simple resistor inductor (RL) model (see Fig. 2.5)

discussed above for the glottis (Z_g) and the wall impedance model (Z_w, see section 2.3 3) is used to model the glottal end termination This 2-pole 2-zero digital model obtained as the result, is discussed in [Scaife, 1989] The derivation of this digital model (Z_g) with the values used for the glottal impedance (R_g, L_g) and the wall impedance (R_w, L_w, C_w) are given in appendix B3

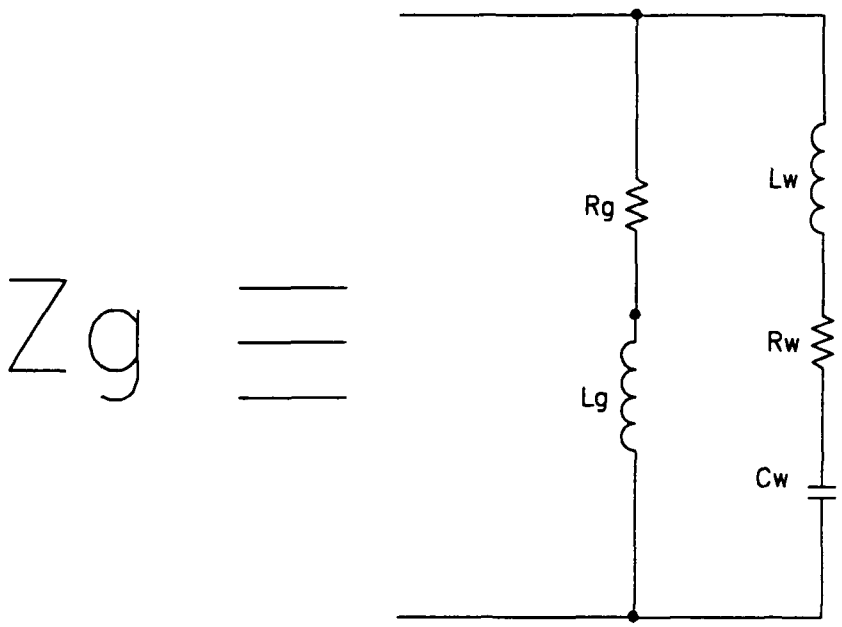


Fig. 2.5 Composite termination at the glottal end, modelled by the parallel combination of wall and glottal impedance

The inclusion of this model (Z_g) at the glottal end of the current SPM improves the modelling of the system at low frequencies The frequency dependent reflection coefficient obtained at the glottal end r_G is given as.

$$r_G = \frac{\alpha_{g0} + \alpha_{g1} z^{-1} + \alpha_{g2} z^{-2}}{1 + \beta_{g1} z^{-1} + \beta_{g2} z^{-2}} \qquad 2\ 13$$

where the variables area defined as

A_{lar}: Area of the first vocal tract tube from the glottis,

ρ· is the density of air ;

c . velocity of sound and α_{0g}, α_{1g}, α_{2g}, β_{1g} and β_{2g} are constants depending on A_{lar}, sampling frequency, etc (see appendix B3)

2.3.2.3 Effects on the formant frequencies and bandwidths of varying glottal area

The effects of changing R_g on the frequency response of the SPM are studied here. R_g is increased from 30 Ω to 150 Ω in steps of 15 Ω i.e; glottal opening area A_g is decreased from high to low. Spectra of six vowels for different values of R_g are shown in Fig. 2.6a to Fig. 2.6f.

As expected, when A_g is increased the amplitudes of the formant are decreased and the bandwidth are increased, due to the increase of losses at the glottis. As can be seen from the figures in general, the influence of R_g is more prominent on the lower formants than the higher formants. The effects are more visible on the back vowels */a/*, */o/* and */u/*. This is because the first two formants of these vowels are close to each other. The increase in bandwidth for the first two formants, causes the convergence of these two formants.

As a result of this study, a value of $R_g = 110$ Ω was chosen as the optimum value (i.e; the value which gives better formant frequencies and bandwidths) for the current model.

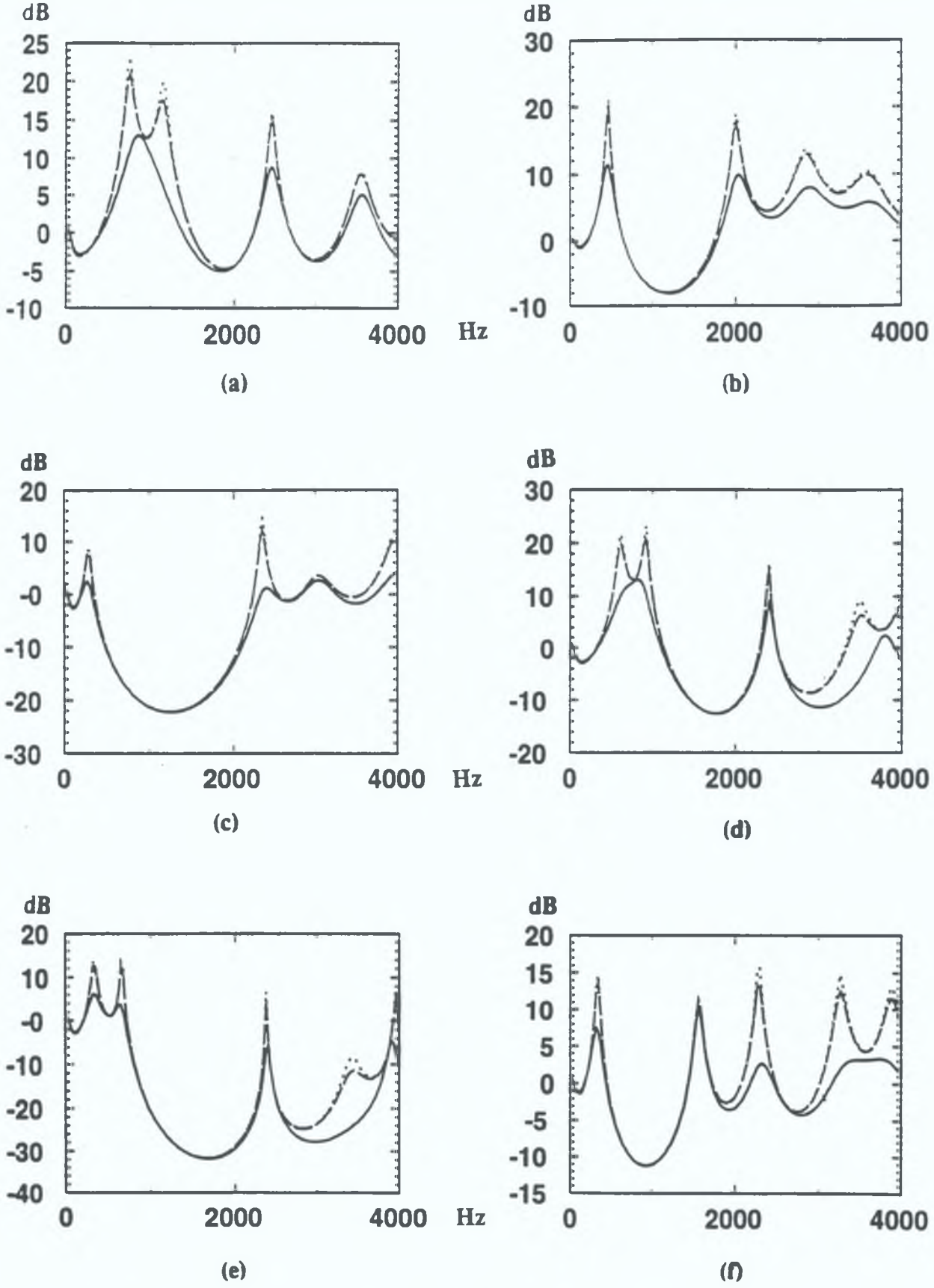
2.3.3 Wall vibration effects

The vocal tract walls vibrate during the propagation of the speech waveform along the vocal tract. The result is to shift (increase) the low frequency formants of the vocal tract.

The shift is related to the comparatively large acoustic mass of the walls effectively shunting the mass of the air in the tract. In wall impedance modelling, this effect is accounted for by an inductor L_w. The loss of energy absorbed in the tissue and partly reradiated through the skin is accounted for by a resistance R_w. A capacitive element is also included in the wall impedance model to include the effect of closures of the vocal tract. In [Liljencrants, 1985], z domain modelling of wall interaction was discussed and digital representations were given.

The wall vibration effects discussed above occur along the vocal tract. Therefore, earlier speech models uniformly distribute the vibration effects (wall impedance) along the tract [Flanagan, 1972; Fant, 1972; Ishizaka et al., 1975; Wakita and Fant, 1978].

Experiments however showed [Badin and Fant, 1976] that the maximum vibrational amplitude occurs at about 4 cm above the larynx with a somewhat weaker



——: R_g = 30 Ω; —·—: R_g = 110 Ω; ·····: R_g = 135 Ω;

Fig. 2.6 Transfer function of the SPM with varying R_g (A_g) values.

a): Vowel *[a]* b): Vowel *[e]* c): Vowel *[i]* d): Vowel *[o]* e): Vowel *[u]*

f): Vowel *[ɨ]*. X- ray vocal tract cross sectional area function values are used.

maximum near the cheek area. This suggests that the wall vibration effects (the wall impedance) can be approximated by placing the equivalent models at the glottis and lip end.

In the present work, a digital model for the wall impedance described in [Liljencrants, 1985] used A parallel combination of this wall model and glottal impedance is placed at the glottal end of the SPM.

2.3.3.1 Effects of placing the wall impedance model at different point along the vocal tract

Using the distributed lossy SPM designed in appendix B4 (a set of reflection coefficients similar to those of lossless vocal tract for the model were derived. The losses in the tract were considered as shunt and series loss factors [Liljencrants, 1985]), the effect of placing the wall vibration losses at different points of the vocal tract is studied

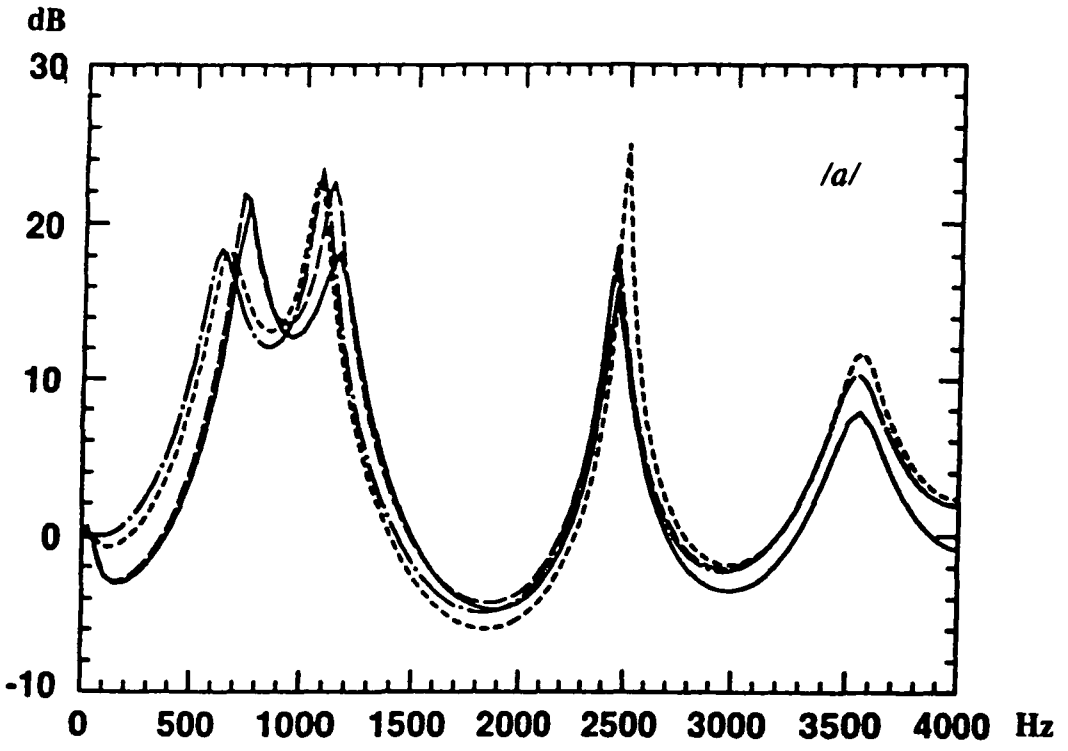


Fig. 2.7. Effects of placing the wall impedance at different points along the vocal tract for vowel /a/. ——: lumped parametric model; – – – –: wall impedance at 1.1 cm from the glottis; - - - - - - - -: 6.6 cm from the glottis; —·—·—: 12.1 cm from the glottis.

Fig. 2.7 shows the spectra for vowel */a/* with wall impedance model placed at 1 1 cm, 6 6 cm, 12 1 cm from the glottis and at the glottis (i e wall vibration effects are lumped at the glottis, current SPM).

Consistent with the previous research findings, placing the wall effects in the middle of the vocal tract does not affect the lower formant frequencies (i e same as the constant coefficient glottal end termination) Whereas, placing the losses near the glottis or lumping the effects at the glottis increases the lower formant frequencies The graphs clearly show that placing the wall losses at 1.1 cm and at the glottis produces similar low formant frequencies

Fig 2.8 shows the spectra of */a/* and */u/*, with and without the z - domain wall model. Inclusion of the wall model, increases the lower formant frequencies (Fig 2 8) The effect is same for all the vowels tested

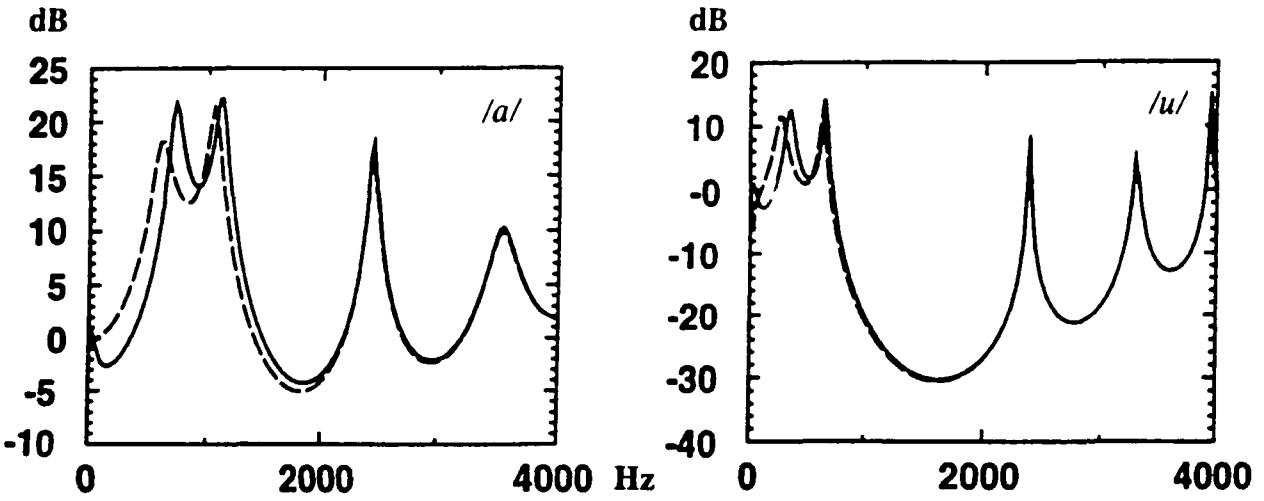
The result of excluding the phase angle of the glottal termination model is shown in Fig 2.9 The difference in lower formants, between the curves is not surprising as excluding the phase angle is equivalent, to modelling the glottal end by a constant coefficient

As a result of this study, we conclude that simply lumping the wall model at the glottis improves the modelling of the speech production system This enables the vocal tract to be treated as rigid and smooth in the modelling of the speech production system Including the wall vibration effects into the rigid and smooth vocal tract model, as mentioned, has the effect of producing formant and bandwidth values that compare more realistically with a true vocal tract

Formant and bandwidth values of the model, for vowels with and without wall impedance are given in table 2.2. The table also shows the formant and bandwidth values estimated from the more comprehensive speech models [Badin and Fant 1984] The table indicates that the inclusion of wall vibration effects in the model produces formant and bandwidth values which are close to the values obtained from the more complex models

2.3.4 Description of the lip radiation section

In the production of speech the volume flow delivered to the lips or/and to the nose is radiated as a pressure waveform at these points A widely used analogy is to represent this effect as an orifice in a sphere The orifice is considered as the radiating surface

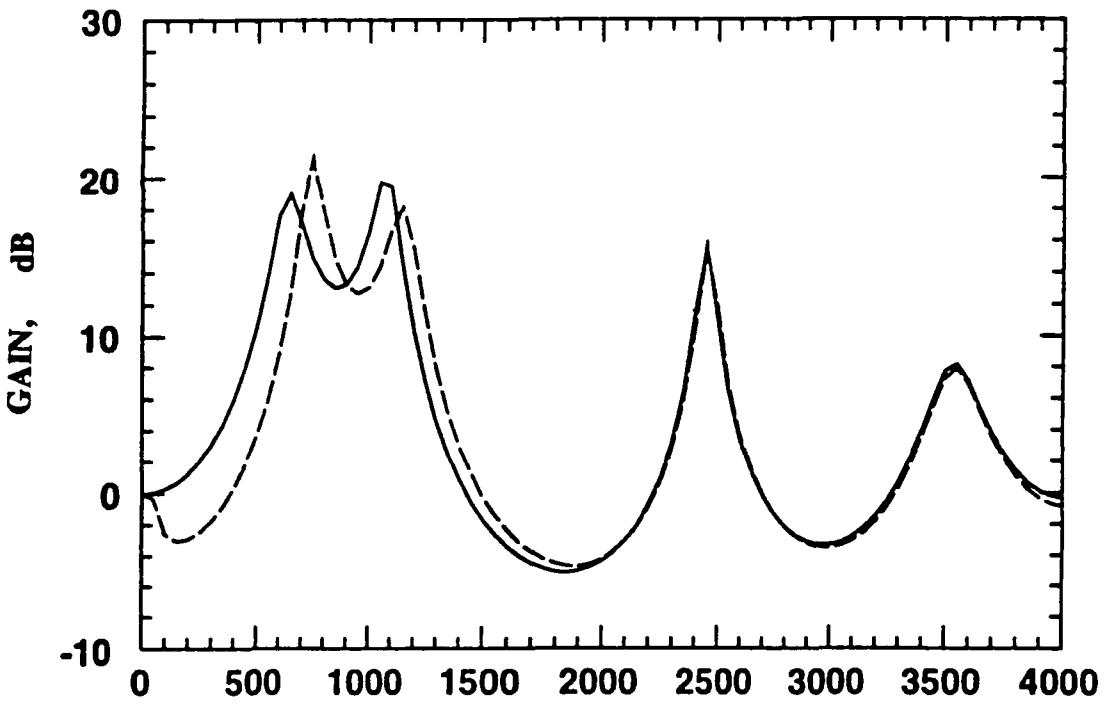


FREQUENCY, Hz

Fig. 2.8 Transfer function of the model with and without the wall model.

a): Vowel /a/ b): Vowel /u/

———: with wall model ; ---------: B, without wall model;



FREQUENCY, Hz

Fig. 2.9 Transfer function of the current SPM for vowel /a/.

_____: without the phase angle of the glottal model

---------: with the phase angle of the glottal model

Vowels		F_1 (B_1) Hz	F_2 (B_2) Hz	F_3 (B_3) Hz	F_4 (B_4) Hz
/a/	A	744 (88)	1147 (136)	2458 (73)	3544 (221)
	B	612 (121)	1057 (73)	2443 (56)	3536 (205)
	C	766 (94.8)	1127 (106 5)	2473 (83.6)	3629 (185 4)
/e/	A	466 (38.5)	2003 (83)	2842 (280)	3586 (501)
	B	419 (30)	1969 (62)	2808 (224)	3562 (406)
	C	460 (35.5)	1993 (94.3)	2834 (224.6)	3650 (408 1)
/i/	A	285 (61)	2348 (62)	3045 (403)	3934 (231)
	B	212 (27)	2307 (26)	3048 (400)	3948 (153)
	C	287 (48.3)	2293 (81.8)	3133 (204.3)	3723 (237 3)
/o/	A	620 (82)	922 (61)	2395 (32)	3512 (257)
	B	498 (104)	879 (37.5)	2384 (251)	3389 (45)
	C	604 (68.3)	901 (59.4)	2394 (42.8)	3494 (246 2)
/u/	A	337 (54 5)	660 (31)	2381 (13)	3440 (*)
	B	258 (38)	627 (21)	2370 (10)	3298 (42)
	C	313 (52)	633 (36.9)	2385 (26.3)	3797 (172 3)
/ɨ/	A	341 (37)	1556 (77)	2289 (71)	3274 (130)
	B	282 (27)	1553 (75)	2242 (47)	3225 5 (53)
	C	340 (40 9)	1543 (78.3)	2422 (60.2)	3476 (99 4)

Table 2.1. Formant and bandwidth data for six Russian vowels.

A: The SPM employed in the present research. Wall losses are placed at the glottis.

B: The SPM without the wall losses

C: SPM used in [Badin and Fant, 1984]. This SPM includes viscous, thermal, wall, radiation and glottal losses.

*** Bandwidth B4, of vowel /u/ is hard to determine as the gain at this point is very low and hence make the measurement unreliable.**

with the radiating sound waves being diffracted by the sphere or the wall at low frequencies. This model is termed as Piston In Sphere (PIS) model [Flanagan, 1972].

Morse has derived equations for the equivalent radiation load and shown that the load is a function of frequency. The actual equations involved in the PIS model are computationally difficult and cannot be expressed in a closed form. However, when the radius of the piston becomes small compared to the sphere, the model approaches to a piston in a wall model. The acoustic impedance of this model can be derived by relating the pressure waveform to the volume velocity at the lips. For small values of *ka* (described in eqn. (2.14)), using some approximations the impedance can be expressed in terms of resistance and reactance part as:

$$Z_l = \frac{(ka)^2}{2} + \frac{j8(ka)}{3\pi} \qquad for \quad ka \ll 1 \qquad 2.14$$

where $k = \omega/c$ and a is the piston radius.

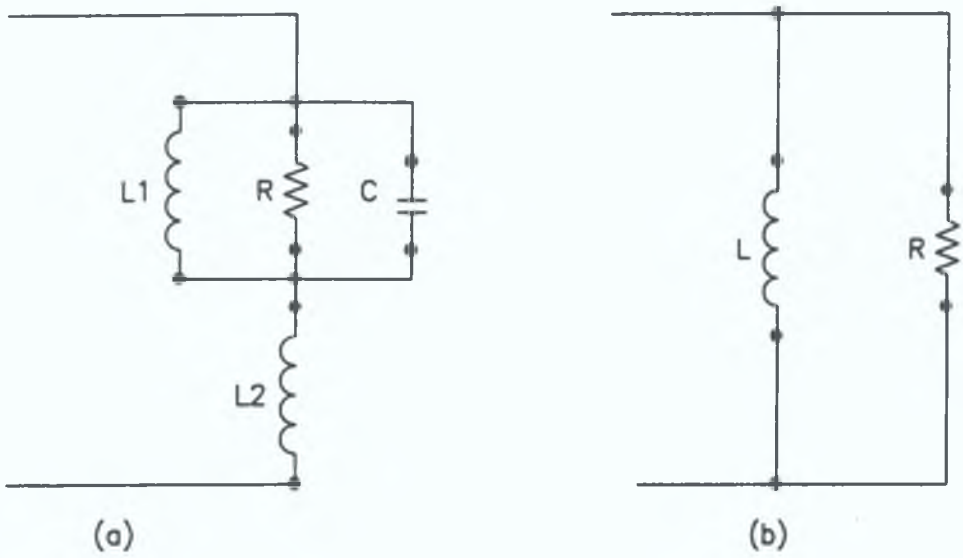


Fig. 2.10 Electrical equivalent circuit for the lip radiation.
a): Stevens, Kasowski and Fant model b): Flanagan model

As mentioned, estimating the radiation impedance using the PIS model is computationally inefficient. Therefore several simple equivalent models were proposed in terms of resistive (R) and reactive (L,C) components. Two widely used models which accurately match the PIS model are Flanagan's RL model [Flanagan, 1972] and

Stevens, Kasowski and Fant's RLC model (SKF model, see Fig. 2 10). It was clearly seen [Lin, 1990] that the match of these two analog models to the PIS model is indistinguishable up to 5 kHz in both the R and L domains.

Several digital domain models were proposed in [Laine, 1982] by minimising the mean square error between the model.

$$Z(\omega) = C(1 - \cos(\omega T)) + B.j.\sin(\omega T) \qquad 2\ 15$$

and the acoustic impedance (eqn. (2.14)) of the PIS model. By means of matching the following values, depending on the lip opening area A, are obtained for C, T and B

$$C = 0\ 674 - 0.00854A\ , \qquad T = \sqrt{(A/54)}\ , \qquad B = 0\ 674$$

By substituting $\cos(\omega T) - j\sin(\omega T) = e^{-j\omega T} = z^{-1}$ (i e. $2.\cos(\omega T) = (z^{1} + z^{-1})$ and $2.j\sin(\omega T) = (z^{1} - z^{-1})$) in eqn (2.15), Laine obtained three z domain models In the z domain models he fixed the value of T, by sacrificing some accuracy in the matching Laine found the values for C and B as .

$$C = C_1.A \qquad \& \qquad B = B_1 \sqrt{A}$$

with C_1 & B_1 being constants depending on the sampling frequencies

By giving equal values to B & C, the following first order model Z_1 was obtained:

$$Z_1 = C(1 - z^{-1}) \qquad 2\ 16$$

A 2nd order model was also proposed by giving proper values to the C_1 & B_1. This improves the modelling of the reactive part.

Laine also proposed a pole/zero model by matching the model to the PIS model. Due to the one more degree of freedom of this model, reactance which controls the formants, is matched better than the previous two models. This pole/zero model, used in the current SPM, is given by .

$$Z_l = \frac{\rho c}{A}.a.\frac{(z-1)}{(z-b)} = \frac{\rho c}{A}.a.\frac{(1-z^{-1})}{(1-bz^{-1})} \qquad 2.17$$

where a & b are a function of lip opening area A and are given by:

$$a = 0.0779 + 0.2373\sqrt{A} \; ; \qquad b = -0.843 + 0.3062\sqrt{A}$$

Laine assumes the radius of the sphere to be 9 cm which is about the average size of the human face. However, a wide range of lip areas and sampling frequency give unacceptable values for a and b and therefore a linear interpolation from a pre-computed table is used for the values a and b [Scaife, 1989].

Both the analog (Flanagan and SKF) and digital (Laine's) models mentioned above approximate the PIS model well at small lip opening area. This is expected as at small piston radii the approximation approaches the piston in a wall model and hence the derived equations closely match the original equation.

2.3.4.1 Incorporating the lip radiation model into the SPM

The pole/zero model mentioned above is incorporated at the lip end of the SPM. Therefore the reflection coefficient r_L at the lip, is a frequency dependent digital model and is given as:

$$r_L = \frac{\rho c/A_N - z_l}{\rho c/A_N + z_l} \qquad 2.18$$

where Z_l is defined in eqn. (2.17).

2.3.4.2 Conclusion

We conclude that the pole/zero model provides a better match for formants at high frequencies. Also the reflection coefficient at the lip using this model can be efficiently calculated. Therefore this model was chosen to represent the lip radiation characteristic.

2.4 COMPLETE MODEL

The SPM developed in the research is a normal lossless acoustic tube model. This model is similar to the LPC model, however the glottal and lip end are terminated

by frequency dependent networks The glottal end is terminated by a parallel combination of the wall and the glottal impedance The lip end is terminated by Laine's lip radiation model A detail description of the wall, glottal and the lip radiation models was given in section 2.3.

By using Fant area function data values of six Russian vowels in the estimation model we showed the relevance of the model. We showed that this model produces formant and bandwidth values which are very much comparable (Table 2.1) with those produced from the comprehensive models [Badin and Fant, 1984].

In section 2 3.2 we have modelled the vocal tract using the Lin-Fant area function model In the next chapter, we employ this SPM to analyse real and synthetic vowels. We also, discuss in detail the various ways of estimating the parameters of this model for speech signals However, to indicate the success of the model to the reader, we show the matches obtained for original and estimated spectra of six vowels in Fig 2.11 X - ray area function data of the six Russian vowels are used for the above test

This simple SPM designed (see also [Scaife, 1989]) above has been used to analyse and synthesize vowel like sounds throughout this research.

2.5 CONCLUSION

Functions of various parts of vocal apparatus were considered and equivalent digital models were discussed In section 2.3, a lumped parameter SPM was described Vocal tract section of the speech production system was modelled by a constrained area function model [Lin, 1990]. The effects of including the wall impedance at different points along the vocal tract are discussed. A parallel combination of wall impedance and glottal impedance is used to model glottal end termination Laine's pole - zero model is used to model the lip radiation characteristic The SPM clearly shows the relationship between the parameters A_c, X_c, A_0 with F_1 & F_2 formants (see chapter 3) The SPM is an extended version of the LPC model, with glottal and lip end are modeled by frequency dependent network. However, the inclusion of the loss mechanisms wall model, radiation model and glottal model improves the modelling of the system over the LPC modelling. This SPM is being used for the analysis of speech signals throughout the thesis

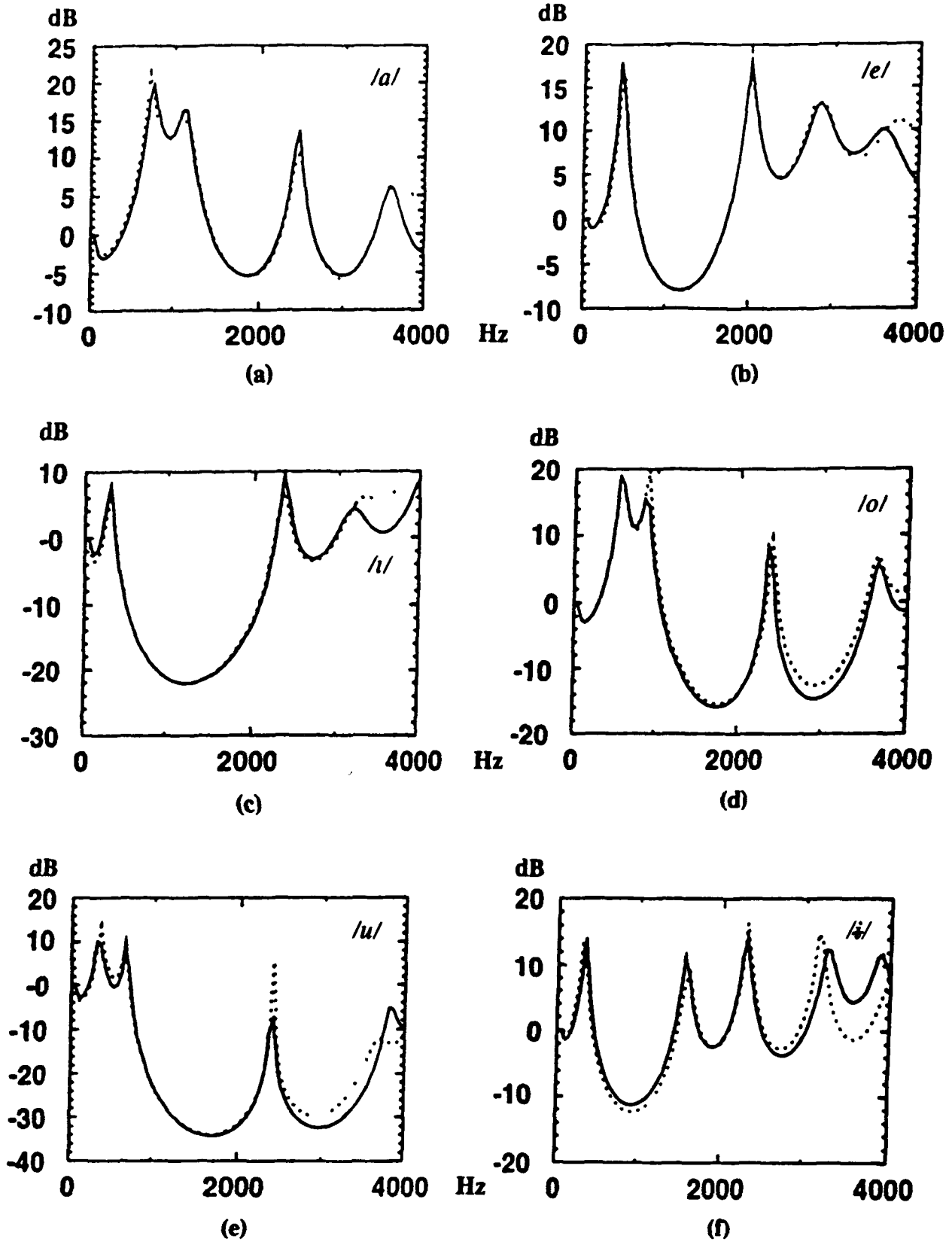


Fig. 2.11 The original and the estimated spectra of six Russian vowels.

a): Vowel */a/* b): Vowel */e/* c): Vowel */i/* d): Vowel */o/* e): Vowel */u/*

f): Vowel */ɨ/*. X- ray vocal tract cross sectional area function values are used.

_____: original; .....: estimated;

CHAPTER 3

ESTIMATION OF VOCAL TRACT PARAMETERS FROM ACOUSTIC DATA USING LINKED CODEBOOK

3.0 INTRODUCTION

This chapter describes methods to estimate, from the speech signal, vocal tract parameters for the analysis/synthesis model designed in chapter 2

Initial attempts at estimating the model parameters using optimisation procedures are discussed in sections 3.1, 3.2 and 3.3 Because of the highly non-linear mapping of acoustic to articulatory parameters, optimisation algorithms may choose local minima of the cost function and hence produce incorrect values for the parameters This is overcome by restricting the search area to the vicinity of the global minimum Discussions on how to reduce the search area and the results obtained using this method are given in section 3.2 and 3.3

Extending these initial ideas and methods lead us to generate a large acoustic to articulatory codebook for the model The technique, used here is similar to the linked codebook technique, firstly introduced in [Schroeter, Larar and Sondhi, 1987] Since then, the technique has been successfully applied to different articulatory models for analysis/synthesis of speech [Larar et al., 1988; Schroeter and Sondhi, 1989, Schroeter et al., 1990].

Section 3 4 discusses the advantages of the codebook techniques by referring to speech coders Section 3.5 discusses the generation of a linked codebook for the current model Section 3 6 discusses the estimation of parameters for given segments of a number of real utterances and synthetic vowels using the codebook The results and possible future improvements to the techniques are discussed in the rest of the chapter

It should be noted that only half of the estimation of the complete model, namely the estimation of the vocal tract parameters, is described in this chapter The estimation of the remaining half of the model parameters i e the glottal flow/source parameters is a major study involving inverse filtering, and is described in chapter 4 A block diagram of the current research is shown in fig 5 2

3.1 INITIAL ATTEMPTS AT ESTIMATION OF PARAMETERS

The estimation of parameters of any model involves matching the real data with the model data The parameters of SPMs in speech research are generally estimated by matching some form of speech signal with the model signal. Different techniques use

different error criteria to estimate the parameters either in the time or the frequency domain There are several error functions e g mean square error in time domain, squared spectral distance, likelihood ratio, mean square cepstral differences etc , available for estimation purposes depending on the requirements [see appendix C5]

3.1.1 Estimation of parameters using the direct optimisation approach

Initially, a direct optimisation algorithm was used to estimate the parameters The "squared log spectral distance" between the model and the input speech is chosen as the cost function [Papamichalis, 1987] The "simplex" method of optimisation, which does not require estimation of derivatives [Press, et al , 1988], is used to estimate the five parameters of the model. This method requires $(n+1)$ starting points for an n - dimensional search (see appendix C1).

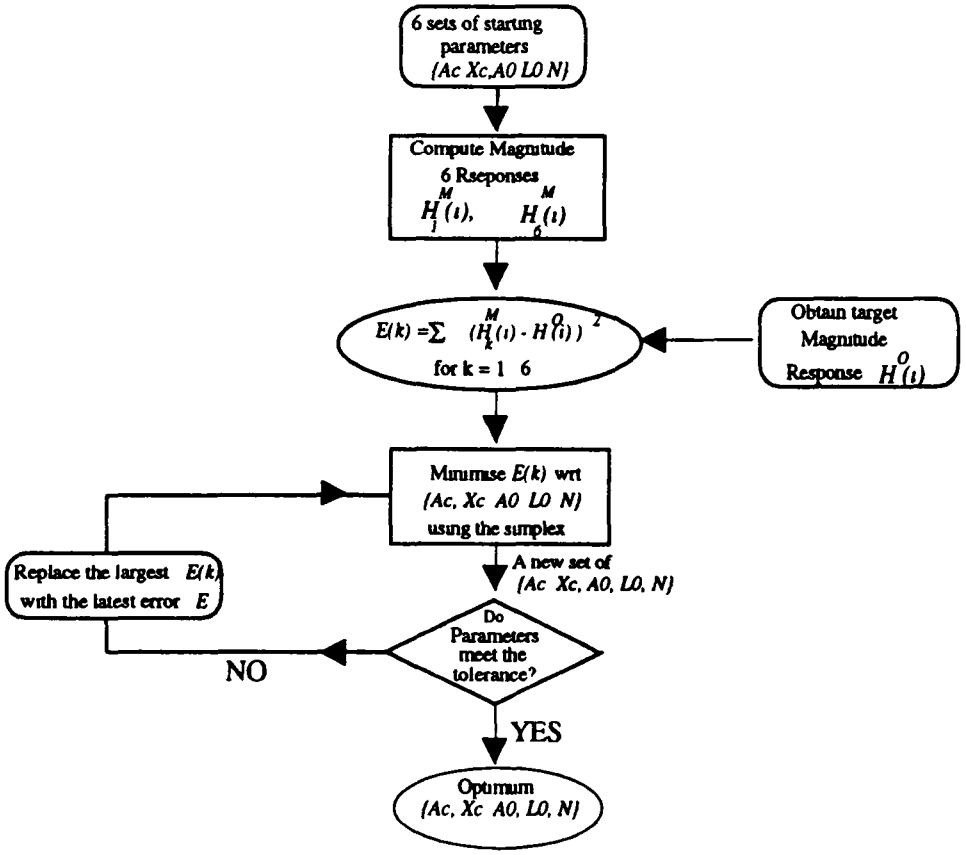


Fig. 3.1 Estimation of the area function parameters A_c, X_c, A_0, L_0 and N using the simplex method of optimisation

The estimation methods and algorithms were tested on magnitude responses i.e.

spectra of the impulse responses, of six Russian vowels */a/, /e/, /i/, /o/, /u/* and */ɪ/* estimated using the SPM (fig 2 2a) The magnitude response was estimated by substituting $z^{-1} = e^{-j\omega t}$ in the model transfer function of eqn. (B1 26) derived in appendix B1 X-ray area function data [Fant, 1960] of six vowels quantised to a tube length appropriate to a sampling rate of 16 kHz were used in the estimation. This magnitude response was then considered as the target function for the optimisation algorithm.

The optimisation of five parameters using the "simplex" method is described in the following paragraphs (see also fig 3 1).

a) Estimate the original magnitude response $|H^O(i)|$ where $i=1$. 80, of the SPM for a given area function using the synthesis model (Fig. 2.2a) over 0 to 4 kHz at 50 Hz steps i e. a vector of 80 spectral parameters to represent the magnitude response This $|H^O(i)|$ is the target magnitude response for the optimisation The flow chart to to estimate the magnitude responses is given in appendix C2

b) Since an n - dimensional search of simplex method requires $(n + 1)$ starting points, initialise the optimisation algorithm with six different sets of magnitude responses $|H_1^M(i)|$, $|H_2^M(i)|$ . . $|H_6^M(i)|$ where $i=1$. 80, of the model Random set of values were given to the five area function parameters $\{X_c, A_0, A_c, L_0, N\}$, to estimate the above six starting points. The six magnitude responses $|H_1^M(i)|$, $|H_2^M(i)|$ . . $|H_6^M(i)|$ were then used in the calculation of the errors $E(k)$ (defined below) at these points

c) Minimise the cost function $E(k) = \sum_{i=1}^{80} (H^O(i) - H_k^M(i))^2$ $k=1$ ..6 in the 5 parameter domain using the simplex method of optimization Very briefly, the method tries to find a new set of parameters $\{X_c, A_0, A_c, L_0, N\}$ (equivalently a new magnitude response $H^M(i)$) which gives a minimum value for the cost function E from the initial starting points The procedure (algorithm) is repeated until the error E converges to within reasonable tolerances on the model parameters. The tolerances for each parameter is defined by the changes in values of the subsequent model parameters The following tolerance values are given $d(X_c) = 0\ 15$, $d(A_c) = 0.15$, $d(A_0) = 0\ 5$ and $d(L_0) = 0.15$, where d denotes the differences in values To illustrate the nature of convergence of the model parameters, typical values obtained for the parameters during

the optimisation procedure for vowel /i/ are given in Table 3.1.

It was observed that the algorithms converges after 8 - 18 iterations for all the vowels tested. The flow chart of the optimization process is shown in fig. 3.1.

Iteration	A_c	X_c	A_0	L_0
1	0.1	5.00	4.00	1.10
2	0.15	4.00	5.00	1.15
3	0.25	3.25	4.75	1.15
4	0.46	3.84	5.16	1.18
5	0.40	2.67	4.46	1.16
6	0.39	3.32	4.95	1.18
7	0.31	3.49	5.20	1.21
8	0.40	3.61	4.43	1.13
9	0.50	3.30	4.77	1.19
10	0.65	2.69	3.55	1.10
11	0.51	2.50	4.71	1.20
12	0.49	2.78	4.64	1.18
13	0.44	2.88	4.54	1.17
Target values	0.42	3.03	5.21	1.40

Table 3.1 A typical set of values obtained for the parameters during the downhill simplex method of optimisation for vowel /i/. The optimisation is stopped after 13 iterations due to insignificant changes in the parameters.

3.1.2 Conclusion from the random parameter optimisation

The results obtained by initialising the optimisation with random sets of values for the

parameters were unsatisfactory Most of the time, spectral and area function matches obtained are not even remotely similar to the original functions Therefore, starting the optimisation procedure with a random set of parameters was abandoned as a method of estimation of the parameters

However, when the optimisation was started with different sets of initial values, it invariably produced matches which were similar to the given function This suggested that the cost function must have several local minima, and the failure to start the optimisation procedure near the global minimum leads to incorrect estimates of the parameters. This implied that some method had to be found to restrict the search area to near the global minimum To be able to know the correct search area (area which contains the global minimum) we examine the nature of the cost function with each of the model parameters separately

3.2 DEPENDENCE OF THE SPM ON AREA FUNCTION PARAMETERS

One would expect that each parameter of the model has different degree of effect on the SPM The knowledge of these effects on the model should allow us to initiate the optimisation at the correct initial point Therefore a case study was done to investigate the effects of each parameter on the model by individually varying the values of each parameter The results of this case study are discussed in the following subsections

3.2.1 Cost function vs X_c

For all six vowels, the cost function $E(k) = \sum_{i=1}^{80} (H^{O}(i) - H_k^{M}(i))^2 \quad k = 1 \quad \frac{L}{2}$ where L is the length of the vocal tract in millimetres, was calculated by varying the parameter X_c in steps of 2 mm and assigning some neutral values to the other model parameters. The graphs of cost function against X_c for each vowel are shown in Figs 3 2a to 3 2f

All these graphs show a distinct global minimum value for the cost function at a particular value of X_c. When the optimisation is started with an X_c value near the global minimum, the algorithm produces good matches for both spectra and area functions.

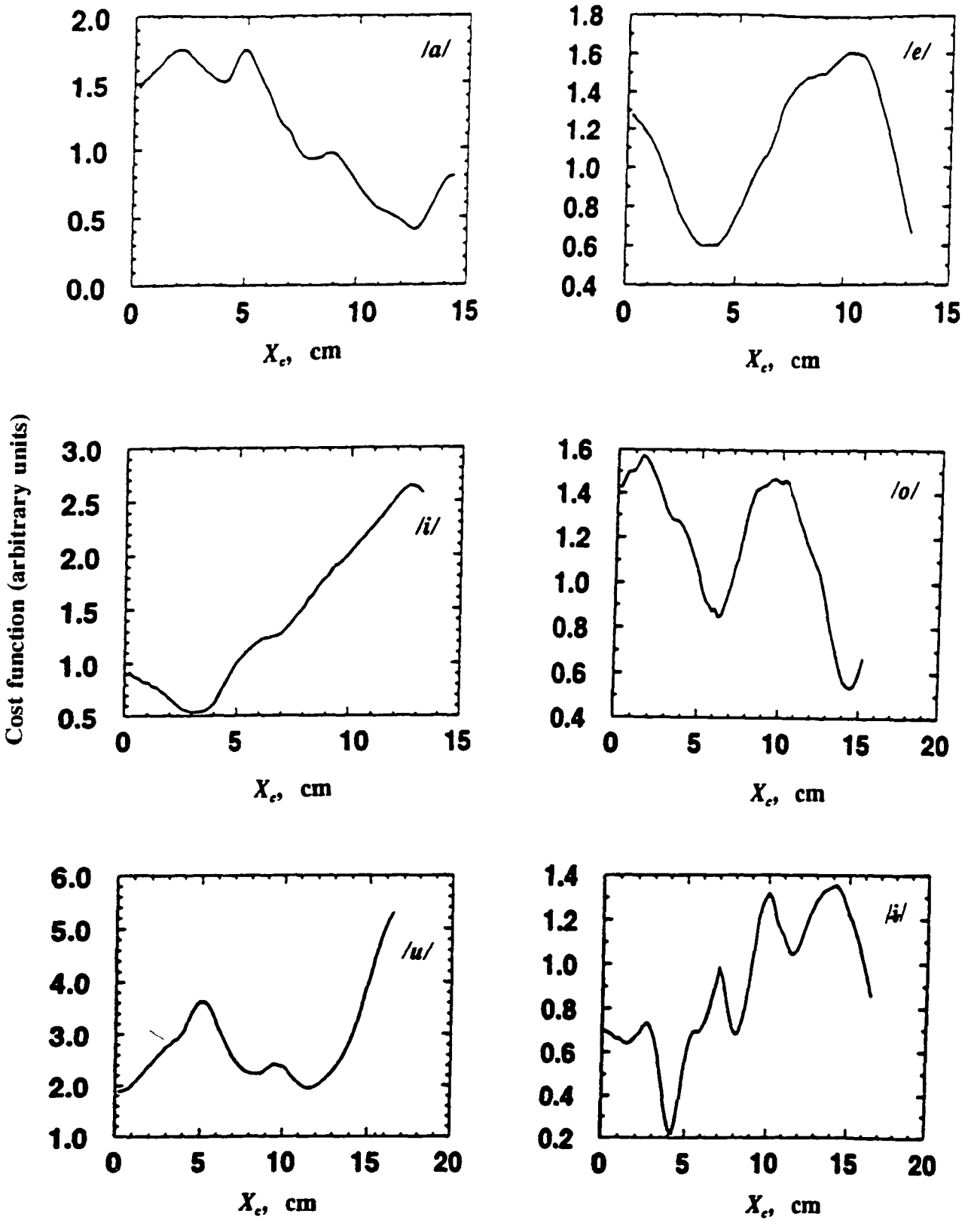


Fig. 3.2 Cost function Vs X_c for six vowels. X_c is increased in steps of 2 mm from the lip end to the glottal end. Other parameters A_c, A_0 and L_0 were given typical average values.

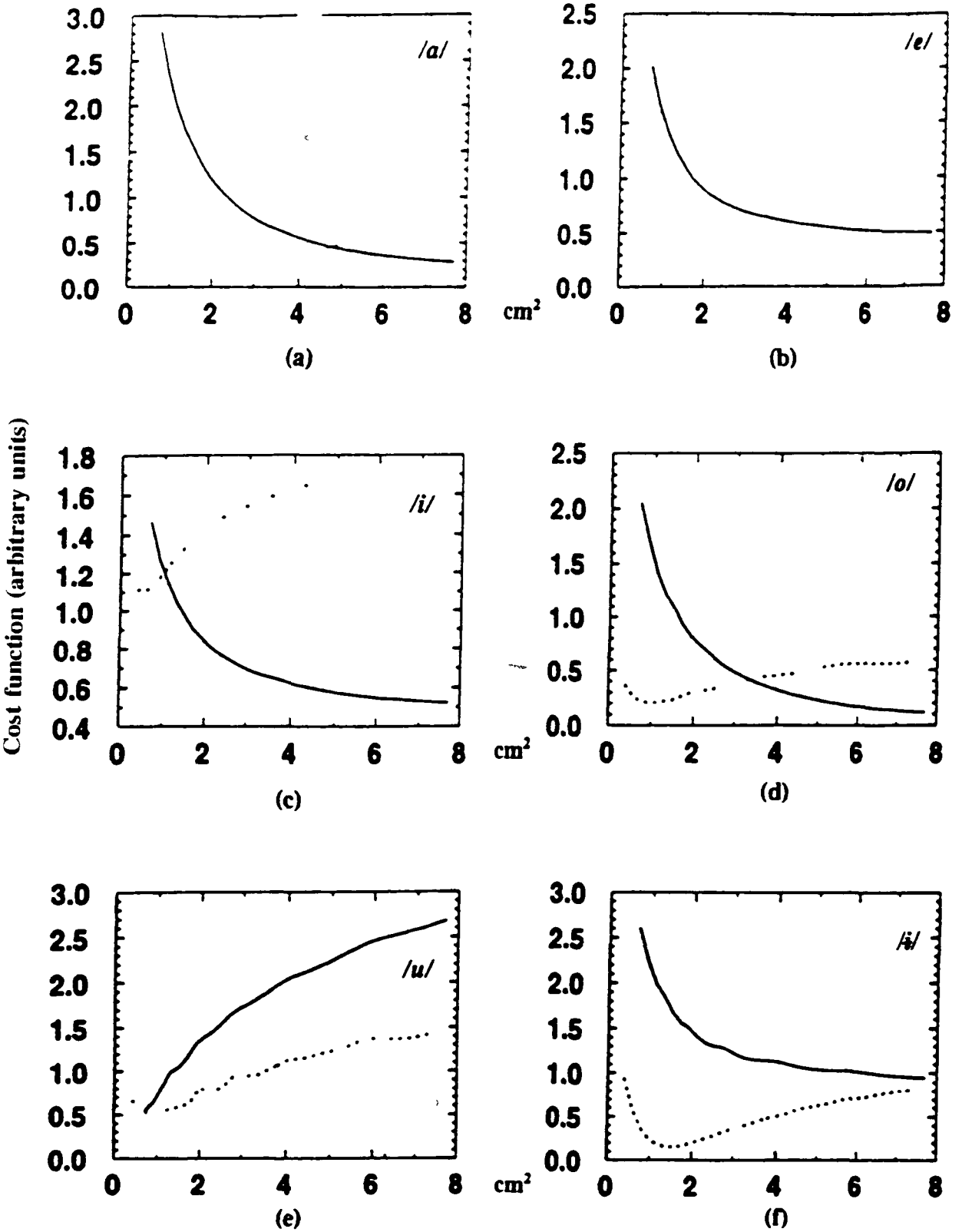


Fig. 3.3 Cost function Vs A_c and Cost function Vs A_0 for six vowels. Other parametes were given average values. a:) Vowel */a/* b:) Vowel */e/* c:) Vowel */i/* d:) Vowel */o/* e:) Vowel */u/* f:) Vowel */ɨ/*

.....: Cost function Vs A_c; _____: Cost function Vs A_0

3.2.2 Cost function vs A_0; Cost function vs A_c; Cost function vs L_0

Similarly, graphs of cost function vs A_0 were plotted by varying A_0 in steps of 0 2 cm^2 and by fixing some constant values to the other parameters The graphs (Fig 3 3a to Fig 3 3f) obtained are smoother than those obtained for the cost function vs X_c, with a few or one minima This suggests that the optimum value for A_0, which gives the global minimum for the cost function in multi dimensional space, can be estimated by giving the smallest and the largest possible values to A_0 in the initial set of parameters in the optimisation algorithm

The range over which A_c and L_0 vary, is small (about *0 to 4 cm^2* for A_c and *0 2 cm to 2.5 cm* for L_0) Also the cost function, when plotted against A_c or L_0 (fig 3 3a to fig 3.3f), behaves in a regular fashion i e there is only one minimum

3.2.3 Conclusion of the case study

From the study of the cost function vs each of the individual parameter and from the experience of estimating the parameters, it was concluded that the sensitivity of the articulatory parameters to the spectra can be given in decreasing order as X_c, A_0, A_c and L_0 When the optimisation of five parameters is started near the global minimum value of X_c, the algorithm produces good matches for both the spectra and the vocal tract shapes for all synthetic vowels. Although there is no guarantee that starting the optimisation with the global minimum of the individual parameters would always converge to the global minimum in the multi dimensional space, the method consistently produces good matches for both area function parameters and for spectra

3.3 GENERATION OF A SMALL LINKED CODEBOOK

How can one provide optimum initial values for the optimisation? The optimum initial values can be provided through a small size (180 shapes) linked codebook (see section 3 4) Since the parameter X_c has more effect on the spectral changes of the model than the other parameters, a codebook was generated by varying X_c in steps of 4 mm from the glottis to the lip end for five different number of vocal tract tubes (15,16,17,18 and 19). The other three parameters were assigned some neutral values All the vocal tract shapes (180 shapes in all) and their associated 80 points (0-4 kHz at 50 Hz steps) magnitude response were stored The magnitude response of each tract shape is found

using a program written in "C"(see appendix C2 for the flow chart).

3.3.1 Estimation of the parameters using a *linked codebook* and the optimisation procedure

In searching for the optimum tract shape for initialisation, the original speech spectrum is compared (using $E = \sum_{i=1}^{80} (H^O(i) - H^M(i))^2$) with the spectral part of the *linked codebook*. The shape which is closest to the original spectrum is then selected as one of the six initial set of points for the optimisation. The other five initial points are selected by giving random values to the parameters in their acceptable range of values. The acceptable range of values for the parameters are as follows: A_0 : 0 to 8 cm^2; A_c : 0 to 3 cm^2; L_0 : 0.1 to 2.5 cm and N: 15, 16, . . . 19.

In fact, we found that the value for the parameter N (number of vocal tract tubes) selected from the codebook always turned out to be the best value N, even after the optimisation procedure. Therefore the value of N found from the codebook was chosen to be the optimum value. This reduces the optimisation complexity to four from five dimensions. The estimation of the parameters are performed using an optimisation program written in "C".

3.3.2 Results and conclusions from the initial estimation attempts at parameters

Spectral matches obtained using the above method were shown in fig. 2.11. The corresponding area functions obtained are shown in fig. 3.4. Both of these figures clearly show that the matches obtained for all six vowels are good in both the area function and the frequency domain. Table 3.2 lists the first four formant frequencies of the estimated shapes and their original values.

In conclusion, the method discussed above produces good estimates for the model parameters. However, the algorithm efficiency can be improved in a number of ways:

a) A sufficiently large codebook adequately spanning both the articulatory and spectral spaces, would eliminate the need for any subsequent optimisation.

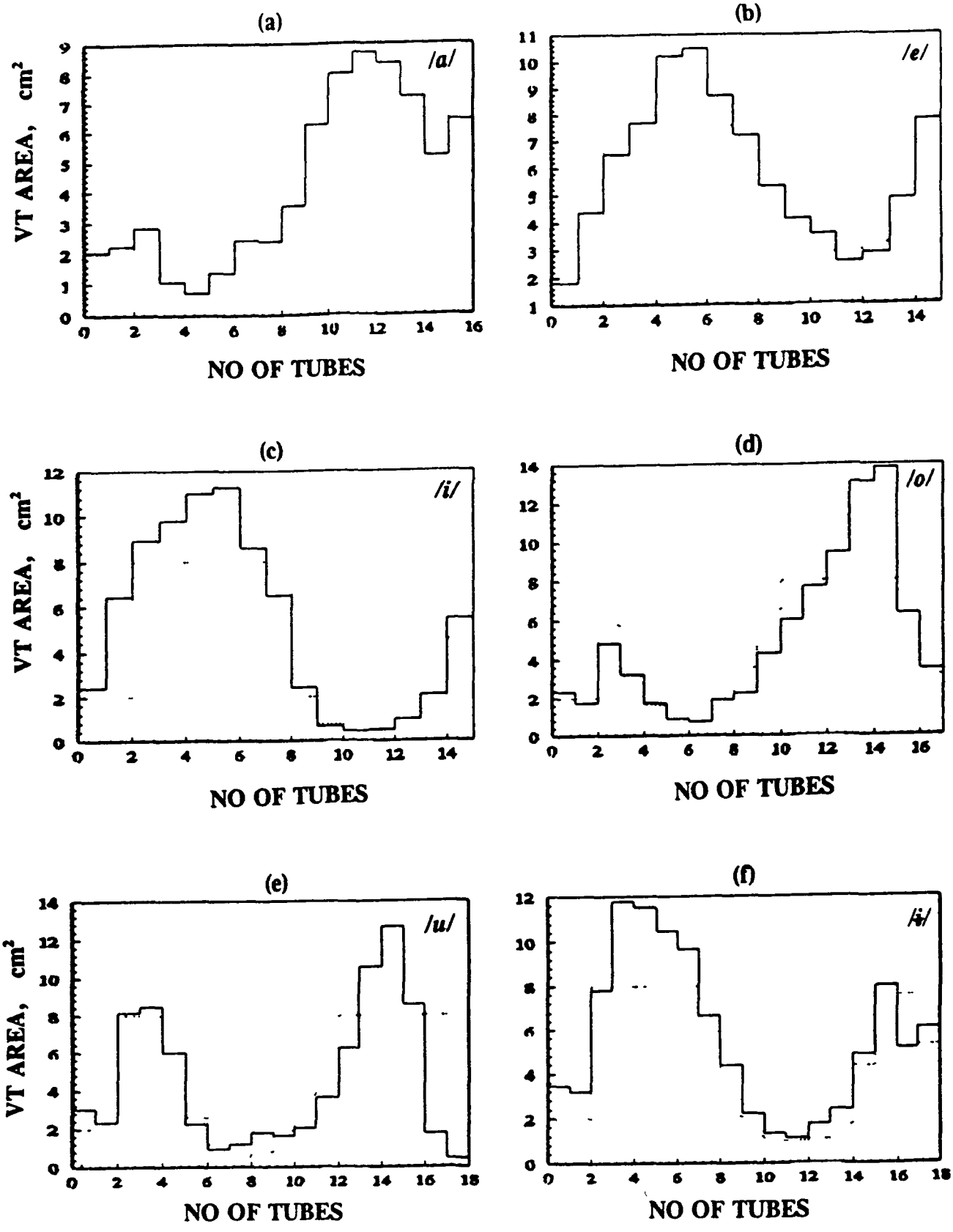


Fig. 3.4 Area function matches of six synthetic vowels using the optimisation.

a:) Vowel */a/* b:) Vowel */e/* c:) Vowel */i/* d:) Vowel */o/* e:) Vowel */u/*

f:) Vowel */ɨ/* .....:Estimated; _____:Original

b) A better criterion could be chosen to match the parameters For example the mean square cepstral distance measure which is equivalent to the Log Spectral Distance (LSD) can be used. The proof of equivalency of the criterion is given in appendix C3. The cepstral measure will reduce the number of spectral points to be compared (from 80) to a small number (between 20 to 25) at 8 kHz sampling rate. This will reduce codebook size and the analysis time i e. codebook access time. In addition, the cepstral measure separates the source characteristics from the vocal tract characteristics (see also section 3.6 1) This is an essential requirement for the accurate estimation of the SPM parameters [Schroeter, et al , 1987].

Extensive work has been carried out here, based on the above two ideas This method called the linked codebook technique is a major part of this thesis work and hence a detailed description of this work is given in the following sections

Vowel	F_1	F_2	F_3	F_4
/a/	692 (744)	1180 (1147)	2438 (2458)	3613 (3544)
/e/	466 (466)	2007 (2003)	2823 (2842)	3553 (3586)
/ɪ/	286 (285)	2248 (2348)	2719 (3045)	3934 (3975)
/o/	604 (620)	933 (922)	2426 (2395)	3626 (3512)
/u/	377 (377)	618 (660)	2375 (2381)	3467 (3440)
/ɨ/	348 (341)	1577 (1556)	2341 (2289)	3183 (3274)

Table 3.2 Estimated and original (shown in the brackets) first four formants of Russian vowels [Scaife and Kanagaratnam, 1990].

3.4 INTRODUCTION TO LINKED CODEBOOK TECHNIQUES

The linked codebook consists of two sections. One contains articulatory/vocal tract area function parameters and the other describes the acoustic properties of the corresponding tract shapes The linked codebook approach is used here to estimate the articulatory

parameters from the speech signals. This technique which is relatively new and similar to the codebooks used in existing speech coders, has been previously employed successfully [Schroeter et al , 1987]

The advantages of accurate estimation of physiological shapes in synthesis, coding and medical therapy were discussed in section 1 1. The benefit of using the linked codebook in the estimation of vocal tract shape using articulatory models is that it largely eliminates the possibility of unrealistic area function that arises in other techniques (e.g LPC techniques). This is discussed in section 3 4 2.

3.4.1 Codebooks in speech coders

The use of a "codebook" in a speech coder reduces the data bit rate in the transmission of speech [J Haagen et al., 1991, Schroeder and Atal, 1985; Wong et al , 1982; Muller, 1990]. Different types of codebooks containing acoustic parameters as their entries, are used in speech coders

A 2 4 kbps speech coder designed in [J Haagen et al , 1991] uses two different codebooks, one for the short term predictive parameters (LPC parameters) and the other for the long term predictive coefficients (excitation signals)

In CELP (Codebook Excited Linear Predictive coders) so called, innovation sequences (excitation signal) are generated by Gaussian distributed random numbers and stored as entries in the codebook. In synthesis of speech, the optimum innovation sequence (for the excitation) is selected by exhaustive search from the codebook, using a given fidelity criterion. This innovative sequence is then passed through two recursive filters, one for introducing voice periodicity and the other for spectral envelope, to generate the synthetic speech signals. The spectral envelope is defined by short term predictor coefficients filter and the voice periodicity is defined by long term predictor coefficients filter, estimated from the speech signal. The fidelity criterion is chosen by comparing the original and synthetic speech samples [Schroeder and Atal, 1985]

In other cases, codebooks are designed from, a large set of training speech data, the speech corpus, spanning a wide acoustic space [Wong et al , 1982] The required acoustic parameters, e.g. LPC coefficients, cepstral coefficients etc , are then estimated from this speech data and stored in a codebook. The codebook size is then reduced by

Vector Quantisation (VQ) techniques

Fig 3 5 shows a simplified block diagram of a low bit rate speech coder At the encoder (Fig. 3 5), incoming speech is compared with the speech synthesised from the codebook entries according to some criterion e.g likelihood ratio, cepstral difference, mean square error difference [Wong et al , 1982, Schroeter et al , 1987] The codebook entry which gives the minimum error (the one which is most similar to the incoming speech) is then chosen as the representation of the incoming speech In speech coders, the best matched parameter set is transmitted as an index together with the voiced/unvoiced information and pitch information At the decoder, the corresponding spectrum with the same index is retrieved from the codebook and used to resynthesise the original speech.

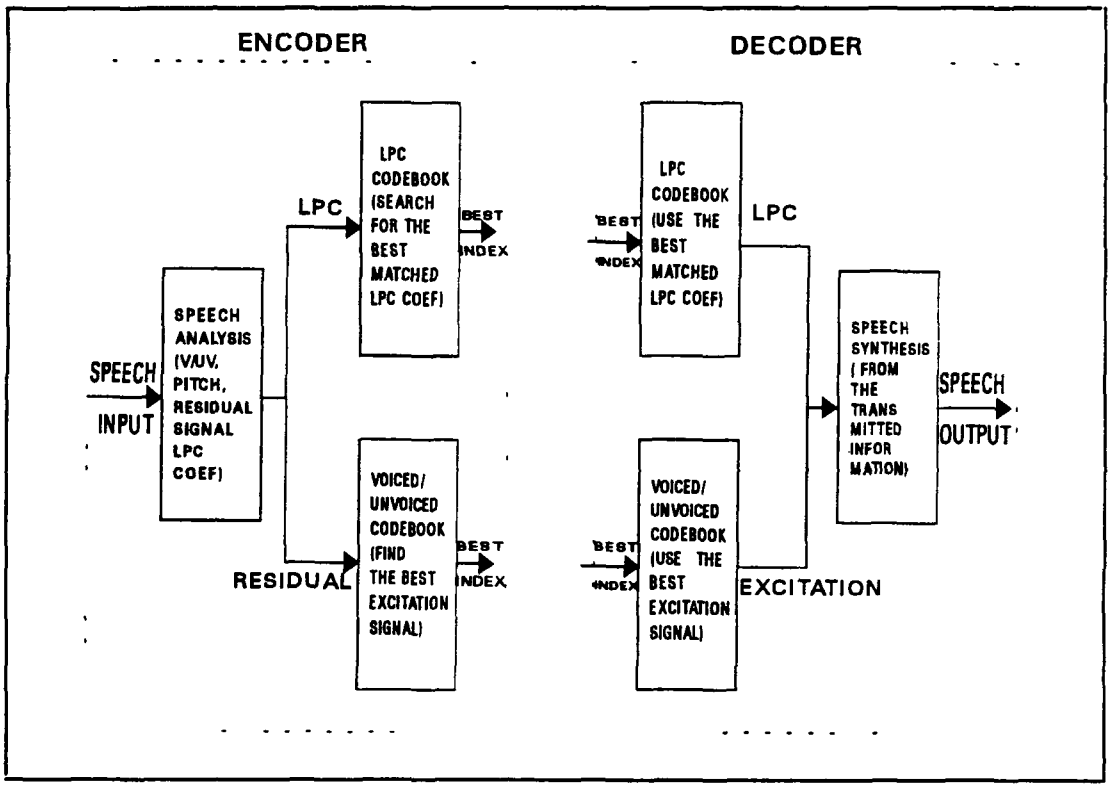


Fig.3.5 Schematic diagram of a generic speech coder

3.4.1.1 Linked codebook and codebooks in speech coders

The linked codebook has similarities with a codebook used in a conventional speech coder.

a) acoustic section of the linked codebook entries are similar to that of codebook of a conventional speech coder

b) matching criterion used for choosing the best codebook entry is usually similar

However, the generation of the linked codebook is very different from the generation of a conventional codebook. This is discussed in section 3.5.

3.4.2 Benefits of using a linked codebook technique over LPC techniques, for the estimation of area function parameters

The generation of the current codebook is discussed in section 3 5. The main advantages are that the linked codebook can be generated from a) models which include the major losses in the vocal apparatus b) realistic area function shapes. Therefore the use of a linked codebook in the estimation of the parameters eliminates the possibility of the unrealistic area function that arises in the LPC techniques Simply converting the codebook of LPC parameters of the speech coders to area functions [Wakita, 1973] can produce unrealistic physiological shapes for the following reasons:

a) LPC model does not properly include the important energy loss mechanisms (wall loss, radiation loss, etc) of the vocal apparatus and hence the bandwidth information are not modelled adequately However, for the determination of area function the accurate information about the bandwidth is essential [Sondhi, 1979]. Failure to include the loss mechanisms can therefore produce unrealistic area functions

b) It is well known that two entirely different set of LPC parameters can produce the same speech sound [Sondhi, 1979] Therefore the area functions, obtained by the direct conversion of LPC parameters of a particular sound can be different from the true tract shape. Also, it is possible that the tract shapes obtained by the conversion of the LPC coefficients may be unattainable by the human speech production system

c) The difficulty of normalising the area function from the reflection coefficients without a reference area

As mentioned above, these difficulties are to a certain extent overcome by the linked codebook approach. In addition, the linked codebook approach in speech analysis exploits both the benefits of codebook techniques (as in speech coders) and the

benefits of articulatory models

The generation of the linked codebook is computationally expensive Fortunately, the codebook is generated only once, so this does not affect the estimation time of the parameters

In estimating the true physiological parameters, as in the case of speech coders, incoming speech is compared with the acoustic part of the linked codebook and the best parameter set is chosen. But unlike the conventional coders, a set of articulatory parameters/area function parameters of the SPM corresponding to the best matched speech in the codebook is chosen to represent the incoming speech This chosen parameter set or in some cases a parameter set obtained by optimizing the chosen set is assumed to represent the true physiological shape which produced the speech

3.5 ANALYSIS METHOD

Here, a large set of speech signals are generated by varying the vocal tract shapes of the SPM The spectral parameters (e g cepstral coefficients) of the speech signals are stored as the acoustic section of the linked codebook and the vocal tract shapes (parameters) corresponding to these speech signals are stored as the articulatory section of the linked codebook

Two major decisions had to be made in the generation of a linked codebook Firstly, a method had to be found to generate a reasonable set of vocal tract shapes in the articulatory domain Secondly a decision had to be taken to choose the most suitable spectral parameter set eg LPC coefficients, Cepstral coefficients etc , to represent the corresponding acoustic data of the vocal tract shape

3.5.1 Codebook formation

An infinite number of vocal tract shapes can be generated from vocal tract models However, not all the shapes generated are physiologically attainable by the human speech production system Since the aim is to estimate true vocal tract shapes from speech signals, the unattainable vocal tract shapes which produce the similar speech signals should not be included in the codebook For this reason, the articulatory space is sampled non uniformly by sampling the so called most reasonable regions [Sondhi et al, 1987]

3.5.1.1 Nonuniform sampling of the articulatory space (see Fig. 3.6)

The initial tract shapes were chosen from the X-ray data of fifteen different voiced and fricative sounds including the six Russian vowels The cross sectional area of the vocal tract was measured every 1 mm from the glottis to the lips The X-ray data was obtained through a private correspondence with Dr. Pierre Badin [Badin, 1990] This area function data was then quantised using a tube length appropriate to a sampling frequency of 16 kHz (about 1.1 cm). The modified area function model parameters A_c, X_c, A_0, L_0 and N (number of vocal tract tubes) of these X - ray area function data were estimated by matching the spectra of the X - ray data with the model spectra. The optimisation technique (simplex method) used in section 3 3 1 was applied for the estimation of the parameters

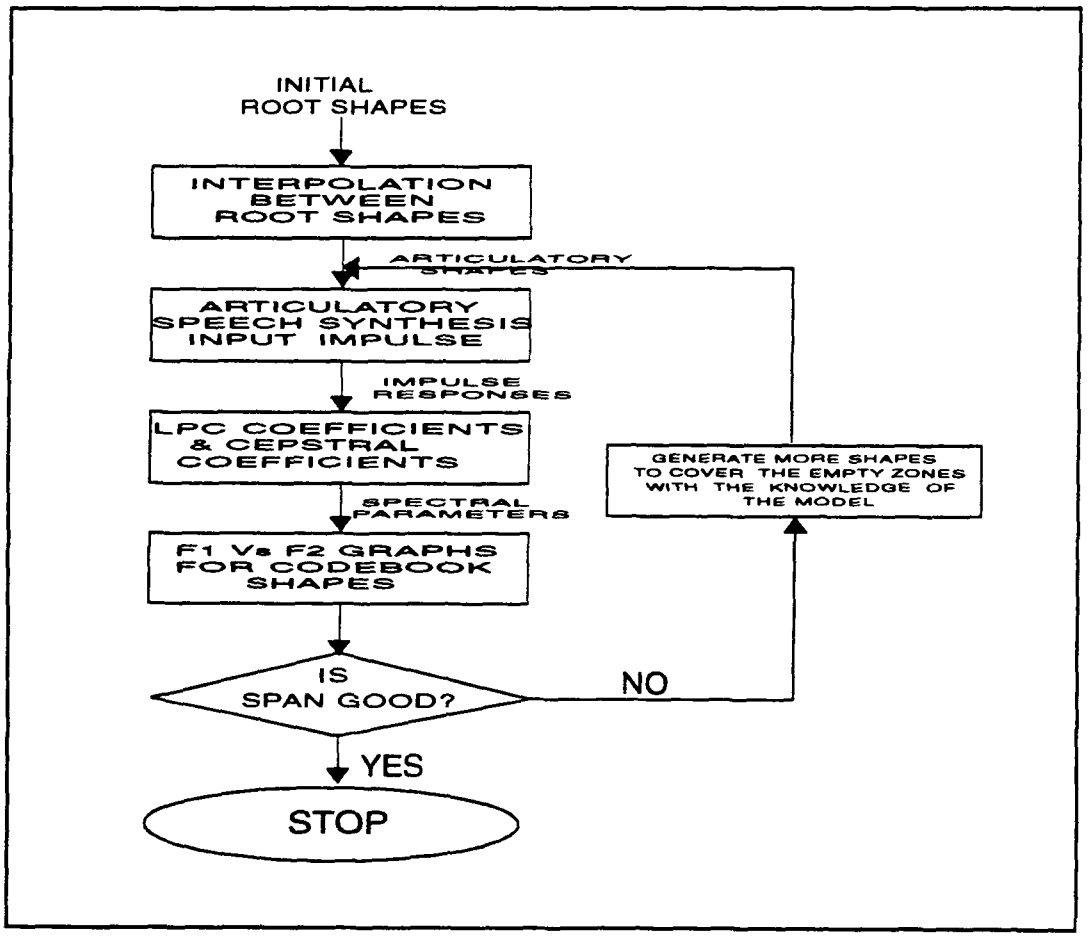


Fig. 3.6 Procedure for generating a linked codebook

IPA symbol (example)	A_c	X_c	A_0	L_0	no of VT tubes
a (*a*rm)	1 18	10.99	4 5	1.20	16
e (h*ea*d)	2.56	2 87	7.92	1.10	15
ɪ (pol*i*ce)	0.37	3.03	5.21	1.45	15
o (*o*ld)	0 51	11.32	4 53	1 20	17
u (tr*u*th)	0.47	10.43	0.44	1.13	18
I (b*i*t)	0.88	6.35	5.45	1.30	18
cx	0.10	5 65	6 48	1 30	16
f (*f*ish)	1.85	9.42	1.66	1.19	17
fp	0.43	2.74	1.24	1 30	16
ȷ (*y*es)	0 16	3 93	4.84	1.20	17
s (*s*un)	0.13	0.05	6.21	1.10	16
sc	0.62	5.37	4.71	1.40	16
sp	0.19	5.30	3.70	1.30	16
sx	4 14	10 30	3.45	1.20	17
x	0.28	7.40	4.96	1.00	18

Table 3.3 Root shapes and their corresponding Fant-Lin model parameters. IPA (International Phonetic Alphabet) symbols and examples of the corresponding sounds are given in the first column. In IPA symbols ƥ and fp "p" stands for palatalisation. In cx and sx "x" stands for circumflex.

The root shapes found using the aforementioned method for the 15 sounds, show good matches both in the articulatory domain (area function) and in the acoustic domain (spectra). The spectral and the area function matches obtained for the vowels */a/*, */e/*, */i/*, */o/*, */u/* and */i/* are shown in fig. 2.11 and fig. 3.4. The articulatory parameters A_c,

X_c, A_0, L_0 and N (the number of vocal tract tubes) estimated for the fifteen root shapes, six vowels and nine fricatives, are shown in Table 3 3

As a result of this analysis 15 vocal tract shapes, which are physiologically attainable by the human vocal apparatus, were chosen as the initial/root shapes All other shapes in the codebook (named cdbk1) were generated by interpolating between one root shape to the another in only one direction in *9* steps This method has the drawback mainly that the ((15*14/2)*9 + 15 = 960) 960 shapes produced by the interpolation are not guaranteed to fully span the acoustic space This problem is to a certain extent overcome by adding more tract shapes to the interpolated codebook (see section 3.5.3)

Once the vocal tract shapes were generated for the linked codebook the corresponding spectral parameters were obtained as follows.

a) For each vocal tract shape in the codebook, the impulse response (glottal flow to lip volume flow) was found using the SPM (see Fig 2 2a) at 16 kHz sampling rate Laine's lip radiation impedance Z_L (a pole/zero model given by eqn. (2 15)) was also included in the SPM, to model the lip radiation characteristics of real speech signals This would give approximately +6 dB rise at high frequencies. The +6 dB trend is removed by an integration (filter with $(1 - \alpha z^{-1})$ characteristics) process Therefore, the generated impulse response had spectra with zero trend (0 dB falloff) in the frequency domain.

b) The impulse responses were digitally filtered using a 3 - pole low pass Bessel filter designed at a 16 kHz sampling frequency As it is well known that the spectral characteristics of speech signals below 4 kHz must be preserved for the modelling of speech production system, a filter with a 2.5 kHz cutoff frequency and a gradual roll off to -8 dB at 4 kHz was designed The Bessel filter was chosen for its linear phase response This characteristic is essential for inverse filtering in which time domain signals (including both the phase and magnitude characteristics) are obtained [Karlsson, 1988] The characteristics of the low pass filter are given in appendix C4 The low passed speech is then down sampled to 8 kHz.

c) The number of LPC coefficients computed during the LPC analysis on real speech was varied from 10 to 12. It was found that 10 LPC coefficients consistently

underestimates the number of formants in the given spectra, while 12 LPC coefficients consistently overestimates the number of formants. However, the use of 11 LPC coefficients was found to give optimal spectra at an 8 kHz sampling frequency Therefore 11 LPC coefficients were used to generate the codebook The LPC spectra were normalised to 0 dB at DC level (see section 3.5.3), before being converted to cepstral coefficients.

d) Finally, the LPC coefficients were converted to cepstral coefficients and stored in the codebook The reasons for using the cepstral coefficients are given in section 3.3.2 and 3.6 1. These cepstral coefficients now represent the acoustic section of the linked codebook. The conversion of LPC coefficients to cepstral coefficients is discussed in appendix C3.

3.5.1.2 Normalisation of the spectral part of the linked codebook to 0 dB at 0 Hz

Different vocal tract shapes produce different output signals, and hence different energy contents, for the same input flow at different frequencies. Therefore, the spectra (magnitude responses) of the impulse responses produced using different vocal tract shapes in the codebook, will have different magnitude at DC level (zero frequency) However, we have normalised all the spectral shapes estimated using the LPC model, although it is not strictly necessary, to 0 dB at DC level. This is achieved as follows:

Let the transfer function of the LPC filter $H(z)$ be.

$$H(z) = \frac{1}{1+\sum_{k=1}^{P} a_k . z^{-k}} \qquad 3.1$$

Substituting $\omega=0$ *i.e* $z = e^{j\,0t}$ in the transfer function $H(z)$ gives the magnitude at DC $|H(z)|_{\omega=0}$ (say G_0) as:

$$|H(z)|_{\omega=0} = G_0 = 1 / (1 + a_1 + a_2 + \ldots + a_P)$$

where a_1, a_2, . . . a_p are set of LPC coefficients and they are different for different signals This gives different values of gain i e. $20*log(G_0)$ dB at zero frequency

However, multiplying eqn. (3.1) by $1/G_0$ normalises the transfer function $H(z)$ to 0 dB gain (unity gain) at DC level (0 Hz) as

$$H'(z) = \frac{G'_0}{1+\sum_{k=1}^{P} a_k z^{-k}} \qquad 3\ 2$$

where $G'_0 = 1/G_0 = (1 + a_1 + a_2 + \ . . + a_P)$

As the zeroth cepstral coefficient is given by (see appendix C3) $c(0) = log(G)$, the zeroth cepstral coefficient of the normalised transfer function in eqn (3 2) is given by

$c(0) = log(G'_0)$

3.5.2 A Test to determine the adequacy of the linked codebook

Spectrograms convey vital information about the speech signals i e. they show the energy contents of the signal at various frequencies. The formant frequencies in the spectrograms are critical elements of voiced speech signals and are very closely related to the human vocal tract shapes. Therefore, the formant frequency space covered by the codebook of vocal tract shapes, can be used as a tool to justify the adequacy of the codebook

Fig. 3.7 shows the F_1,F_2 span of the codebook (cdbk1) of vocal tract shapes The first two formants of the fifteen root shapes overlaid on fig 3.7 indicate that the majority of the generated shapes have their first two formant frequencies (F_1,F_2) around the overlaid points. This implies that the generated vocal tract shapes have spectra similar to the root shapes' spectra

However it is noticeable from Fig 3.7 that the sounds */o/*, */u/*, */s/*, */cx/*, */j/*, and */sp/* lay near the boundary of the F_1 Vs F_2 space covered by the codebook Therefore, these sounds will not be covered adequately by the current codebook cdbk1. This problem is, to a certain extent, overcome by adding more shapes (which are also humanly attainable) around these empty zones.

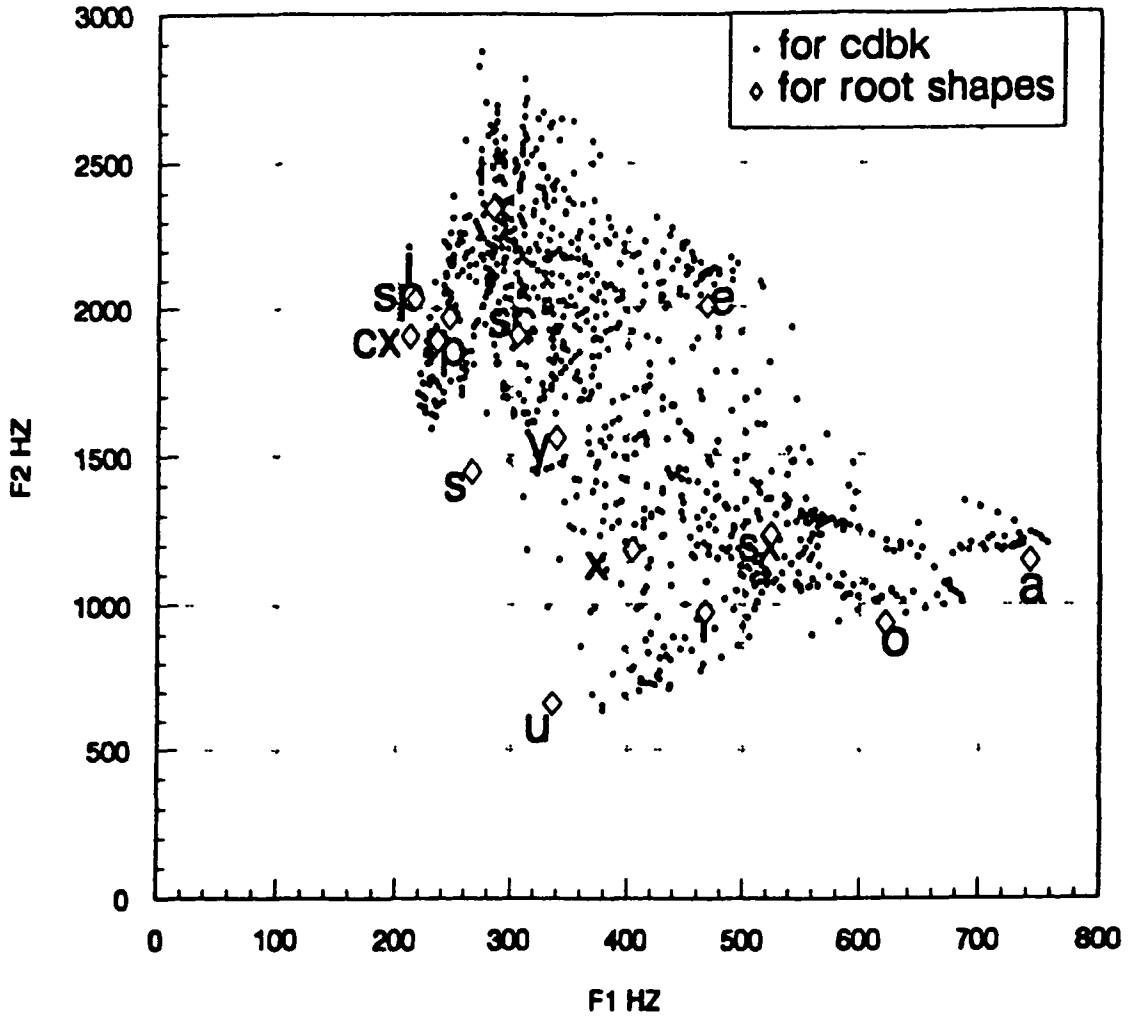


Fig. 3.7 F_2 Vs F_1 span of 960 vocal tract shapes from cdbk1. The codebook was generated by interpolating from each root shape to the next in 9 steps.

3.5.3 Extending the codebook

To extend the codebook in the empty zone one has to know the relationship between the articulatory parameters and the acoustic formant frequencies.

3.5.3.1 Analysis of (F_1,F_2) with respect to (A_0,X_c)

The extended codebook (cdbk2) is divided into three regions with respect to the second formant frequency as follows

a) $F_2 < 1000$ Hz

b) 1000 Hz $< F_2 <$ 2000 Hz

c) $F_2 > 2000$ Hz.

For each region, we found the articulatory shapes which produced the above F_1 and F_2 patterns The graph of one articulatory parameter against another in the articulatory domain indicates that the parameters A_0 and X_c play a major role in generating spectral shapes in the acoustic domain Furthermore, the set of graphs shown in Figs. 3 8, 3.9 and 3 10 show that each formant frequency region (F_1,F_2) has a clear relationship with the parameters A_0 and X_c.

A_c, the constriction area, of course has a major role in the vocal tract model. Increasing the value of A_c in general increases the second formant, and decreases the first formant with a reduction in both bandwidths However by giving reasonable values to A_c (0 1 cm^2 < A_c < 2 cm^2) in the root shapes a good quality codebook can be generated. Whereas it was observed that there were large changes in the formants F_1 and F_2 with the varying X_c and A_0 values.

The fig. 3.3 (dotted curve) showed that the change of error function value (spectral difference measure) with A_c is small and has a smooth relationship with A_c. Whereas, fig. 3 2 and 3.3 (solid curve) showed that the change of error was large with A_0 and X_c Fig 3.2 further showed that the error function has an irregular relationship with X_c. This further supports our view that the exclusive study of the first two formant frequencies (F_1,F_2) with (X_c,A_0) is more relevant than with that of A_c parameter, in order to produce a good quality codebook.

3.5.3.2 Analysis for F_2 < 1000 Hz

Analysing the (F_1,F_2) patterns of the extended codebook (cdbk2) for F_2 < 1000 Hz region showed, that the entire set of these patterns (Fig 3 8a) was mapped into a certain region of the articulatory parameter set (A_0,X_c) with A_0 and X_c taking the values 0 cm^2 < A_0 < 1.5 cm^2 and 5 cm < X_c < 13 cm (Fig 3 8b). Carefully sorting the (A_0,X_c) data of Fig. 3 8b further showed that very small lip opening areas (A_0 < 0.5 cm^2) and high X_c's produced very low F_1 & F_2 formant frequencies High X_c & moderate A_0 (0 5 cm^2 < A_0 < 1 5 cm^2) produced high F_1 & low F_2, and low X_c (X_c around 8 cm) & low A_0 produced high F_2 and low F_1 From this analysis we drew a useful conclusion that the vocal tract shapes with very low A_0 and high X_c produce low

F_1 and low F_2 values This fact was used for extending the codebook, by adding shapes with low F_1 and low F_2 to the original codebook (cdbk1).

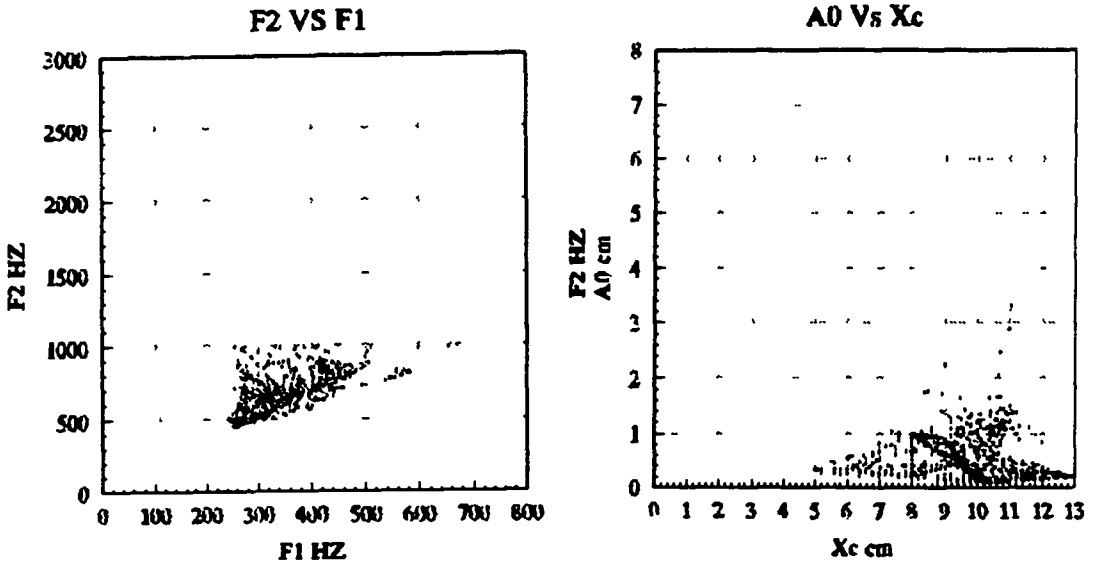


Fig. 3.8 a): Set of F_2 Vs F_1 points of cdbk2 for F_2 < 1000 Hz
b):Corresponding A_0 Vs X_c

3.5.3.3 Analysis for F_2 > 2000 Hz

Analysing the results for F_2 > 2000 Hz of Fig. 3.9 again showed a (A_0,X_c) relationship with the formants. Here it was seen that low X_c (2 cm < X_c < 6 cm) produced high F_2. Further analysis of the sorted data showed that low F_1 in Fig. 3.9a was due to low A_0 and high F_1 was due to high A_0.

3.5.3.4 Analysis for 1000 Hz < F_2 < 2000

Analysis of the data in the range 1000 Hz < F_2 < 2000 Hz gave results that were consistent with the other two regions (Fig 3 10) It showed that a strong relationship exists between the parameters (A_0,X_c) and the formant frequencies The conclusion

drawn from Fig 3 10 is that medium values of X_c & medium values of A_0 produce values of F_2 between 1000 and 2000 Hz

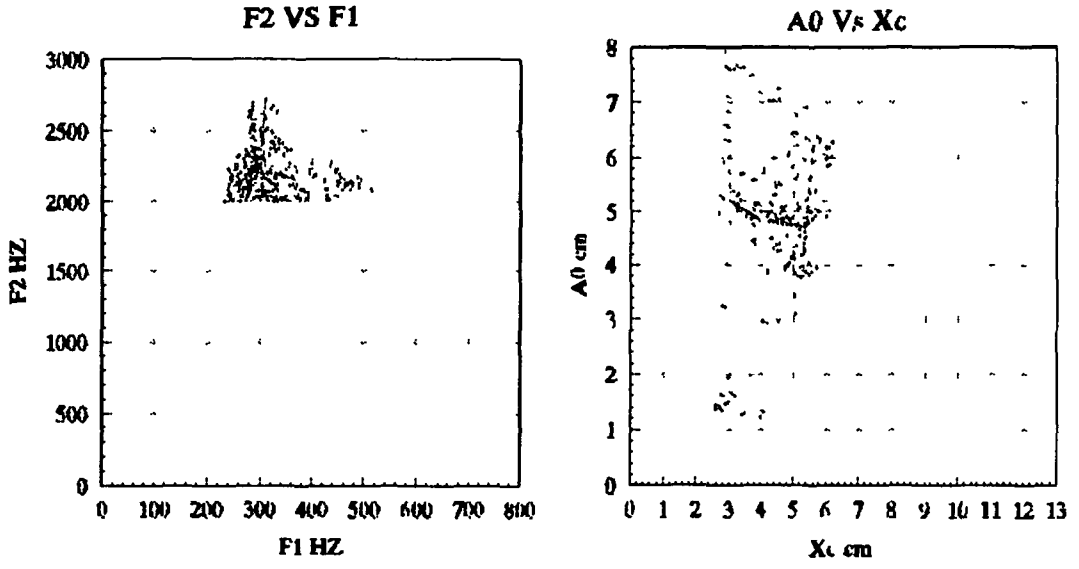


Fig. 3.9 a): Set of F_2 Vs F_1 points of cdbk2 for F_2 > 2000 Hz

b):Corresponding A_0 Vs X_c

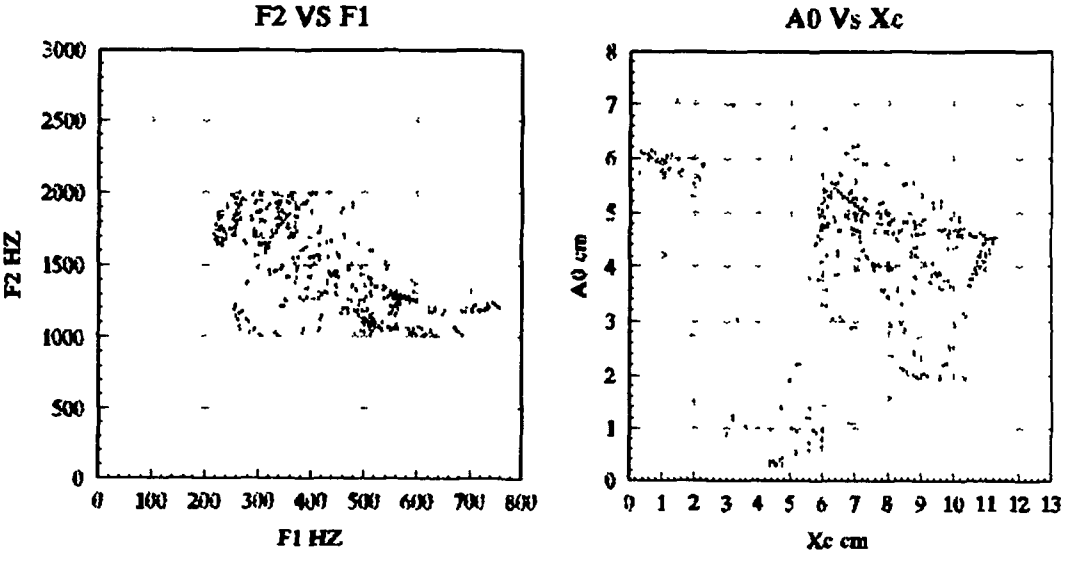


Fig. 3.10 a): Set of F_2 Vs F_1 points of cdbk2 for 1000 Hz < F_2 < 2000 Hz

b):Corresponding A_0 Vs X_c

	A_0	X_c	F_2	comments on F_1
from fig 3.8	$A_0 < 2$	> 8	< 1000	very low A_0 gives low F_1
from fig 3 10	$2 < A_0 < 6$	$6 < X_c < 11$	$1000 < F_2 < 2000$	low A_0 gives low F_1
from fig. 3 9	$1.2 < A_0 < 8$	< 6	>2000	as above

Table 3.4 Quantitative description of the relationship between (A_0,X_c) and (F_1,F_2)

This analysis of sorting (F_1,F_2) into three regions and taking the corresponding articulatory parameters, clearly show the relationship between the area function parameters and the spectral parameters By ignoring the effects of parameters A_c and L_0 on F_1 and F_2 formant frequencies enables (F_1,F_2) to be expressed as a function of (A_0, X_c) as

$$(F_1,F_2) = g(A_0,X_c) \qquad 3\ 3$$

where the nature of the function is quantitatively described in Table 3 4

3.5.4 Extension of the original codebook cdbk1

Using the above known relationship (eqn. (3.3), more vocal tract shapes were added to the codebook "cdbk1" to cover the sparsely populated zones. For this reason, two more codebooks with 415 shapes ("415exd") and 260 shapes ("260exd") were generated These codebooks were generated in similar manner to the original codebook "cdbk1" (section 3.5.1) but using ten and eight root shapes respectively Suitable sets of values for the parameters of the root shapes, were chosen from the experience of dealing with the above model and the relationship which we discussed in the previous section Table 3 5 shows the values of ten initial shapes These root shapes which have very low A_0 values and low to high X_c values, when non uniformly sampled

(codebook named "415exd"), produce low F_1 and medium to high F_2 values. Therefore this codebook when added to the original codebook "cdbk1" covers the empty zones in the low F_1 and the medium to high F_2 regions. Table 3.6 shows the eight root shapes used for producing the codebook "260exd". In this case low A_0 and high X_c values are chosen to produce low F_1 and low F_2 values to cover the sparsely populated zone of "cdbk1".

The frequency span of the large codebook "cdbk2" generated by adding codebooks "cdbk1", "415exd" and "260exd", containing total of 1635 vocal tract shapes is shown in Fig. 3.11. It is evident from fig. 3.7 and fig. 3.11 that the codebook "cdbk2" spans a large acoustic space. Therefore, this linked codebook (cdbk2) is chosen to estimate the model parameters from speech signals.

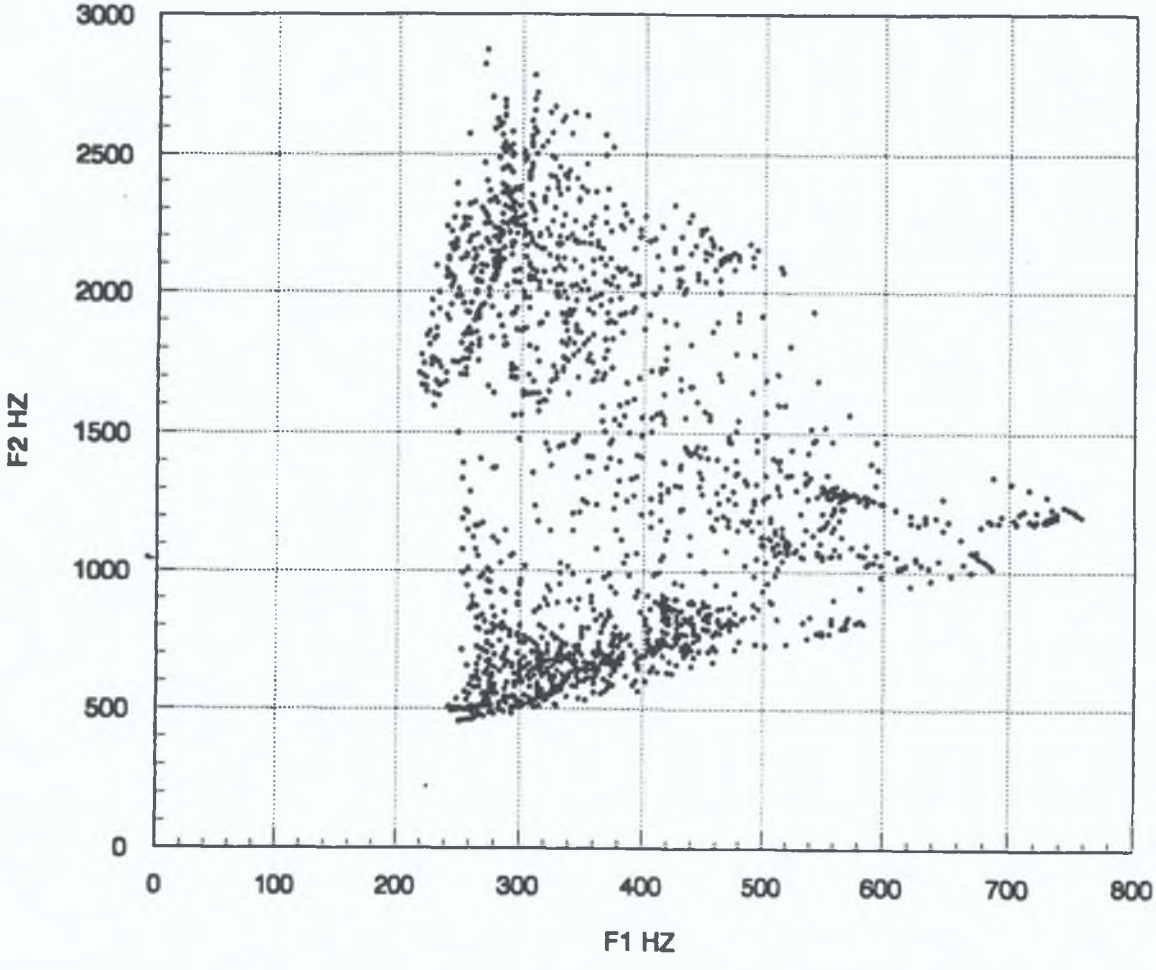


Fig. 3.11 F_2 Vs F_1 span of cdbk2 containing 1635 vocal tract shapes. This extended codebook is an addition of codebooks cdbk1, 415exd and 260exd.

	A_c	X_c	A_0	L_0	*no of VT tubes*
shape 1	0.20	10	0 1	1.30	18
shape 2	0.10	9	0 9	1.30	19
shape 3	0 50	4	0 3	1 10	18
shape 4	0 45	8	0.2	1.15	19
shape 5	0 30	2	1.5	1.30	16
shape 6	0 80	8	1	0 90	17
shape 7	0 20	12	0 15	1 10	16
shape 8	0 70	6	0.34	1 20	19
shape 9	0 60	9.45	0.46	1 12	18
shape 10	0.90	11	1 2	0 80	18

Table 3.5 Root shapes for the codebook "415exd"

	A_c	X_c	A_0	L_0	*no of VT tubes*
shape 1	0.20	10.43	0 1	1 30	18
shape 2	0.10	9	0 9	1 30	19
shape 3	0 50	12	0 3	1 10	18
shape 4	0 45	13	0 2	1 15	19
shape 5	0 30	11	1 5	1 30	17
shape 6	0.80	8	1	0 90	17
shape 7	0 20	10	0 15	1 10	19
shape 8	0 30	9 8	0 34	1 20	19

Table 3.6 Root shapes for the codebook "260exd"

3.6 THE EVALUATION OF MODEL PARAMETERS

In articulatory modelling of speech, the model speech Y' can be notionally expressed by a function $f(X)$ of the articulatory parameter set X as

$$Y' = f(X) \qquad 3\,4$$

For the current model vector X is defined as: $X = \{A_c, X_c, A_0, L_0, N\}$ The parameters A_c, X_c, A_0, L_0, N are defined as previously in chapter 2.3.1.

The function $f(X)$ in equation (3 4) is a non linear function and thus estimating the vector X through the inverse function of Y' (i.e $f^{-1}(Y')$) is not a trivial task Therefore the problem of estimating the parameter set X, always develops into minimising the error (distortion measure) between some form of real speech signal Y and the model speech signal Y' with respect to X.

The distortion measure must be (a) subjectively meaningful in the sense that small and large distortions correspond to good and bad subjective quality (b) tractable in the sense it can be mathematically modified to emphasize the important regions in the signals [Gray et al, 1980] Several spectral distortion or error measures such as mean squared spectral differences, likelihood ratios, mean squared cepstral differences etc., are available for optimisation of X (see appendix C5). Most commonly used distortion measure or error function is the mean squared error between the signals and can be mathematically expressed as

$$E(X) = (y - y')^2 \qquad 3\,5$$

where y, y' are some sort of representation of Y and Y' e g in the current work, y correspond to cepstral coefficients of the speech signal Y

Eqn (3 5) shows the basic form of the mean square error criterion used in signal processing techniques to estimate the parameters It is well known that speech signals below 4 kHz frequency band contain very important features about the production system. Therefore, the cost function $E(X)$ is invariably modified to emphasize this region by a weighting function W as given below:

$$E_m(X) = W\,E(X) \qquad 3\,6$$

where $E_m(X)$ is the modified cost function and W is a weighting function usually chosen to amplify the low frequency (below 4 kHz) features and to attenuate the high frequency features

3.6.1 Cepstral distances

For a successful estimation of the model parameters it is imperative to use a good distortion measure between the model speech and the given speech The cepstral Euclidean distance measure is chosen as the criterion to compare the model speech with the real speech in the current work, for the following reason.

The separation of the source and the tract characteristics from speech signal is not easy due to the source and tract interaction during the speech production However in searching for the best tract shape from the codebook it is important to use a measure which is, in general, insensitive to the source parameters or the excitation signals [Schroeter, et al., 1987]. As can be recalled, the linked codebook in the present work is generated using impulse responses of the model These impulse responses include no information about the excitation signal and therefore the distortion measure used with this codebook to compare the real speech should easily discriminate the source characteristics Since the cepstral criterion, at least partially, has these attributes it was chosen as the criterion for the estimation of parameters.

A recent paper [Meyer, Schroeter and Sondhi, 1991] discusses various lifter windows which can be used for weighting the cepstral distortion The authors attempt to obtain optimal weighting functions to reduce glottal effects on the distortion measure

3.6.2 The use of the linked codebook in the estimation of parameters from real speech.

As mentioned (section 3 5), the linked codebook in the current work was generated with an impulse input and have flat spectra However real speech signals, in general, have an -6 dB fall off trend at high frequencies due to the approximately -12 dB roll off of glottal flow and the approximately +6 dB rise due to lip radiation characteristics Often, in speech analysis, this -6 dB roll off in the speech spectrum is removed by a differentiation ($1 - \alpha z^{-1}$) process Performing this procedure called *pre-emphasis* (appendix C6) on speech signals, to a good approximation, eliminates the glottal flow

characteristics and the lip radiation effects from the speech signals

To justify that the analysis method mentioned in the above paragraphs can be used with the linked codebook, which does not include any glottal characteristics, the following simple test is performed on a synthetic speech signal Russian X-ray area function data of vowel */a/* and */e/* are used in the test

a) Produce an impulse response for a given vocal tract cross sectional area, using the SPM (fig. 2 2a) Radiation effects at the lips are also included in the SPM (fig 2 2a) by Laine's lip radiation model Z_L. This gives +6 dB/oct rise in the high frequency region This effect is removed by an integration process The LPC spectrum of this signal is then found using the autocorrelation LPC analysis

b) For the same area function synthesise speech pressure signal using the *Rosenberg* glottal flow [Rosenberg,1971] This speech signal has -6 dB/oct roll off trend Therefore, the LPC spectrum is found, after the removal of the -6 dB/oct through the *pre-emphasis* process

Fig 3 12 shows the spectra produced using both procedures (a) and (b) It is evident from the figures that the spectra produced using both methods have same formants and very similar spectral shapes

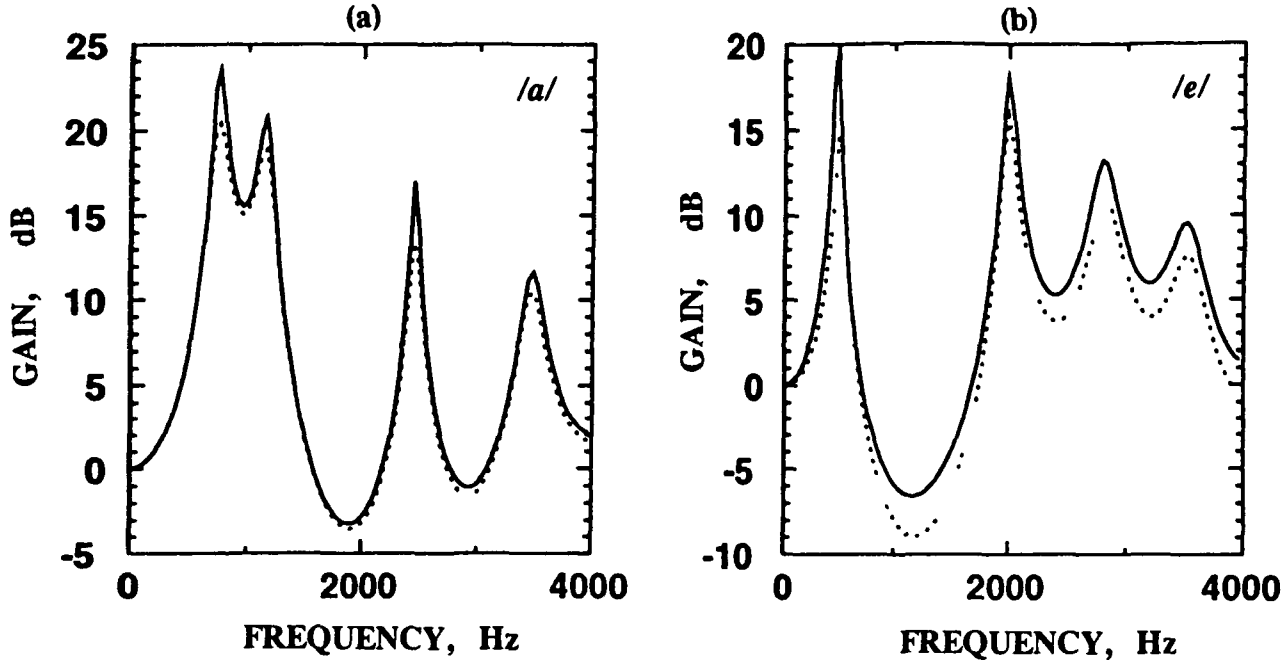


Fig. 3.12 a:) Spectra of vowel */a/* b:) Spectra of vowel */e/.

....: Spectra of the impulse response of the SPM

_____ Spectra of the synthetic speech produced from the Rosenberg glottal flow

Therefore, the *linked codebook* generated using the impulse responses, in the present work can be employed to estimate the SPM parameters from real speech signals

In this research two different types of cepstral measures are used for estimating (retrieving) the shapes from the codebook

a) The Euclidean cepstral distortion measure

b) The weighted cepstral measure

3.6.3 Estimation of parameters using the Euclidean cepstral distortion measure

The following cepstral distortion measure $d(c,\hat{c})$ which is equivalent to the Log Spectral Distance (L.S.D) between two frames is used for estimating the articulatory parameters from the codebook.

$$d(c,\hat{c}_i) = (c_o - \hat{c}_{io})^2 + 2\sum_{k=1}^{kmax} (c_k - \hat{c}_{ik})^2 \qquad i = 1,2,...N \qquad 3.7$$

where c = $(c_0, c_1, \ldots c_{kmax})$ is a vector of cepstral coefficients for input speech.

$\hat{c}_i$ = $(\hat{c}_{i0}, \hat{c}_{i1}, \ldots \hat{c}_{ikmax})$ is a vector cepstral coefficients derived from the codebook

i = index of the codebook entries

N = number of vocal tract shapes in the codebook

$kmax$ = number of cepstral coefficients

3.6.3.1 Speech material used for testing the linked codebook method

Speech material consisting of 11 strings, 6 synthetic vowels and 5 natural vowels, were chosen to test the estimation method discussed.

Vowels */a/, /e/, /i/, /o/, /u/, /ɪ/* and all other root shapes used in the generation of the codebook, were synthesised at 8 kHz sampling rate using the synthesis SPM designed in chapter 2 3 (Fig 2 2a) A Rosenberg glottal flow waveform [Rosenberg, 1971] synthesised at 100 Hz fundamental frequency was used as the excitation source X - ray vocal tract cross sectional area data of all 15 root shapes were chosen to represent the vocal tract section of the SPM.

Natural vowels were extracted from five meaningless words "bala", "bele",

"bili", "bolo" and "bulu" recorded at 48 kHz in an anechoic chamber These speech data were then down sampled to 8 kHz using a 2.5 kHz IIR digital low pass Bessel filter (see appendix C4 for low pass filter characteristics).

3.6.3.2 Test with synthetic vowels

Twenty five cepstral coefficients $(c_1, c_2, \ldots c_{25})$ were obtained for the synthetic vowels The derivation of cepstral coefficients, using LPC coefficients, from speech signal was described in section 3.5.2 These coefficients were compared with the cepstral coefficients of the linked codebook using eqn. (3 7) The codebook shape (i e the value of i) which gave the minimum $d(c,\hat{c}_i)$ of eqn (3 7) was chosen as the predicted area function shape for the given speech

Both the spectral matches and area function matches obtained using the above Euclidean distance measure from the *linked codebook* for vowels */a/ and /e/* are shown graphically in Figs 3.13 and 3 14 It can be observed from the graphs that the estimated spectra closely match the formants and overall spectral shape of the given synthetic speech spectra The matches obtained for the vocal tract cross sectional area were also very good

In fact, very good matches were obtained for all the root shapes except for the sounds [s], [sx] and [x]. However, this is not surprising as these sounds have more complex vocal tract cross sectional area and thus, cannot be modelled by the simple area function model described above Fig 3 15 illustrates the complex nature (two constriction points) of the vocal tract shape of [sx] and [x] Area function matches obtained for the sounds [sx] and [x] are also shown in Fig 3 15

The good matches obtained for the shapes in both the acoustic and the articulatory domain imply that

a) The simple Speech Production Model (SPM) described for the voiced speech correctly estimates the vocal tract shapes

b) The cepstral distortion measure chosen for the estimation purposes correctly estimates the parameters of the SPM

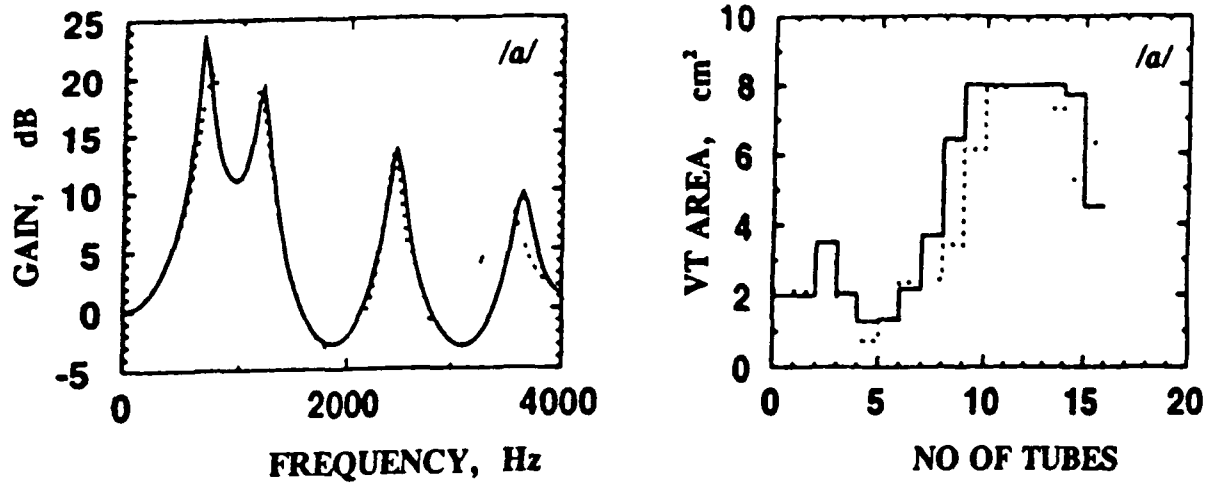


Fig. 3.13 Spectral match and area function match of the synthetic vowel */a/*

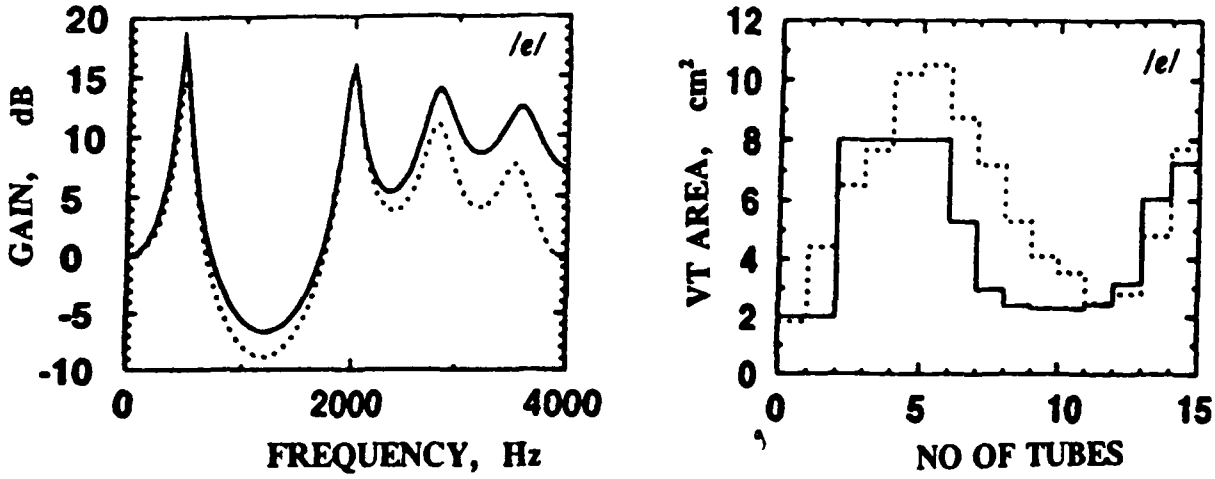


Fig. 3.14 Spectral match and area function match of the synthetic vowel */e/*

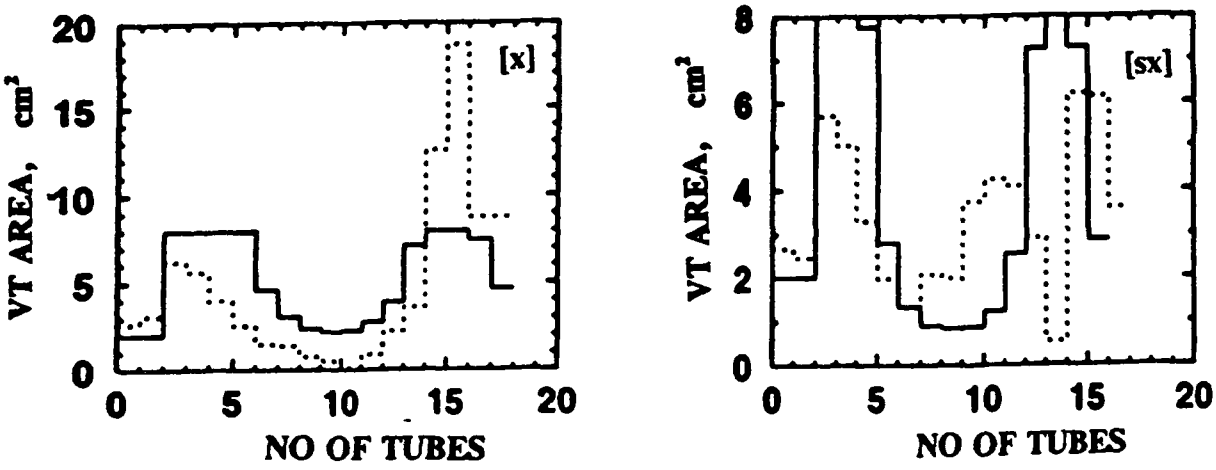


Fig. 3.15 Area function match obtained using the area function model for [x] and [sx].

.....: Original; ——: Estimated

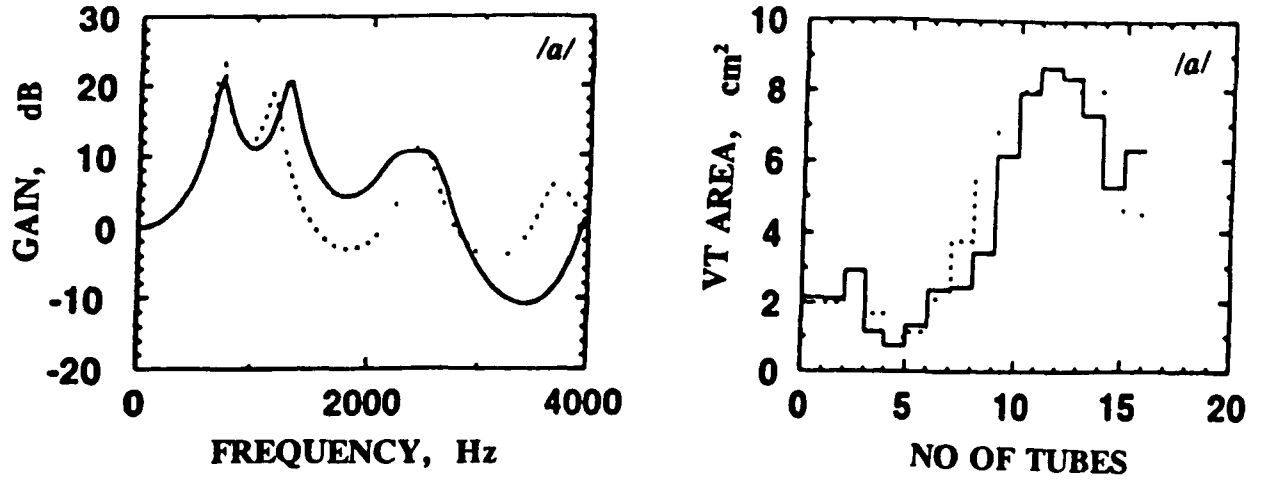


Fig. 3.16 Spectral and area function matches of the natural vowel */a/*

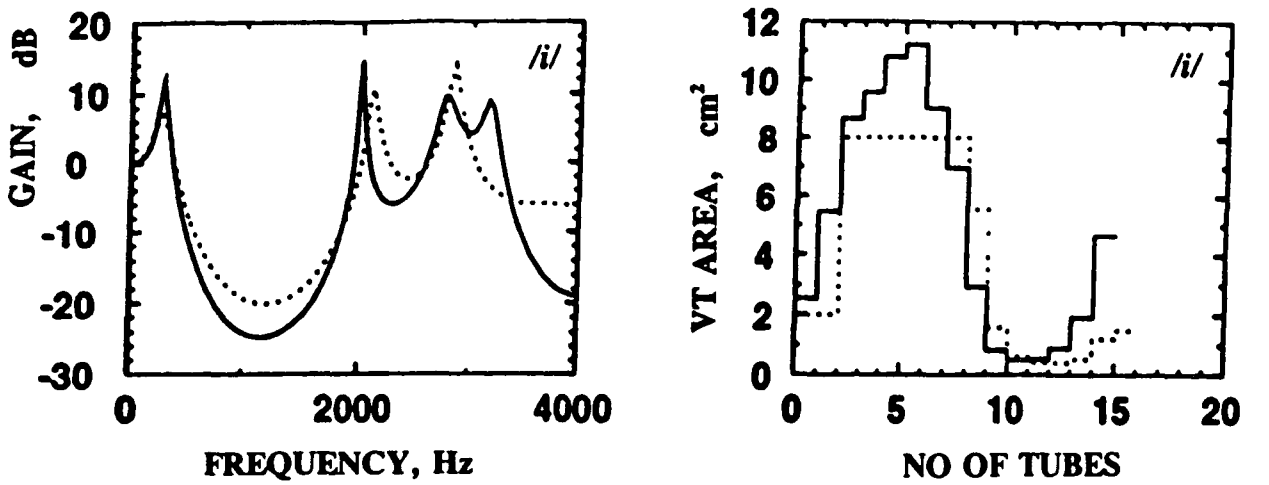


Fig. 3.17 Spectral and area function matches of the natural vowel */i/*

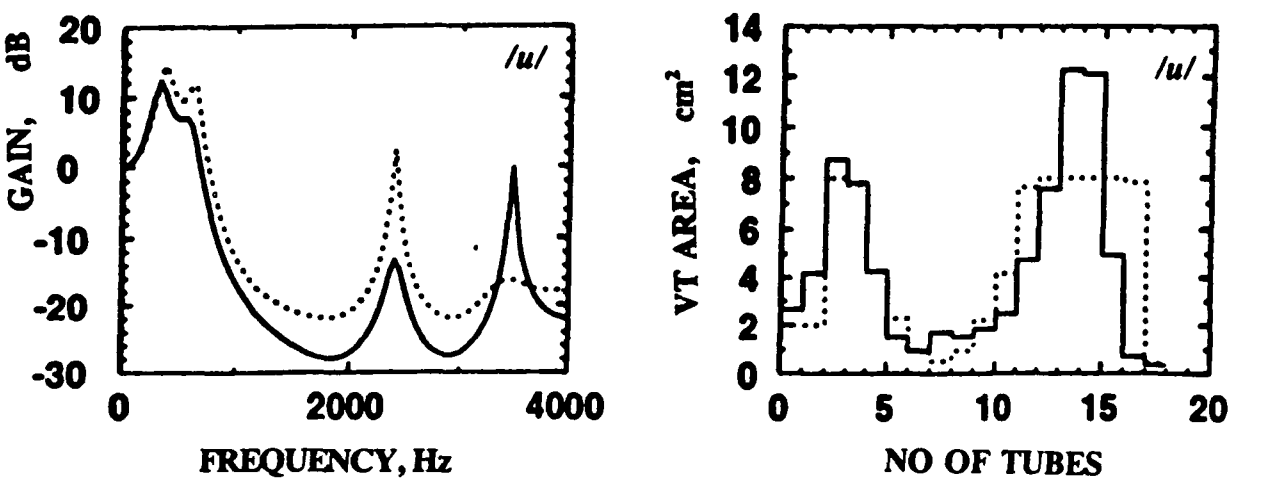


Fig. 3.18 Spectral and area function matches of the natural vowel */u/*

———: Original; .....: Estimated

3.6.3.3 Test with natural vowels

The test used for the synthetic speech was used for studying the natural vowels */a/*, */e/*, */i/*, */o/*, and */u/*.

The spectral matches and area function matches for the vowels */a/*, */e/* and */u/* are shown in Fig 3 16, 3.17 and 3.18. The area functions estimated from the natural vowels are compared with X - ray area function data. The figures show that the general shape of the spectra and the first two formant frequencies of the vowels match well with those of natural vowels In the area function domain, important parameters such as the constriction area (A_c), the constriction point (X_c) and the lip opening area (A_0) match well with those of natural vowels In figs. 3 16 - 3 18 the estimated area functions are shown together with the area functions of corresponding Russian vowels estimated from X - ray data, as presented in [Fant, 1960]

However, the matches obtained for the vowels */e/* and */o/* were unsatisfactory. This may be due to the mismatch of the lower formants F_1 and F_2 (Fig. 3 19 and 3 20). Therefore, a weighting function was used in the matching criterion (eqn.(3.7)) to increase the match in the low frequency region This method is employed in the next section.

3.6.4 Use of weighted cepstral measure in the estimation of SPM parameters

LPC analysis produces the spectral tilt in the high frequency region [Juang, et al , 1987] In addition, the glottal flow characteristics and vocal cord duty cycles are included in the low frequency region Therefore, applying a weighting function to reduce the above effects would improve the matching of spectral and hence the vocal tract area function shapes. Thus, the following band pass lifter window (eqn (3 8), [Juang, et al , 1987]) is used to reduce the sensitivity of the cepstral distance measure to the above characteristics

$$w_k = 1 + 0.5k_{\max} \sin\left[\frac{k\pi}{k_{\max}}\right] \qquad k = 1,2,\ldots k_{\max} \qquad 3\ 8$$

The use of the above window (eqn (3 8)) in eqn. (3.7) gives the weighted measure as

$$d(c,\hat{c}_i) = (c_o - \hat{c}_{io})^2 + 2\sum_{k=1}^{kmax} w_k^2(c_k - \hat{c}_{ik})^2 \qquad i = 1,2, \ldots N \qquad 3.9$$

The use of eqn. (3 9) on the vowels markedly improves both the spectral and the area function matches. Figs 3.19 and 3 20 show the matches for */e/* and */o/* using the Euclidean cepstral distance (eqn. (3.7)) and the weighted cepstral distance measure (eqn. (3 9)) The matches for vowel */i/* are improved (fig. 3.22) but not very distinctly from those obtained using the cepstral distance measure. The matches obtained for vowels */a/* and */u/* are (figs 3 21 and 3.23) the same for both measures

3.7 CONCLUSION

The formant frequencies of the model and its parameters X_c, A_c, A_0, L_0 and N, were studied It was found that the sensitivity of the articulatory parameters to the spectra can be given, for the purposes of generating the codebook, in decreasing order as X_c, A_0, A_c and L_0. This is consistent with previous studies This result was used, in extending the codebook generated from the root shapes It was demonstrated that the codebook obtained was adequate for the estimation of the area function parameters from voiced speech.

Two different cepstral measures were used in the estimation of the parameters. The parameters estimated using the current method and the constrained area function model, for synthetic voiced sounds were excellent. Estimated parameters from the natural vowels compare well with the X - ray area functio7n data for those vowels (Russian vowels). A simple Euclidean cepstral distance measure provided good area function matches for vowels */a/*, */i/* and */u/* Area function matches were unsatisfactory for */o/* and */e/* When constraints were imposed in the acoustic domain by weighting the cepstral coefficients (to increase the match in the lower and middle frequency region) good matches were obtained for */o/* and */e/*. Improvements for /a/, /i/ and /u/ were not noticeable. This is probably, because a very close match for the vocal tract cross sectional area had already been obtained using the basic technique

Improvements can be made to reduce the codebook size and hence to decrease

the estimation time of the parameters These methods, e g clustering the codebook using k - means algorithms, tree type codebooks, and preliminary results are discussed in chapter 5

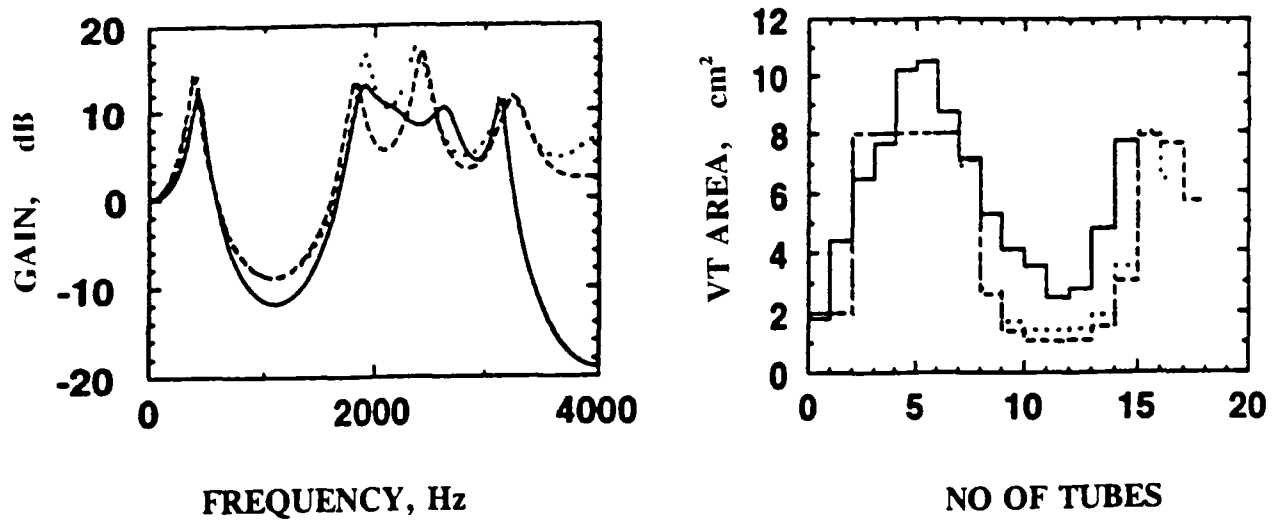


Fig. 3.19 Spectral and area function matches of the natural vowel /e/

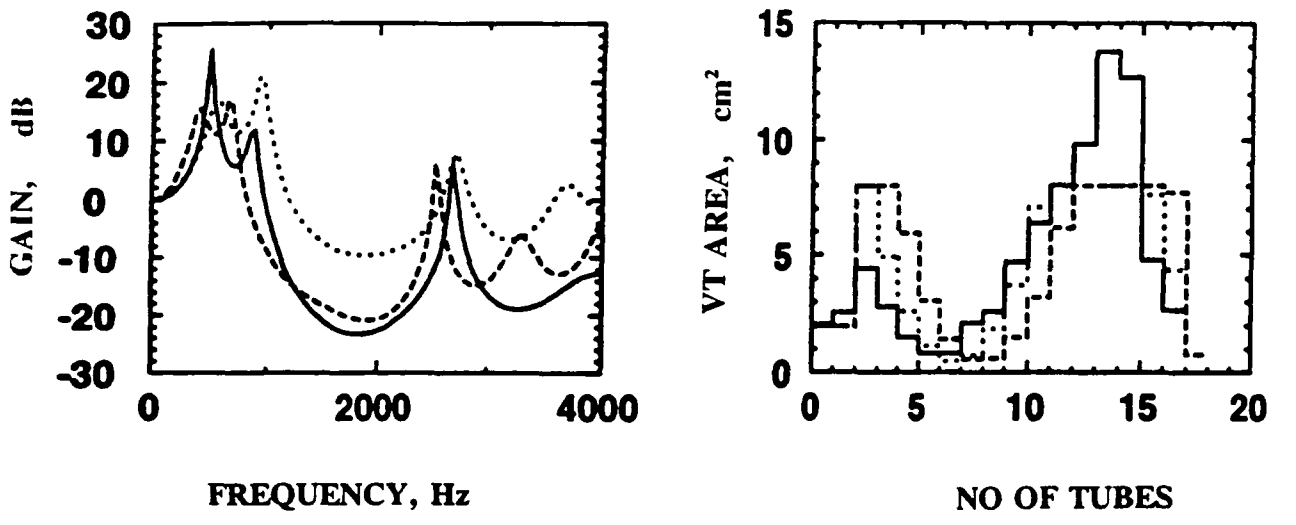


Fig. 3.20 Spectral and area function matches of the natural vowel /o/

. - .....: Using the weighted cepstral distance measure.

-------: Using the unweighed cepstral distance measure.

——: LPC spectra of the given vowels; X - ray area function [Fant,1960] data.

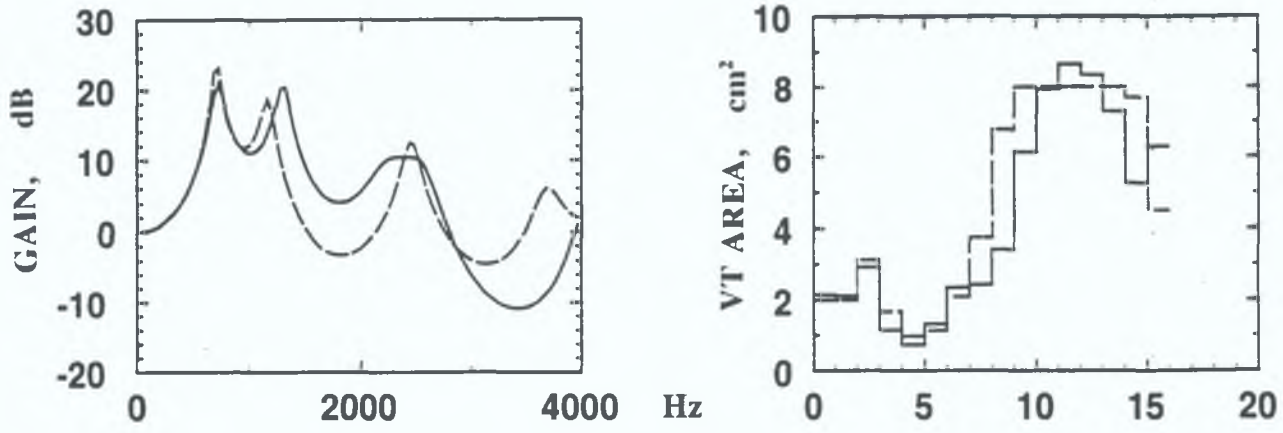


Fig. 3.21

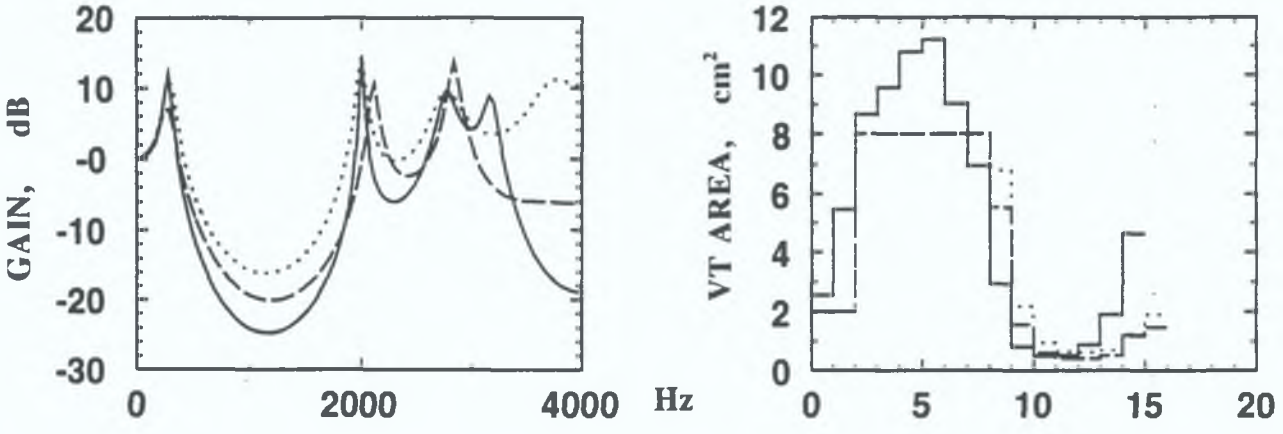


Fig. 3.22

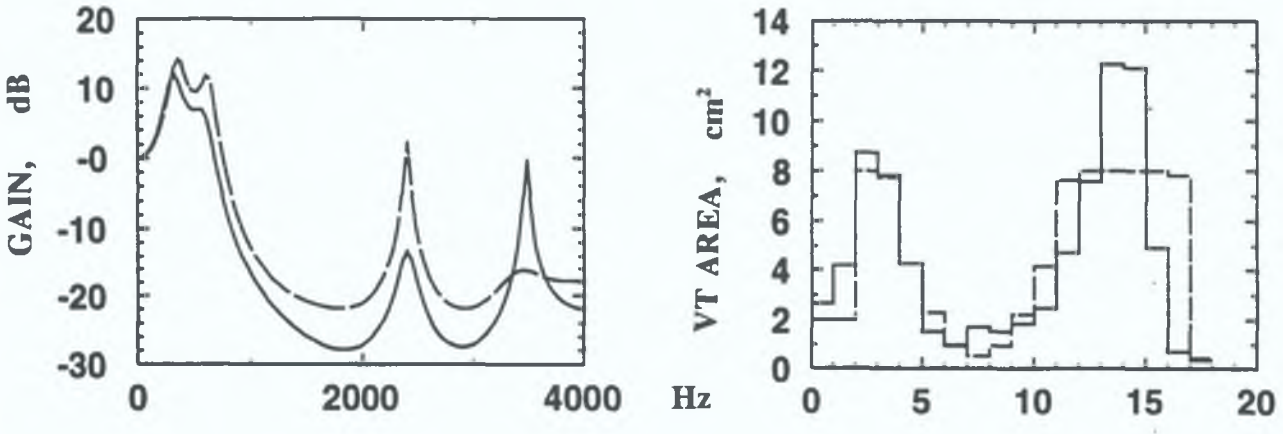


Fig. 3.23

Figs. 3.21, 3.22, 3.23 Spectral and area function matches of the natural vowels */a/*, */i/* and */u/* respectively.

.....: Using the weighted cepstral distance measure.

– – –: Using the unweighed cepstral distance measure.

———: LPC spectra of the given vowels; X - ray area function [Fant,1960] data.

CHAPTER 4

ESTIMATION OF GLOTTAL FLOW PARAMETERS USING VOCAL TRACT AREA FUNCTIONS IN INVERSE FILTERING TECHNIQUES

4.0 INTRODUCTION

Estimation and parameterisation of the air flow through the vibrating vocal folds are discussed in this chapter. The Speech Production Model (SPM) used in this work was discussed in section 2.3. In chapter 3, the vocal tract parameters of the SPM were estimated from speech signal using linked codebook techniques These estimated SPM parameters are used in the calculation of the input signal of the human vocal tract apparatus i.e. glottal flow.

Estimation of the glottal flow using the inverse filtering method is discussed in sections 4.2 and 4 3. The inverse filter is, in this project, constructed using the area function estimated in chapter 3 The use of the area function is not common in the previous inverse filtering techniques

Different glottal flow signals produce speech with various quality characteristics and voice types. Reliable estimation of the glottal flow has been a subject of study for several years [Flanagan et al , 1975; Sondhi, 1975; Rosenberg, 1971; Milenkovic, 1986; Alku, 1992]. The object of this study is to estimate the glottal flow by noninvasive and uncomplicated procedures A simple parametric model and an accurate characterisation of glottal flow have a number of possible applications in synthesis, coding and recognition systems For example

a) Use of an accurate model of the excitation signal in the SPM produces better quality speech than using impulses and white noise [Cummings and Clements, 1992]

b) Although glottal wave analysis is not often applied in the speech coding area, a simple parametric model for the flow can produce high quality low bit rate speech coders. [Hedelin, 1986 & 1988, Alku, 1990] discuss the use of the glottal flow signal in low bit rate coders

c) Laryngeal pathology (vocal fold abnormalities) may be detected by monitoring the shape of the flow signal [Milenkovic, 1986, Price,1989]

d) Different voicing types e g. normal, breathy, vocal fry, modal, falsetto etc , can be classified from the shape of the flow waveforms (i e width, skewness and closing time) [Karlsson, 1988; Childers and Lee, 1991; Holmberg et al , 1988]

4.0.1 Glottal flow measurement techniques

As in the case of direct area function measurements, direct glottal flow measurements are difficult to make but can be obtained. Available methods can be divided into four categories [Sondhi, 1975]·

a) cinematography

b) optical method

c) electrical and mechanical methods

d) acoustical methods.

In measuring the vocal fold movements using the cinematography technique, the vocal cords are illuminated and filmed with the help of a small mirror that is placed at the back of the subject's mouth [Farnsworth, 1940, Alku, 1992] In the optical method, images of the vocal cord movements are captured by using a light source and an optical sensor positioned near the vocal cords using a fibrescope [Kiritani et al , 1990] These images can then be processed to obtain information about the vocal cord movements

In the electrical study, the Electro Glotto Graph (EGG) was used to study the vocal fold movements [Lee and Childers, 1989] The EGG instrument measures the electrical impedance variations of the larynx using a pair of plate electrodes held in contact with the skin on both sides of the thyroid cartilage. A radio frequency current of 1 MHz is supplied to one skin electrode and detected by the sensing electrode. The radio frequency current is amplitude modulated by the varying tissue impedance due to the vibration of the vocal folds. This modulated current is then demodulated by the detector circuit of the EGG instrument This demodulated waveform (i e the EGG signal) represents the opening of the glottis as an upward deflection and the closing of the glottis as an downward deflection [Krishnamurthy and Childers, 1986]

Due to practical difficulties and the invasive nature of these equipments acoustical methods are preferable to the above methods

A simple acoustical method was suggested [Sondhi, 1975], in which a microphone was directly used to record the glottal waveforms The effects of vocal tract resonances on the glottal flow were eliminated by speaking into a reflectionless closed uniform tube A hard walled uniform tube with the cross sectional area identical to that of the vocal tract is used The subject is then asked to utter a non nasalised neutral vowel into the tube The output signal measured at any point along the tube,

by a microphone represent the glottal flow.

Shortcomings of this method are obvious Flow is measured by fixing a reflectionless uniform tube at the lip for a fixed tract shape Therefore, the method cannot be applied to normal speech Also, the vocal tract is assumed to be hard walled, which is not true. For these reasons, inverse filtering techniques [Alku, 1992, Price, 1989, Veeneman and BeMent, 1985] have been employed for the estimation of flow signals from speech signals

4.0.2 Inverse filtering technique for glottal flow measurements

The general acoustical method used for obtaining the glottal flow signal is inverse filtering. A very simplified idea of this method is depicted in the block diagram fig 4 1 Making some assumptions about the input signal G(n) (G(n) is the digitised form of the continuous glottal flow signal G(t)), a SPM with transfer function H(z), is found from the speech signal S(n) As shown in fig 4 1b, the speech signal S(n) is then filtered through the inverse transfer function $H^{-1}(z)$, to obtain the input signal G(n)

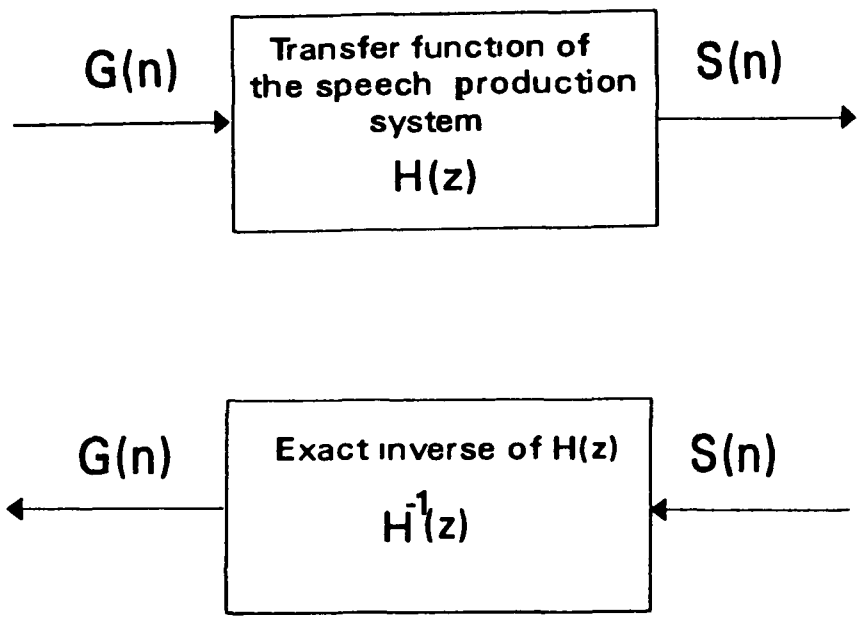


Fig. 4.1 a): Simplified functional model of speech production system

b): Inverse filtering model of the speech production system

Inverse filtering methods based on LPC type models [Alku, 1989], have been applied to speech signals to separate source characteristics from vocal tract characteristics The separation is performed by the cancellation of the vocal tract characteristics from the speech signal. In other words, speech signals are filtered through the inverse transfer function of the vocal tract Therefore, in theory, what is left in the speech signal after the inverse filtering, is the glottal source characteristics and the lip radiation characteristics [Price, 1989]. Further cancellation of the lip radiation characteristics would then give the glottal flow alone

The quality of the inverse filtered signal (i e how well it matches the true flow) depends on the precise cancellation of the formant frequencies. Several different methods have been used in the construction of LPC type models The speech signal during the closed phase of the glottal cycle, during which the source/tract interaction is minimum, has been used for estimating the LPC filter [Veeneman and BeMent, 1985, Wong et al , 1979]. These filter coefficients can then be used in the inverse filter

Recently an Iterative Adaptive Inverse Filtering (IAIF) method was used for the estimation of flow signals [Alku,1992] In this method the speech production system is modelled by three separate parts· glottal excitation, vocal tract and lip radiation sections The interaction between the sections was assumed to be negligible in the modelling.

4.0.3 Inverse filtering technique used in the current research

In this work, the vocal tract area function shapes, estimated from the speech signals (c.f. chapter 3), are used to construct the inverse filter This method, which uses the area function parameters, is not common [Ananthapadmanabha, 1984] in previous speech research

A method is described here, to extract the glottal pulse waveform from both the Inverse Filtered (IF) flow and the Differentiated Inverse Filtered (DIF) flow signals of the SPM. This method automatically decouples the vocal tract and the radiation effects from speech signals to produce the glottal flow, using the inverse lattice filter model The glottal flow is then parameterised using the LF glottal flow model [Fant, 1985]

4.1 REVIEW OF GLOTTAL FLOW MODELS

Unlike the complex speech signal, the glottal flow is a relatively smooth waveform This waveform can be approximated by simple parametric models in the time domain The glottal flow is related to the actual movements of the vocal folds The movements of the vocal folds are usually described by the opening phase, closing phase, open phase, closed phase, fundamental period of oscillation, maximum flow etc Thus, these same terms are used to describe the parameters of glottal flow models

A number of flow models have been suggested - the derivative of the glottal flow, the glottal flow itself and the glottal area function [Gupta and Schroeter, 1991]. One of the earliest is the Rosenberg glottal flow model (see appendix D1) defined by sinusoidal segments [Rosenberg, 1971] Several other models are also available, for example vocal fold movements approximated by a two mass model [Ishizaka and Flanagan, 1972], glottal flow derivatives described by sinusoidal and polynomial segments [Lobo and Ainsworth, 1989], a five parameter model [Ananthapadmanabha, 1984], a three parameter model [Liljencrants model, 1984], the "F" model [Fant, 1979] and a four parameter model also called LF model [Fant, Liljencrants and Lin, 1985], etc In this work, the later four parameter model (also called the LF model) which closely matches the natural flow is chosen to parametrise the inverse filtered speech (see also section 4.3 1 for the basis of this choice).

Another important feature of the flow signal is the turbulent noise component This turbulent flow which occurs due to high air flow rate through the glottis near glottal closure, is included in a model [Childers and Lee, 1991] described by combining the LF model and a turbulent noise generator

4.1.1 Two mass models of glottal excitation

Vocal fold vibration during speech production is due to the subglottal pressure that drives the vocal cords apart; and the muscular, elastic and Bernoulli restoring forces that draw them together These effects (i e, the vocal fold movements) are approximated by a two mass model in [Ishizaka and Flanagan, 1972] This model, described in terms of glottal area, sub-glottal pressure, cord tension parameter and the current vocal tract shape, can be used to estimate volume flow at the glottis (vocal folds).

Modified two mass models were described in [Koizumi et al., 1987] These models were used to analyse the effects of interaction between the source and tract e g. *skewing, truncation, dispersion* and *superposition* Skewing is a phenomenon in which glottal pulses are skewed to the right compared to the glottal area functions, due to the inertance of the vocal tract load The damping of the vocal tract resonances during the glottal open phase due to coupling of the subglottal system and the time varying nonlinear glottal impedance, is termed as the truncation [Nord et al., 1984] The effect of truncation can be observed in spectrograms of natural speech The term superposition expresses the fact that acoustic energy is carried over from one glottal period to the next.

The two mass model and the modified two mass models described in [Koizumi et al., 1987], are more suited for interactive modelling of speech and its glottal flow classifications For the modelling of glottal flow in the current work, the models defined in terms of flow parameters are more relevant. These models are described below.

4.1.2 A four parameter model of derivative of glottal flow

A model for the derivative of the glottal flow model (eqn. 4 1) with 4 independent parameters was proposed in [Fant, Liljencrants and Lin, 1985] The parameters t_p, t_e, t_a and E_e are shown in fig. 4.2. Three of these parameters describe the angular frequency ω_g of the sinusoid (equals to π/t_p), scaling factor of the model E_0 (depending on E_e and t_e) and the exponential growth constant α (depending on E_e and t_e) of the sinusoid. The fourth parameter (t_a) is defined by the projection on the time axis, of the tangent to an exponential approximation to the return curve. The exponent ϵ in eqn (4 1b) is related to the parameter t_a. The fundamental criterion used in the estimation of the four parameters is that the flow return to zero in each cycle

$$\frac{dU_g(t)}{dt} = E(t) = E_0 e^{\alpha t} \sin \omega_g t \qquad 0 < t < t_e \qquad \text{4 1a}$$

$$\frac{dU_g(t)}{dt} = E(t) = \frac{E_e}{\epsilon t_a}\left(e^{-\epsilon(t-t_e)} - e^{-\epsilon(T_0-t_e)}\right) \qquad t_e < t < T_0 \qquad \text{4 1b}$$

The LF model is, in fact an improved version of models previously developed by [Fant, 1979] and [Liljencrants, 1984] The model allows for an incomplete closure or a residual phase of progressive closure after the major discontinuity occurs (at the excitation point t_e) in the natural speech. The model produces optimum results for non interactive glottal flow parameterisation [Fant et al , 1985] Fig. 4 2 shows a typical glottal flow and the differentiated glottal flow of the model

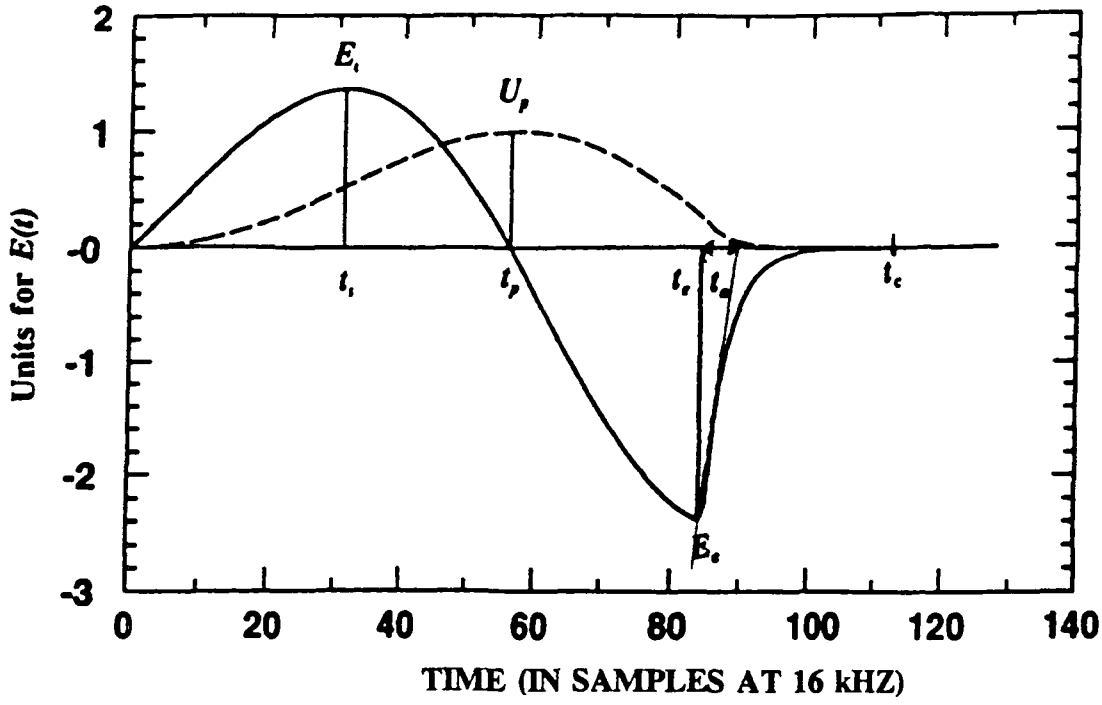


Fig. 4.2 Differentiated LF glottal flow *E(t)* (solid) and the corresponding glottal flow *U(t)* (dashed). The glottal flow are normalised to maximum excursion of one.

U_p = peak flow; E_i = maximum positive rate of change in the flow

E_e = negative level of flow derivative at the excitation point

t_p = glottal flow peak position; t_e = excitation point

t_a = time constant of the return phase from the maximum closing continuity towards the maximum closure

The following values are used to synthesize the signals shown above:

f_s = 16 kHz, $t_c = T_0 = 1/F_0$ = 8 ms, t_e = 5.3 ms, t_p = 3.5 ms, t_a = 0.2 ms.

where T_0 and f_s are fundamental period and sampling frequency.

4.1.3 Other models used in the parameterisation of glottal flow

Like the four parameter model mentioned above, non abrupt flow termination is also allowed in a model described by five parameters [Ananthapadmanabha, 1984].

The L model described by three parameters is given in appendix D2 [Liljencrants, 1984].

A model, proposed in [Fant, 1979] by ascending and descending branches of the glottal flow is also used for parameterising glottal flow signals. This model, described by cosine segments, has a discontinuity at the flow peak, thus giving secondary weak excitation at this point This model shares the same t_p parameter with the L model

4.2 ESTIMATION OF GLOTTAL FLOW FROM SPEECH SIGNAL

For the SPM described in fig 4.1, we need an estimate of both the input signal G(n) (source signal) and the transfer function H(z) (this includes the effects of vocal tract and lip radiation) of the SPM from the output speech signal S(n) Estimation of the transfer function of the system from the output signal only (i.e without the input signal), is defined as the indeterminant problem in control theory. However, in speech analysis the transfer function H(z) is estimated by making some special assumptions about the excitation signal The usual assumption made for these purposes is the white noise/impulse model. Estimation of the vocal tract filter during the closed phase (during which no excitation exists) of the glottal cycle also produces good results The use of these assumptions in speech analysis improves estimates of the vocal tract transfer function

An estimate of H(z) together with the original speech signal S(n) can then be used to re-estimate the input (excitation) signal G(n) Fig 4 1b shows the basic inverse filter model used in different systems.

4.2.1 Construction of inverse filter model

The SPM designed in chapter 2 was used for the inverse filtering of 5 natural and 6 synthetic vowels Parameters of the inverse filter (equivalent to estimating H(z)) were calculated using the area functions estimated for these vowels in chapter 3 The inverse filtered signal obtained is parameterised here, using the LF [Fant et al , 1985] glottal flow model

The construction of the inverse filter model involves two steps

a) Extraction of SPM (area function) parameters A_c, X_c, A_0, L_0 and N from the speech signal The estimation of these parameters from voiced speech signals using the linked codebook was discussed in section 3 6 The SPM parameters are estimated by minimising the total mean squared cepstral distance $d(c,\hat{c})$ between the given cepstrum and the codebook cepstra (equivalent to Log Spectral Distance (LSD) between the given and the codebook spectral shapes) given in eqn 4 3 (see section 3 6) As discussed in section 3 6.3, the cepstral measure is weighted for the estimation of the parameters

$$d(c,\hat{c}_i) = (c_o - \hat{c}_o)^2 + 2\sum_{k=1}^{kmax} w_k^2 (c_k - \hat{c}_k)^2 \qquad i=1\ldots1635 \qquad 4\ 3$$

where c_k, ,$\hat{c}_k$ are the given and estimated cepstral coefficients and $kmax$ (equal to 25 in the current work) is the number of cepstral coefficients used for the estimation

b) Construction of the inverse filter model The area functions estimated in section 3 6 are used for the construction of the inverse filter model The coefficients $r_k = \dfrac{A_k - A_{k+1}}{A_k + A_{k+1}}$ of the inverse filter are calculated from the estimated area functions; where A_k and A_{k+1} are the cross sectional areas of successive vocal tract tubes Fig 4 3 shows the lattice inverse filter used for the inverse filtering of speech pressure waveforms. This filter is the same as the inverse filter described in fig. 2 2b, except that the inverse characteristic of the lip radiation model Z_L^{-1} (eqn. 4 4) is included at the beginning of the filter The filter Z_L^{-1} is the exact inverse of the radiation model described by Laine's pole/zero digital model in eqn 2 17 The filter Z_L^{-1} is given by.

$$Z_L^{-1} = \frac{1}{Z_L} = \frac{A}{\rho c a}\frac{(z-b)}{(z-1)} = \frac{A}{\rho c a}\frac{(1-bz^{-1})}{(1-z^{-1})} \qquad 4\ 4$$

where a & b are functions of the lip opening area A The values of a,b and the

constants are the same as described in eqn 2 17.

The inverse filter model has been implemented as a *"C"* program

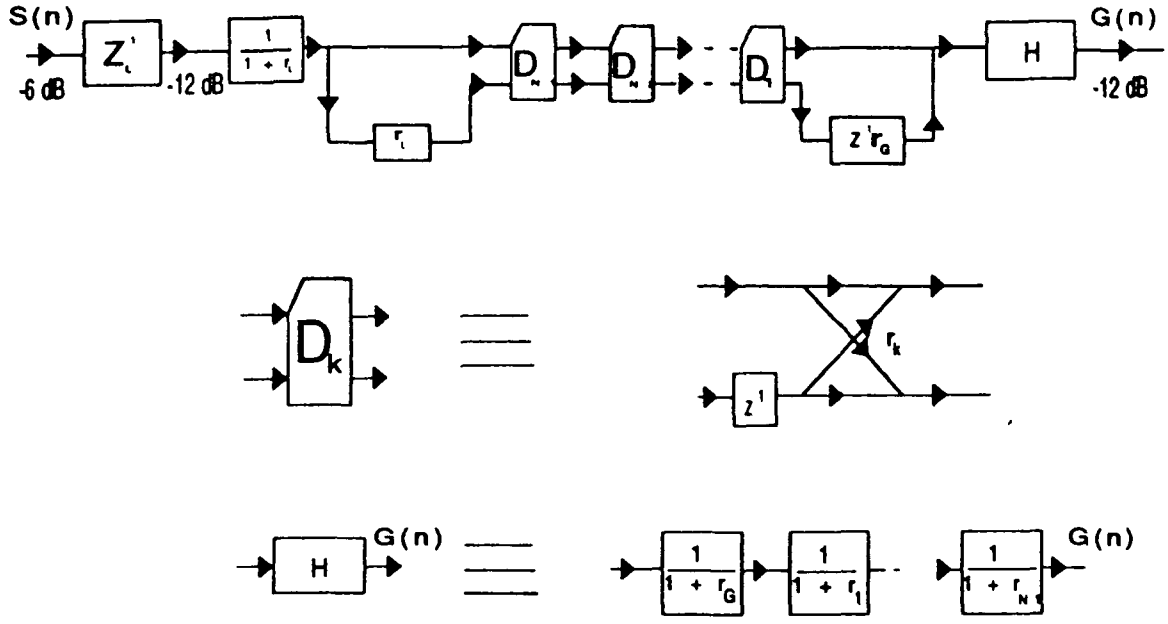


Fig. 4.3 Inverse lattice filter model, inverse characteristic of lip radiation is also included at the front of the lattice filter.

4.3 INVERSE FILTERING OF SPEECH USING THE MODEL

The Inverse Filtered (IF) signal G(n) and the Differentiated Inverse Filtered (DIF) signal G'(n) obtained, using the inverse filter shown in fig. 4.3, is described in this section The Differentiated Inverse Filtered (DIF) signal G'(n) is parameterised using the LF model. Both the G'(n) and the G(n) signals are used in the parameterisation of the signal G'(n) (see section 4.4)

The speech pressure waveform S(n) is passed through the constructed inverse filter (fig. 4 3) Speech waveforms S(n) , as is well known, have a characteristic -6 dB/oct roll off in the frequency domain.

Conversion of lip flow to pressure at the lips (i.e. the lip radiation characteristic) is, as mentioned, modelled by Laine's pole-zero filter Z_L (eqn (2 17)), designed to give an average +6 dB/octave rise in the spectral domain Therefore in the inverse filtering, S(n) is passed through the inverse of the radiation filter Z_L^{-1} This gives a

signal with a characteristic -12 dB/oct fall off, after the first stage of the inverse filtering as shown in fig. 4 3

This signal is then passed through the inverse vocal tract filter The signal obtained at the end of this process describes the IF signal (i e glottal flow signal) G(n) The spectral roll off values during the inverse filtering process at the radiating surface (at the lips, -6 dB/oct) and at the glottis - vocal tract junction (-12 dB/oct) of the model are indicated in fig. 4 3 for clarity

The DIF signal, which is modelled in the current work, is then obtained from the IF signal, by a differentiation process $(1 - z^{-1})$

4.3.1 Selection of a glottal flow model for the parameterisation of DIF signal

The LF model is chosen to parameterise the DIF signals obtained in this work The DIF signal G'(n) is, in general, parameterised with reference to the important factors of the voice signal such as glottal pulse width, pulse skewness and the abruptness of the glottal pulse closure Several glottal flow models, available for the modelling of flow were discussed in section 4 1.

All the models discussed in section 4.1 allow variable pulse width and skewness Only the LF model [Fant et al , 1985] and the five parameter model [Ananthapadmanabha, 1984] allow variation in the abruptness of glottal closure Other models produce abrupt closure. Therefore the LF model which varies all three factors with only four parameters was selected to parameterise the flow signal

4.4 A METHOD TO ESTIMATE THE LF PARAMETERS FROM DIF AND IF SIGNALS

The four parameters t_p, t_e, t_a and E_e which uniquely define the LF model, are measured/estimated from the IF and DIF signals (see fig 4 5) Different matching methods have been used in the estimation of LF flow parameters in the literature In [Childers and Lee, 1991], the least - mean - square error between the DIF and LF model was minimised In [Fant et al , 1985], the ratio of the maximum negative peak to the maximum positive peak of the DIF signal (E_e/E_i) was matched to estimate the parameters These methods would produce, good results for clean DIF signals However, these methods cannot be used for DIF signals with a large number of ripples

The DIF signals obtained here, using the inverse filter model contain several formant ripples, especially in the closed phase of the glottal flow (fig. 4.4). These ripples may be due to the source-tract interaction or improper formant cancelling of the signal. This restricts direct measurement of the LF model parameters from the DIF signals.

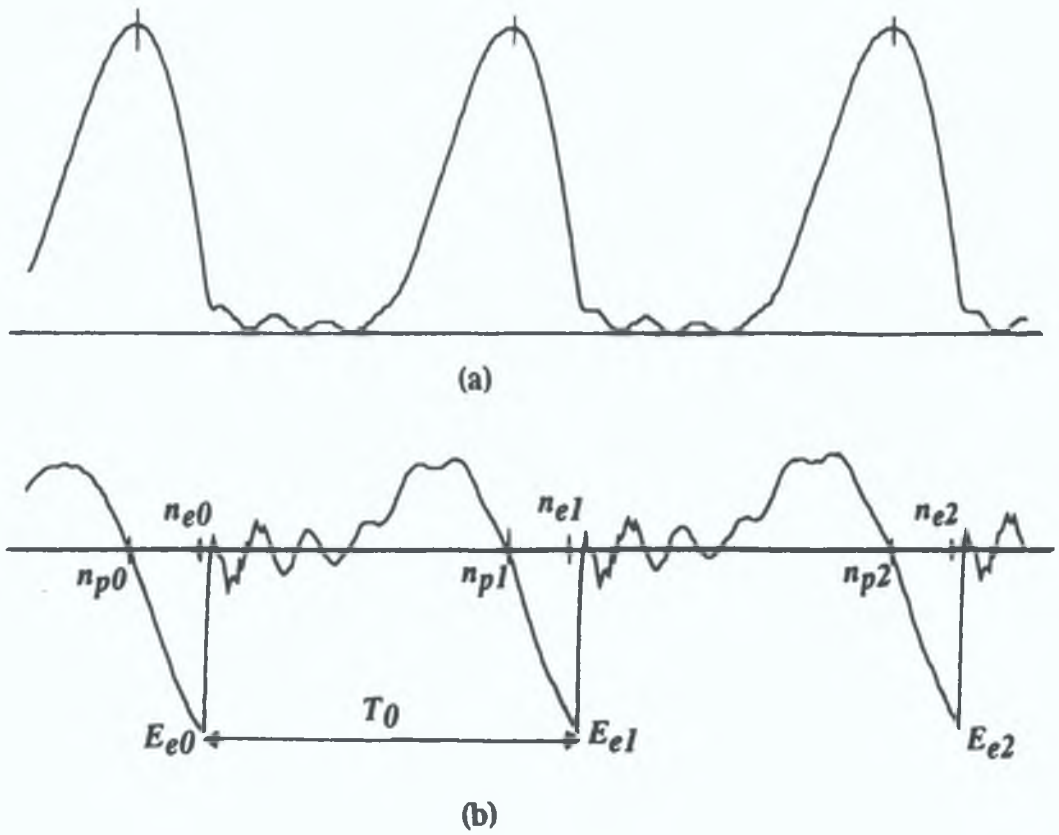


Fig. 4.4 Typical inverse filtered signals shown with critical points, required for the measurements of T_0, t_{dep} and E_e

a) IF signal corresponding to the DIF signal shown in (b)

b) DIF signal

It may be observed from both the DIF signal (G'(n)) (fig. 4.4) and the IF (G(n)) (fig. 4.4) signal that certain parameters of the LF model can be easily determined from IF speech while others may be more easily determined from DIF speech.

The matching method described in the current work is robust for a DIF signal with many ripples. Estimates for the parameters are found from both DIF and IF signals. The use of IF signal for determining the zero crossing point of the DIF signal

(i.e. peak glottal flow point at t_p) is found to be more reliable than the use of DIF signal itself. The method needs only two area values under the DIF signal at two points where they can be estimated precisely. The points t_p and t_e are chosen as these two points

Estimation of each parameter is described in the following sections Section 4.4.2 describes the estimation of the parameters E_e, t_p, t_e and the fundamental period of the signal T_0. Section 4 4 3 describes the estimation of the parameter t_a These algorithms are then used for the parameterisation of the DIF signal, obtained from natural and synthetic speech.

4.4.1 LF model estimation

In LF modelling, only three parameters t_p, t_e, and E_e are required to define the first part of the model, from $0 < t < t_e$ This is basically the L model [Liljencrants, 1984], directly defined by three parameters ω_g (π/t_p), α and E_0 which are closely related to the parameters t_p, t_e and E_e However, the parameter t_a together with other parameters is required (directly or indirectly) to define the return phase ($t_e < t < T_0$) of the LF model (eqn. 4.1b)

4.4.1.1 Difficulties in arriving at an analytical solution for the LF model

In the estimation of the model parameters, the first part of the DIF is (i e $0 < t < t_e$) fitted to the L model. A complete analytical solution for the L model defined by three parameters, requires three equations from the DIF However, obtaining three equations from the DIF or IF signals is not possible This is because the zero glottal flow point ($t = 0$ in fig 4.2) from which all other time parameters are defined, cannot be reliably estimated from the DIF or IF signals, due to several ripples (see DIF and IF signals in fig. 4.4)

4.4.1.2 A solution for the LF model by area matching

Only one equation can be written at t_e in terms of the measurable (known) quantity E_e, This equation is written as·

$$E_e = E_0 e^{\alpha t} \sin \omega_g t_e \qquad 4.5$$

Writing the E_0 term in the left hand side of the equation gives

$$E_0 = \frac{1}{E_e e^{\alpha t} \sin \omega_g t_e} \qquad 4.6$$

where t_e is an addition of t_p (unknown) and a known measurable quantity (t_e - t_p).

The other two parameters α and t_p (equivalently ω_g) are estimated by matching the areas under the DIF curve and model curve at two different points between The knowledge of E_e (i.e eqn. (4 6)) is also used in the above matching of areas The estimation of these three parameters E_e, ω_g and α, is described in the next section As mentioned E_e is measured directly from the DIF curve (fig 4.5)

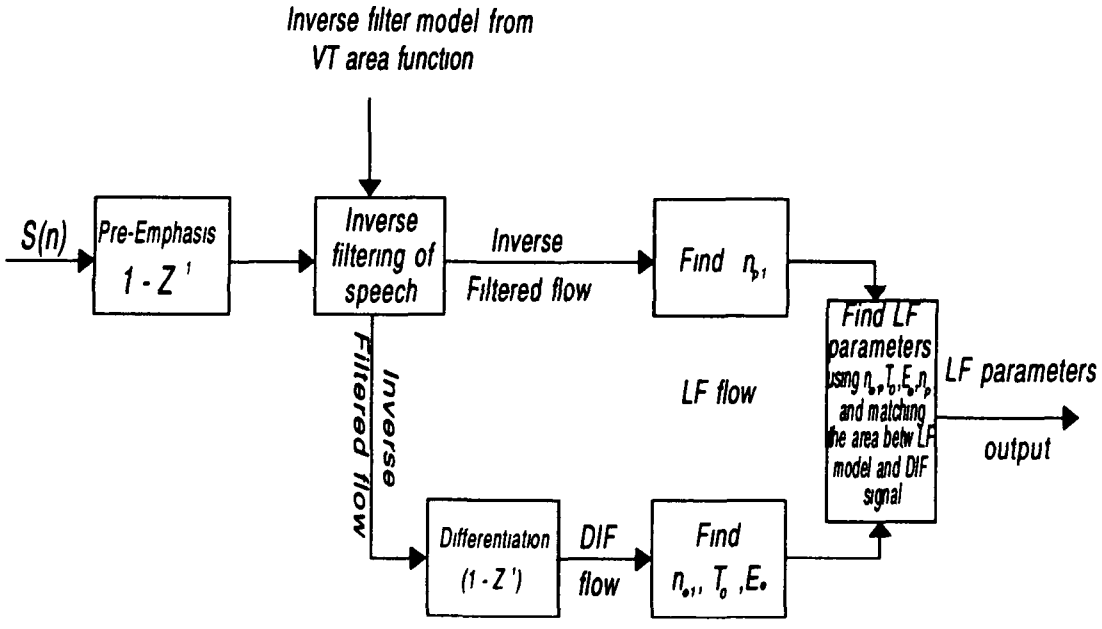


Fig. 4.5 Block diagram of the parameterisation of DIF signal using the LF model

4.4.2 An algorithm to estimate T_0, ω_g, α and E_e from DIF and IF signal

As mentioned previously, both the DIF and IF signals are used in the estimation of the above parameters The quantities E_e, T_0 and t_{dep} (i.e t_e - t_p) are measured directly from

the DIF and IF signals This is described in section 4.4.2.1

However, to fully estimate the values of t_p and t_e, it is necessary to determine the point at which zero flow occurs (see fig. 4.4) The determination of this point is not trivial as the DIF signal has several ripples in the closed phase. Therefore, as mentioned in the last section, the actual value of ω_g (t_p) and the value of α are determined by matching two areas of the DIF signal with the model signal The areas are matched by varying the parameters ω_g and α. The "simplex" method of optimisation is used to obtain the optimum values for the parameters ω_g and α. This is explained in section 4.4 2 2

Note that t and n in the following paragraphs are used to indicate the continuous value (analog) and the corresponding discrete (digital) value Dividing the digital value n by the sampling frequency would produce the equivalent time Therefore the sample values n_{e1}, n_0, etc., are sometimes used in the following section, to mean the time domain signal Time domain values are rounded of to their closest sample point

4.4.2.1 Measurements of E_e, T_0 and t_{dep}

The DIF signals are normalised to the maximum excursion of one by dividing by the absolute maximum value of the DIF signal With reference to the DIF and IF signals shown in fig 4 4, the estimation of the above parameters involved the following steps

a) *Estimation of the fundamental period T_0:* T_0 is estimated from the DIF signal. The distance between the two negative maxima (e g (n_{e1} - n_{e0}) or (n_e - n_{e1})) was searched over a 3 ms to 15 ms range (i e; 48 samples to 240 samples) which reperesents the T_0 range of male adult speakers The range 3 ms to 15 ms contains the usual T_0 value for most speech material This restriction is imposed to avoid the algorithm finding multiples of T_0 (i.e. for example to avoid taking (n_e - n_{e0}), which would give twice the T_0) rather than the actual T_0. The parameters E_e and t_{dep} are measured with reference to n_{e1} , assuming that the T_0 equals (n_{e1} - n_{e0}) If T_0 = (n_e - n_{e1}), the reference point would be n_e

b) *Measurement of E_e.* This is the value of DIF signal at n_{e1}. Since the DIF is normalised to one, this value is close to unity This value is used to find the value of E_0 using eqn. (4.6).

c) *Measurement of* $(t_e - t_p) = t_{dep}$ The difference between n_{el} and n_{pl} The n_{el} value can be easily measured from the DIF signal However, the zero crossing point (n_{pl}) of the DIF signal is difficult to determine, this especially applies to real speech signals, due to ripples around this point. The zero crossing point in the DIF signal corresponds to the point at which maximum flow occurs in the IF signal This point n_{pl}, can be more precisely identified from the IF signal. Rather than taking the point n_{pl}, directly from the IF signal, an interpolation technique is used around this point for more precise determination The parabolic interpolation, of five points around the maximum flow, is performed to find the point n_{pl} The difference between the two points (n_{el} - n_{pl} = n_{dep}) is then stored for later use, mainly to find t_e The values of T_0, E_e and t_{dep}, determined here are used in the next section

4.4.2.2 Determination of ω_g (t_p) and α using an Optimisation method

The values of ω_g and α are estimated by fitting the first part of the DIF signal ($n_0 < n < n_{el}$) to the first segment of the LF model eqn (4.1a) ($0 < t < t_e$), where n_0 is the beginning of the opening phase of the LF model (see fig 4 6b) The exact position of this point n_0 is dependent on the value of t_p and α

4.4.2.2.1 Selection of the beginning of the opening phase for optimisation

Due to several ripples in the closed phase (fig 4.4 and fig 4.6 on pp 95), it is difficult to detect n_0, the beginning of the opening phase, from the DIF signal However it is reasonable to assume that the n_0 lies between $n_{e0} < n < n_{pl}$ With this assumption for the point n_0, the values for n_p and n_e are estimated as.

$n_p = n_{pl} - n_0$; $n_e = n_p + n_{dep}$, where n_{dep} is found in the previous section

With these values of n_p (t_p), $n_e(t_e)$ and E_e (from the previous section) the first segment of the LF model was synthesised

4.4.2.2.2 Matching the areas under the LF model and the DIF signal

The optimum position (i.e. optimum value) of n_0 is found by matching the areas under the DIF and synthesised LF model curve at two points by varying the parameters t_p and α The areas between n_0 and n_{el} (U_p^M and U_p^R), and n_0 and n_{pl} (U_e^M and U_e^R) are calculated for both the model and the DIF signal (the superscripts M in R stand for the

Model and the Real DIF signal). The cost function $(U_p^M - U_p^R)^2 + (U_e^M - U_e^R)^2$ is then minimised with respect to the parameters t_p and R_d; where $R_d = 2\alpha/\omega_g$. The conditions and the steps involved in the optimisation of the parameters t_p and α are described below:

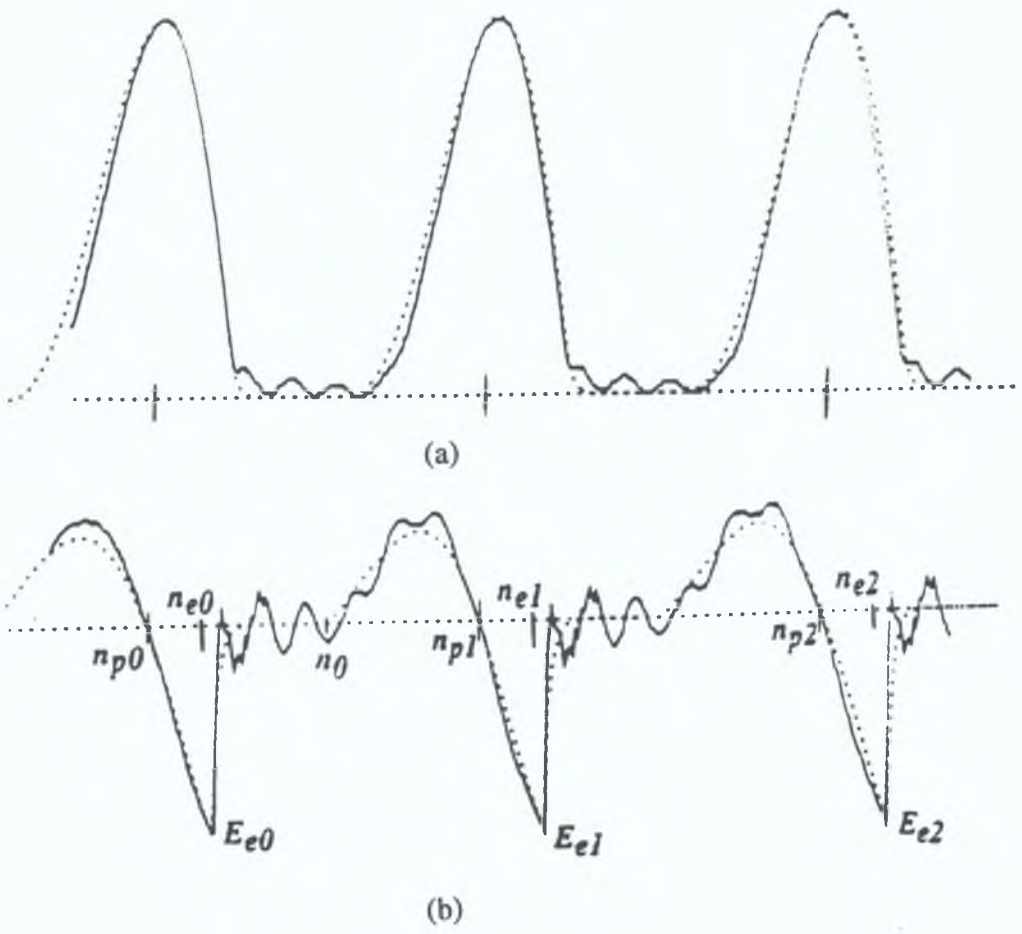


Fig. 4.6 **a) A LF model flow signal is superimposed on an IF signal.**
b) A LF model signal is superimposed on the DIF signal corresponding to fig. 4.6a.
Beginning of the open phase n_0 is shown. However the optimum position of n_0 is found by matching the area under LF curve and the DIF curve. _____: IF filtered curve .....: LF model curve

a) *Range of* t_p (n_p): In the optimisation, n_p (t_p) is allowed to vary between $(n_{dep} + 1$ sample) i.e. $t_{dep} + \delta$ and $(N_0 - n_{dep})$ i.e. $(T_0 - t_{dep})$ range. The value n_{dep} is chosen as the

minimum value for n_p because the glottal flow signals have usually longer opening phase than closing phase (*viz* fig. 4.4 and 4 6). The maximum value is chosen as (T_0 - t_{dep}) to allow for a zero length closed phase

b) *Range of* α *(*R_d*)*: In the actual algorithm R_d, instead of α, is allowed to vary directly. R_d is related to α as: $R_d = 2\alpha/\omega_g$. From past experience of the LF model [Fant *et al.*,1985], it was realised that values of R_d between 0.01 and 2 provide good matches for real glottal flow signals. Therefore, this range is selected for varying the parameter R_d (α)

c) *Targets for optimisation* Areas under n_0 to n_p and n_0 to n_e are matched between the LF model and DIF signal (fig 4.4, 4.6). Knowing (assuming) the values for t_p and α, allows for synthesis of the first segment of the LF model (eqn (4.1a)) Therefore, the areas (value of the flow) between n_0 and n_p, U_p^M, and between n_o and n_e, U_e^M are estimated by integrating the first segment of the LF model The areas at the corresponding points U_p^R and U_e^R are also found for the DIF signal, by integration The trapezoidal rule is applied for integration of both DIF and the LF model signals

d) *Minimisation (cost) function* . $(U_p^M - U_p^R)^2 + (U_e^M - U_e^R)^2$ is minimised with respect to the parameters t_p and R_d using the simplex method of optimisation [Press *et al* , 1988]. Three sets of initial values are given to (t_p,R_d), in the range described in a) and b) The optimisation is stopped when there are no significant changes in the parameter values

The parameters t_p and R_d, determined in the optimisation are then used to find the actual parameters ω_g, α and E_0 as follows:

$$\omega_g = \frac{2\pi}{t_p}; \qquad \alpha = \frac{R_d \omega_g}{2}; \qquad E_0 = \frac{1}{E_e e^{\alpha t_e} \cdot \mathrm{sm}(\omega_g t_e)}$$

where t_e is given by

$t_e = t_p + t_{dep}$

The procedure explained above fits the first segment of the LF model to the first part of the DIF signal The return phase of the DIF model is still needed to be matched to the second segment (exponential section) of the LF model The procedure for this is explained in the next section

4.4.3 Determination of t_a

The second part of the DIF signal ($t_e < t < T_0$) is matched to the second part of the LF model. The determination of the parameters (exponential curve, eqn 4 1b) requires the value of the fourth parameter t_a The parameter t_a is determined from the residual flow U_e, at t_e U_e can be estimated using an exponential curve and is given as [Fant et al.,1985]·

$$U_e = \frac{E_e t_a K_a}{2} \qquad 4\ 7$$

This residual flow U_e is found by the integration of eqn 4 1b The U_e given in eqn (4 7) is an approximation of the integrated flow with K_a taking different values (given below) for different values of R_a where $R_a = \frac{t_a}{(T_0 - t_e)}$ [Fant et al.,1985]

if $R_a < 0\ 1$

$$K_a = 2.0 \qquad 4\ 8a$$

if $0\ 1 < R_a < 0.5$

$$K_a = 2.0 - 2.34R_a^2 + 1.34R_a^4 \qquad 4.8b$$

with K_a taking values $1.5 < K_a < 2\ 0$

if $R_a > 0.5$

$$K_a = 2\ 16 - 1.32R_a + 0.64(R_a - 0.5)^2 \qquad 4\ 8c$$

with K_a taking values $K_a < 1.5$

In the estimation of t_a, U_e is equated to the flow estimated at t_e (U_e^M) using the LF model in the last section The steps involved in the estimation of t_a are described below·

a) Express t_a in terms of R_a as $t_a = R_a\ (T_0 - t_e)$; where the values of T_0 and t_e are estimated in the previous section.

b) Rewrite U_e in terms of R_a alone, using the values of K_a from eqn. 4.8. i.e;

$$U_e = \frac{E_e R_a (T_0 - t_e) K_a}{2} \qquad 4\,9$$

c) The parameter R_a is calculated for all three different values of K_a (eqn. 4.8) with the measured value of E_e, in the previous section and the flow U_e^M, using eqn. 4.9 Only one of the three estimated value of R_a satisfies (would satisfy) the three inequality conditions described in eqn. 4 8 This value of R_a is then used to obtain the required parameter, t_a. i.e; $t_a = R_a.(T_0 - t_e)$

d) For the implementation of the exponential model, we need the value of ϵ This value is estimated by an iterative process on the following eqn. 4.10 as suggested in [Fant *et al.*,1985].

$$\epsilon t_a = 1 - e^{-\epsilon(T_0 - t_e)} \qquad 4\,10$$

Once this parameter is estimated, LF model is fully fitted to the DIF signal using eqn. 4.1.

4.4.4 Conclusion of the new matching method

The method described above is employed in the parameterisation of the DIF signals, obtained from natural and synthetic speech signals (next section).

The matching method described in this work, is robust for DIF signal with many ripples. The LF model parameters are estimated by matching the areas under the LF model and the DIF signals. Both the DIF and IF signals are used in the matching of the parameters The use of the IF signal for determining the zero crossing point (peak glottal flow point at t_p) of the DIF signal is found to be more reliable than finding it from the DIF signal The method needs only two area values of the DIF signals (flow) at two points where they can be estimated precisely. The two points t_p and t_e are chosen as these points

4.5 INVERSE FILTERING AND PARAMETERISATION OF THE SIGNAL, APPLYING THE CURRENT METHOD

Natural and synthetic speech signals S(n), are inverse filtered using the method described in section 4.3. The DIF signal obtained using the inverse filter is then

parameterised (i.e LF model parameters are found) using the algorithm described in section 4.4.

4.5.1 Speech material

As for the estimation of the vocal tract area function parameters in chapter 3, vowels extracted from natural speech signals (section 3.6.2.1) were used to test the inverse filtering method discussed here. However, unlike in section 3.6.2.1, synthetic speech signals here are synthesised using the LF model signals instead of the Rosenberg flow signals. Since the LF model is used in the parameterisation of DIF signals, synthetic vowels are also produced using the LF model flow for comparison purposes DIF signals obtained from synthetic pressure waveforms generated using the Rosenberg flow were also parameterised Good matches were found.

4.5.2 Test with synthetic vowels

Six synthetic Russian vowels */a/, /e/, /i/, /o/, /u/* and */ɨ/* were generated using the SPM described in chapter 2 (fig 2 2a). The lip flow obtained at the end of the model i e, at the lips, is converted to a pressure waveform by passing the signal through the filter Z_L Z_L, the radiation characteristic at the lips is given by Laine's model (eqn 2 15) [Laine, 1982] as described in section 2.3.4. The following data were used in the generation of above synthetic vowels

a) The reflection coefficients of the SPM were found using, the area function data for those vowels (same shapes as those used for the generation of the codebook in chapter 3).

b) The input signal (glottal flow) to the SPM was synthesized by the LF model. The LF model, as mentioned, generates the differentiated form of the glottal flow. Therefore, the model signal was integrated to obtain the glottal flow. The following parameter values were used in synthesis with the LF model fundamental period T_0 = 10 ms, t_p = 4 ms; t_e = 6 ms, t_a = .12 ms, E_e = 1 and the sampling frequency is set to 16 kHz

The synthetic pressure waveforms of the vowels are shown in fig 4 8a - 4 13a.

4.5.2.1 Inverse filtering of synthetic speech

It was shown in section 3 6.3 that the weighted cepstral measure gives a better area function match than that of the unweighted cepstral measure. It was also mentioned that the improvements were particularly noticeable for vowels */e/* and */o/* in both area function and spectral domain.

Fig 4 7 shows the DIF signals obtained for */e/* and */o/* from the area functions estimated using both the weighted and the unweighted cepstral measure. It can be clearly seen, that the size and the number of ripples are very much reduced in the closed phase of the DIF, for both vowels using the weighted measure This reinforces the fact that use of a weighted cepstral measure improves the estimation of SPM parameters. Therefore, as mentioned, the weighted cepstral measure is chosen to estimate the inverse filter coefficients

DIF and IF signals for the synthetic speech waveform are obtained using the inverse filtering method described in section 4 3 1 Figs 4.8b,c - 4 13b,c show DIF and IF signals of the vowels */a/, /e/, /i/, /o/, /u/* and */ɪ/*

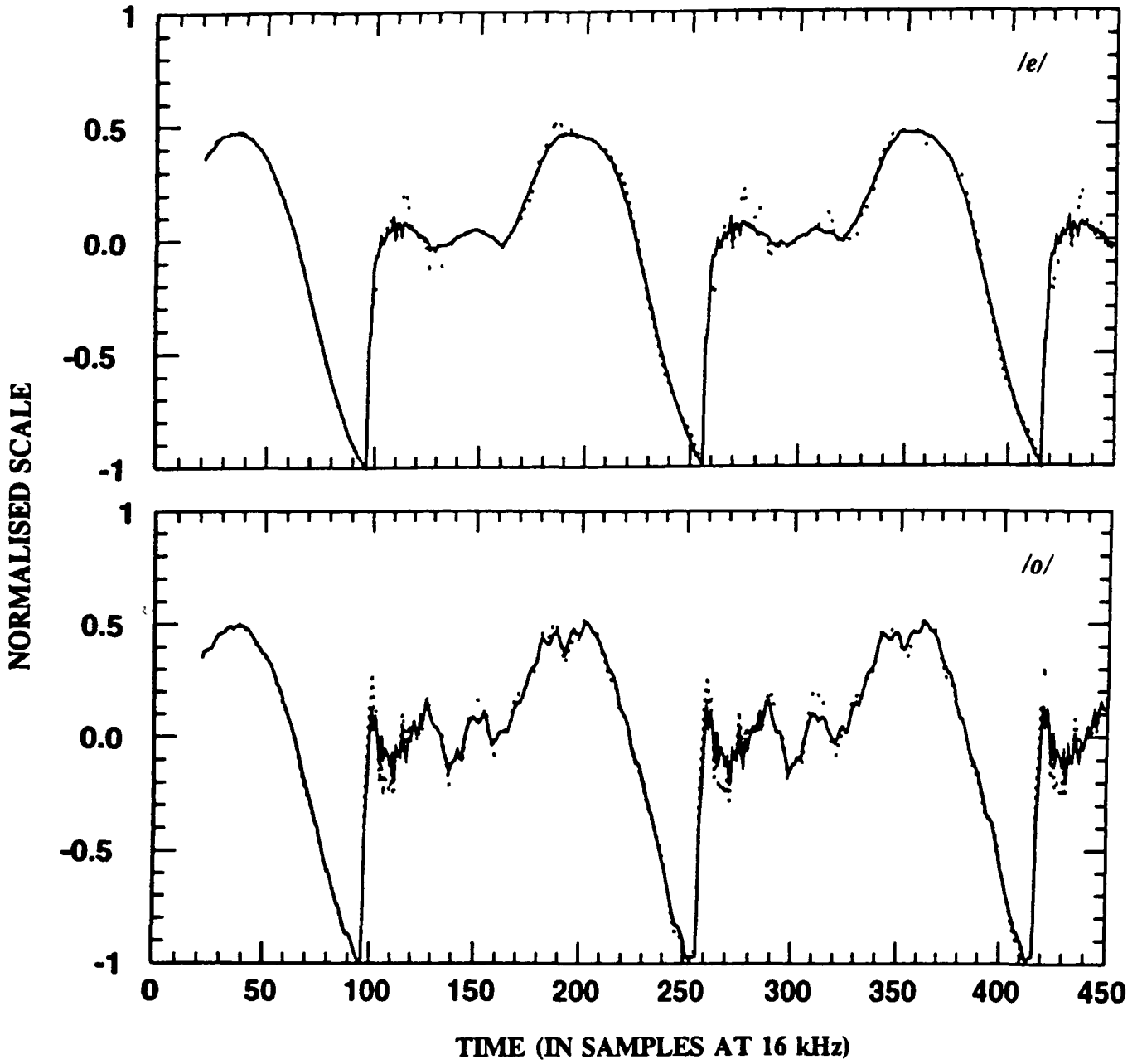


Fig. 4.7 The DIF waveforms of the inverse filter using the cepstral measure (dotted) and the weighted cepstral measure (solid).

a): for the vowel *e/*

b): for the vowel *o/*

4.5.2.2 Matching of LF model to DIF signal

DIF signals of the six vowels are matched to the LF model (figs. 4.8 - 4 13) using the algorithm described in section 4.4 The time based parameters extracted from the estimated source signals are shown in Table 4.1 Figs. 4 8 - 4.13 and Table 4 1 show good matches for t_e, t_p and T_o. However, large errors are found in the estimate of t_a

Description	T_0 (ms)	t_e (ms)	t_p (ms)	t_a (ms)	t_x (ms)
Orig. LF model	10 0	6.0	4.0	0 12	0.313
/a/	10.0	6 074	4.137	0.185	0.863
/e/	10.0	6.006	3 943	0 027	0.017
/i/	9.937	6.278	4.153	0.000	0.000
/o/	9 875	6 077	4.202	0 216	0 6175
/u/	9 823	5.967	4 280	0 362	1 0955
/i/	10 0	6 250	4 063	0 000	0 000

Table 4.1 Estimated time base parameters from the DIF and IF signals, the first row shows the parameters of the original source signal

To relate the original source parameters to the estimated (parameterised) data of the DIF signal, the following factors, Opening Quotient (OQ), Closing Quotient (CQ) and Speed Quotient (SQ) are calculated The above factors defined in terms of time based parameters of the glottal flow are also useful in characterising the different voice types depending on their values e g pressed, breathy, normal, falsetto, vocal fry etc [Alku, 1992, Childers and Lee, 1991]. These terms are defined as follows with reference to the time parameters of fig 4.4.

$$OQ = \frac{open\ phase}{pitch\ period} = \frac{t_e + t_x}{T}$$

$$CQ = \frac{closing\ phase}{pitch\ period} = \frac{t_e + t_x - t_p}{T}$$

$$SQ = \frac{opening\ phase}{closing\ phase} = \frac{t_p}{t_e + t_x - t_p}$$

The errors in the above OQ, CQ and SQ (EOQ, ECQ and ESQ) are defined as $|P_0 - P_e| / P_0$, where P_0 and P_e correspond to the original and the estimate of the parameters Two other parameters oQ and eQ were introduced because it was obvious from the above table that the estimates of t_a are inaccurate and hence this alone would produce large errors in SQ (ESQ) and CQ (ECQ) Therefore ESQ and ECQ would not be a good measure to indicate the good estimates obtained for other parameters t_p, t_e and E_e The parameters oQ and eQ are defined as:

$$oQ = \frac{opening\ phase}{pitch\ period} = \frac{t_p}{T_0}$$

$$eQ = \frac{instant\ of\ excitation}{pitch\ period} = \frac{t_e}{T_0}$$

Table 4 2 shows the percentage of errors observed in T_0, oQ, eQ, CQ, OQ and SQ Low values of EoQ and EeQ show that t_p and t_e are matched accurately with the source signal.

The superimposed model (matched) waveforms and the DIF and IF signals of three glottal cycles, for vowels */a/*, */e/* and */i/* are shown in figs 4.8 - 4.13 The waveforms of vowels */i/*, */o/* and */u/* are slightly out of phase due to a small error in the estimation of the pitch period T_0 However, the error in T_0 would not be amplified or propagated in the actual analysis of speech signals This is because, the parameters of the SPM should be estimated for every 20 - 30 ms period and hence the glottal flow signals (i e, value of T_0) should also be calculated and updated for each analysis frame The matched waveforms (parameterised waveforms) of all the vowels coincide well with the original LF model waveform used However, this is not shown in figs 4.8 - 4 13

VOWELS	*T*	*EoQ*	*EeQ*	*ECQ*	*EOQ*	*ESQ(SQ)*
/a/	0.00	3.300	1 15	21.05	9 89	14.4 (1 48)
/e/	0.00	1.425	0 10	10.07	4 59	9 2 (1 89)
/ɪ/	0 73	0 325	2.300	8 12	0.60	17 3 (2.03)
/o/	1.25	5.050	1.16	7.43	5.93	2.3 (1 69)
/u/	1 77	5 000	0 55	20.28	11.8	11 0 (1.54)
/ɪ/	0 00	1.575	4 17	5 45	1.0	7.5 (1 86)

Table 4.2 Relative errors in the parameters T_0, t_p, t_e and t_a, extracted from the estimated LF source signals. Note that SQ value of the original source signal is 1.73. SQ values for the vowels are shown in the brackets.

4.5.3 Test with natural vowels

As for the synthetic vowels, five natural vowels */a/*, */e/*, */ɪ/*, */o/* and */u/* used in section 3.6.2 1 are filtered using the inverse filter model. Fig 4 14 - 4 18 show the speech, DIF and IF signals and the superimposed model (matched) waveforms of three glottal cycles, for the above vowels The time based parameters extracted from the above DIF and IF signals are given in Table 4 3

Description	T_0 (ms)	t_e (ms)	t_p (ms)	t_a (ms)
/a/	10 125	9.619	6 806	0.371
/e/	9 250	7.641	5 049	0.027
/ɪ/	9.500	6.468	4 530	0.410
/o/	8.375	6.864	5.239	0.360
/u/	6.812	5 821	4 446	0 490

Table 4.3 Estimated time base parameters from the DIF and IF signals

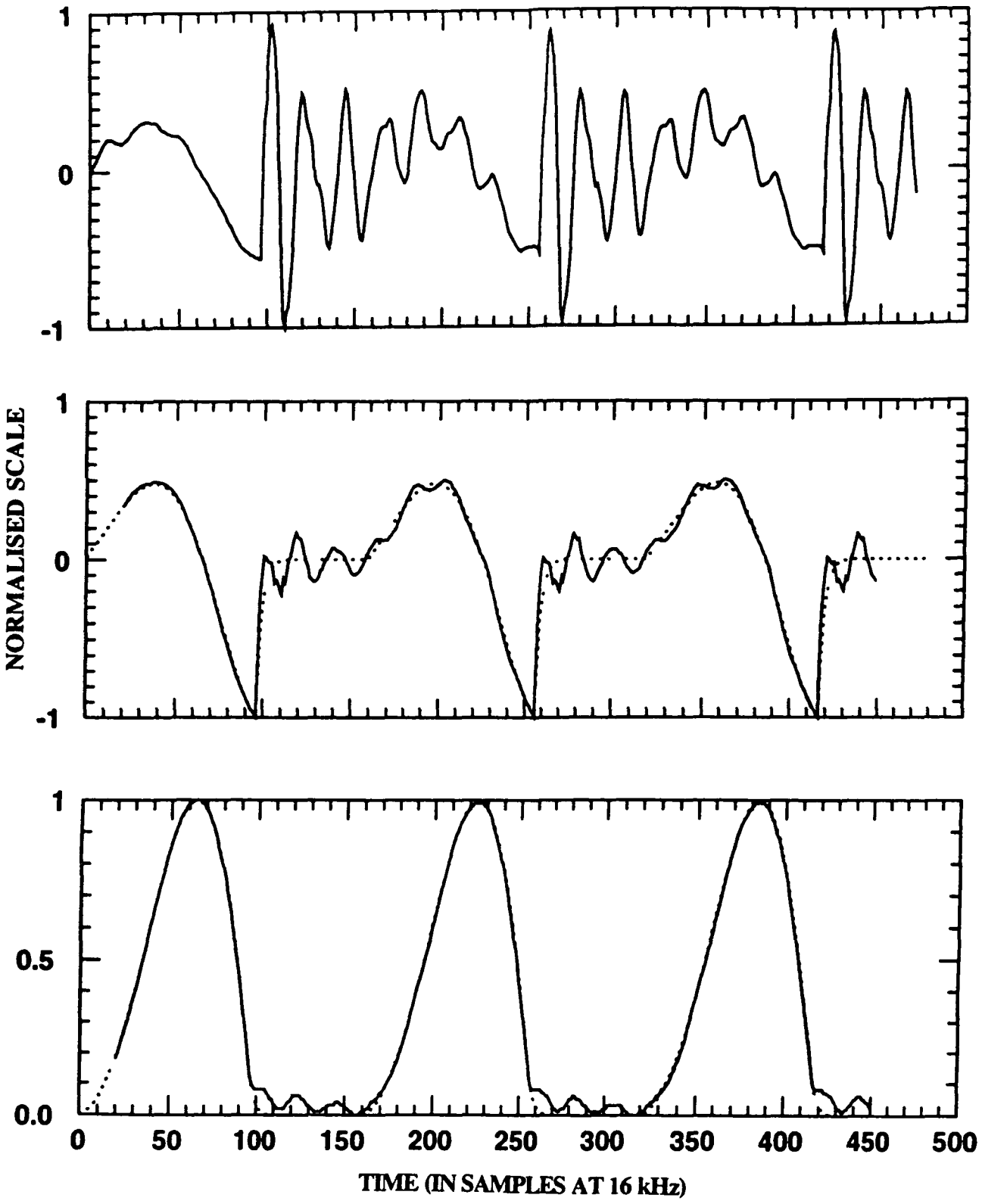


Fig. 4.8 The synchronised matches of the synthetic vowel */a/

a): Speech waveform b): DIF waveform and the superimposed LF model match

c): IF waveform (glottal flow) and the superimposed LF model match

_____: Original waveform .....: LF model matched waveform

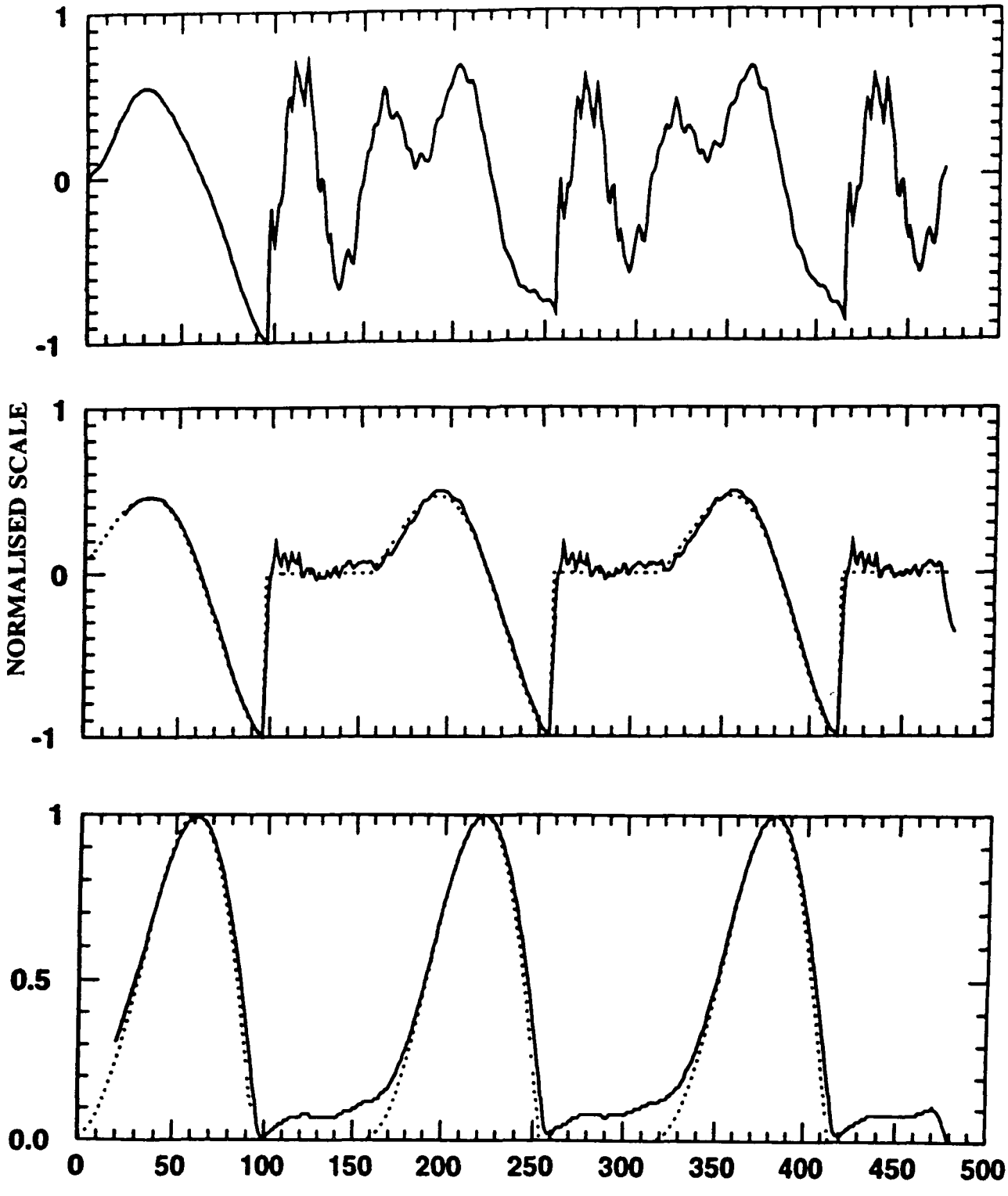


Fig. 4.9 The synchronised matches of the synthetic vowel */e/*

a): Speech waveform b): DIF waveform and the superimposed LF model match

c) IF waveform (glottal flow) and the superimposed LF model match

_____: Original waveform .....: LF model matched waveform

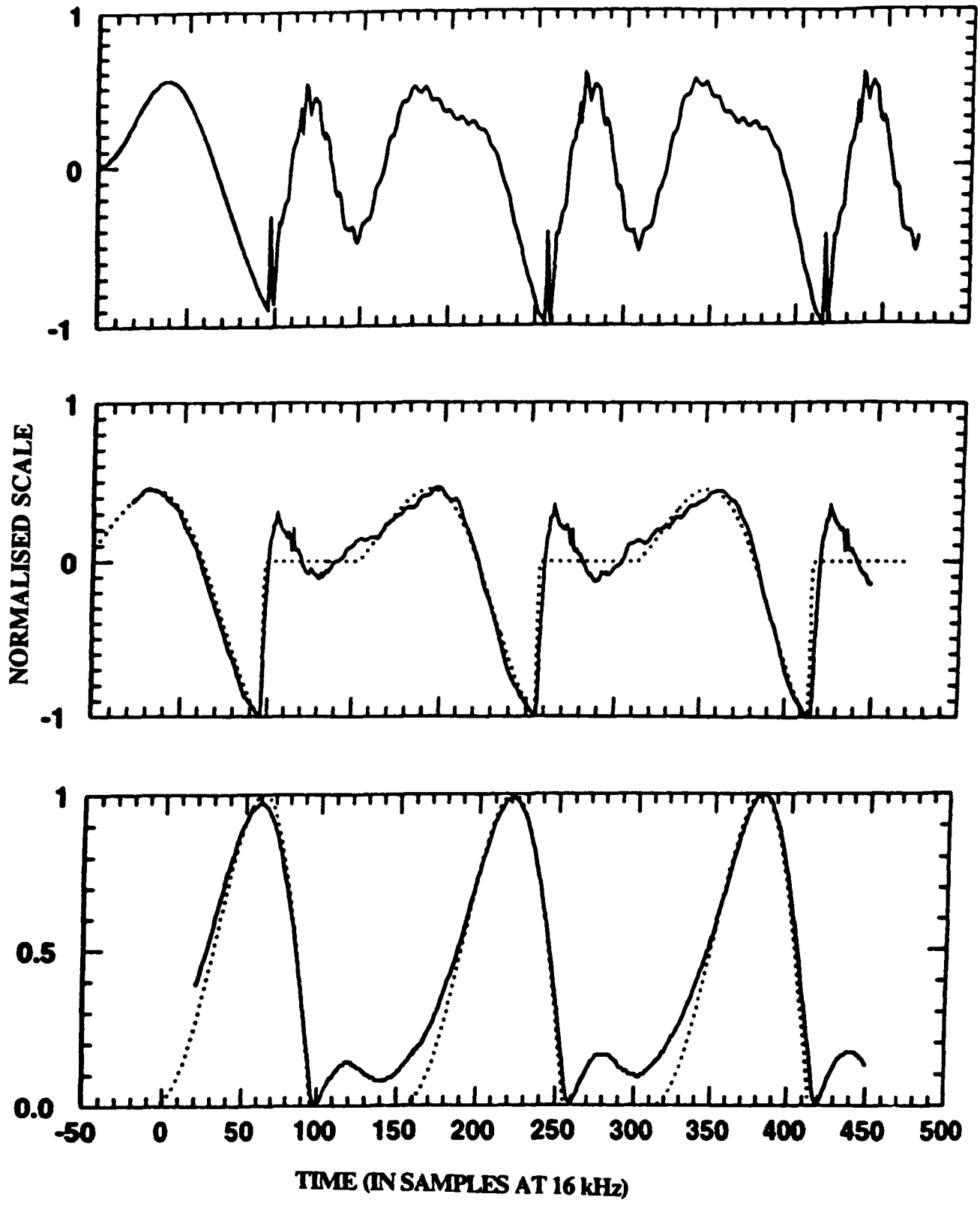


Fig. 4.10 The synchronised matches of the synthetic vowel */i/*

a): Speech waveform b): DIF waveform and the superimposed LF model match

c) IF waveform (glottal flow) and the superimposed LF model match

_____: Original waveform .....: LF model matched waveform

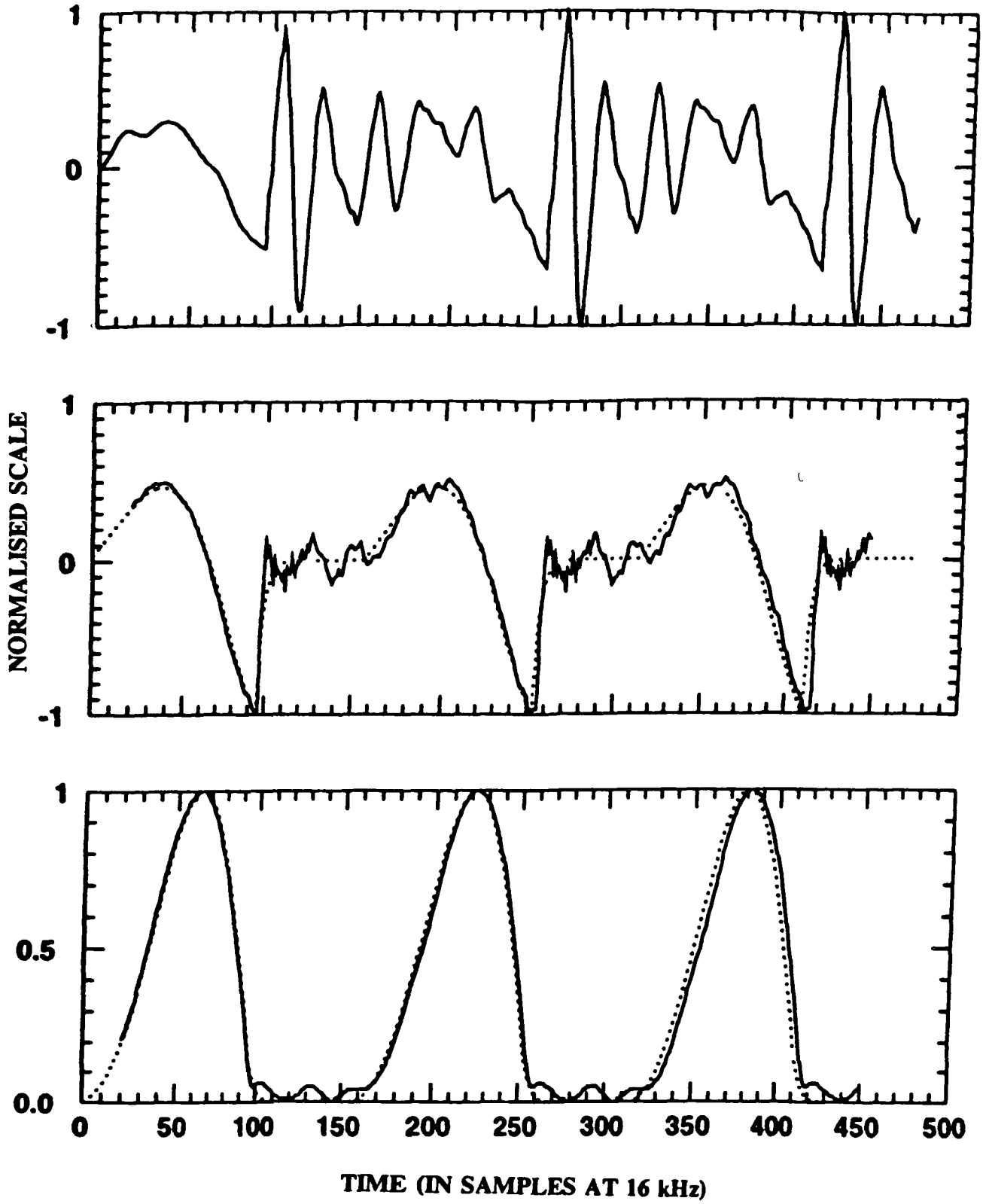


Fig. 4.11 The synchronised matches of the synthetic vowel */0/*

a): Speech waveform b): DIF waveform and the superimposed LF model match

c) IF waveform (glottal flow) and the superimposed LF model match

_____: Original waveform .....: LF model matched waveform

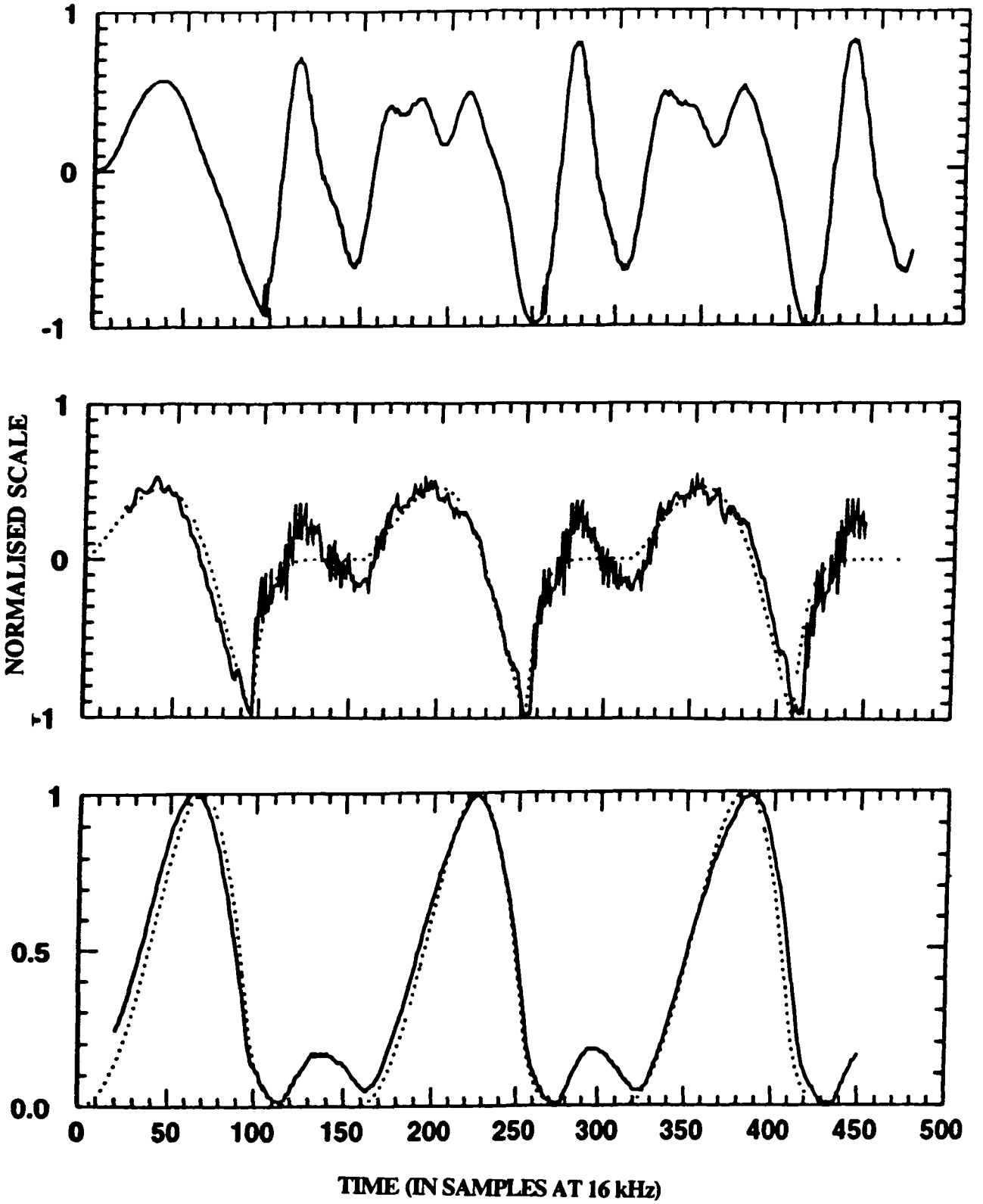


Fig. 4.12 The synchronised matches of the synthetic vowel */u/*

a): Speech waveform b): DIF waveform and the superimposed LF model match

c) IF waveform (glottal flow) and the superimposed LF model match

_____: Original waveform .....: LF model matched waveform

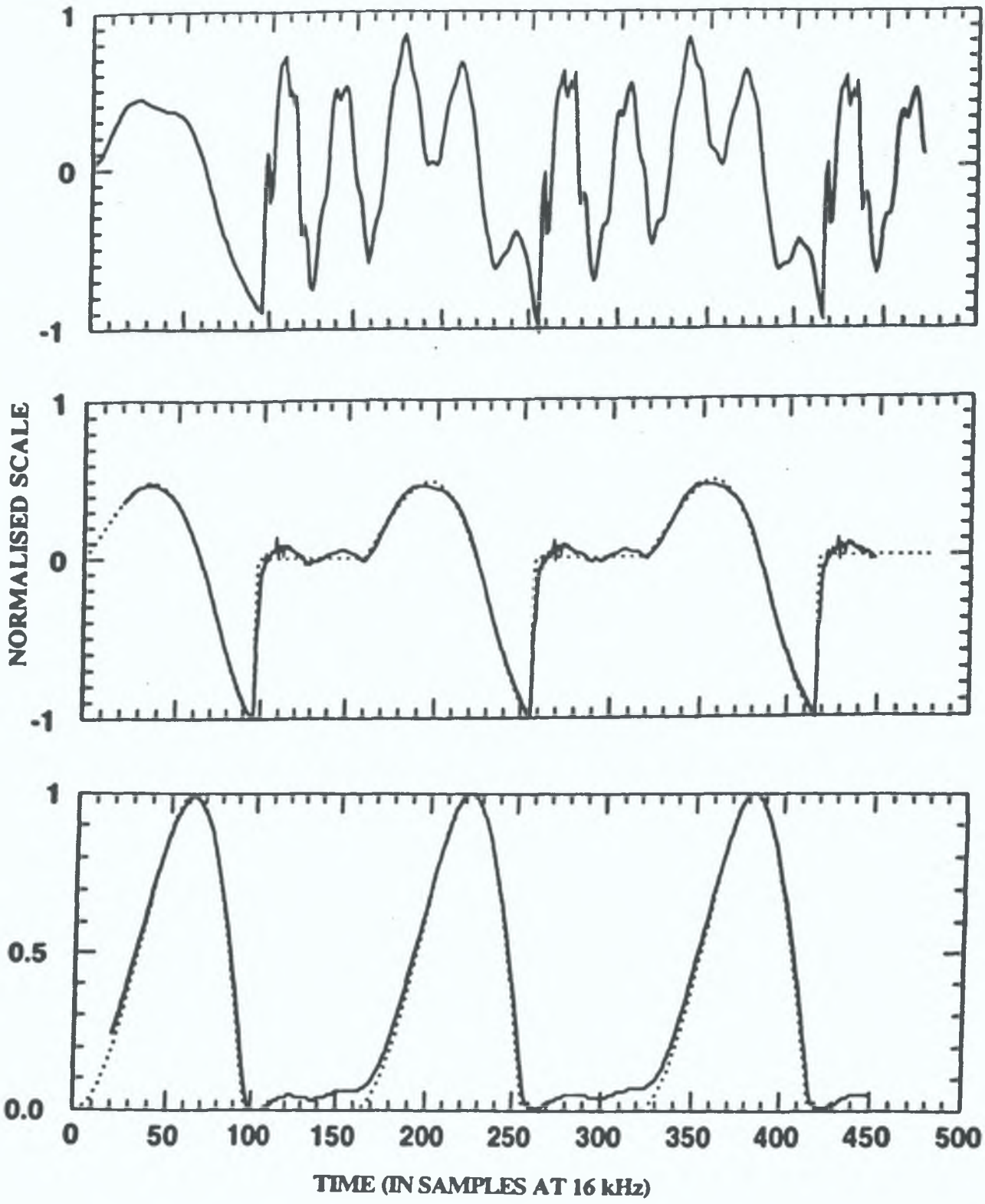


Fig. 4.13 The synchronised matches of the synthetic vowel /y/

a): Speech waveform b): DIF waveform and the superimposed LF model match

c) IF waveform (glottal flow) and the superimposed LF model match

_____: Original waveform .....: LF model matched waveform

Unlike in the case of synthetic speech waveforms, we do not have the true shapes of the glottal flow for comparing the estimated glottal flow signals However, the IF signals obtained are plausible looking and the values of the estimated glottal flow parameters t_e, t_p, t_a, E_e and T_0 seem reasonable for the real speech waveforms.

The correctness of the glottal flow signals is measured indirectly by comparing the natural speech spectra (LPC spectra) with the synthetic speech spectra estimated from the area functions (chapter 3) and the parameterised glottal flow signal For this purpose·

a) The LPC magnitude response of the synthetic speech is found using 11 LPC coefficients.

b) The LPC magnitude response of the natural speech is also found using the same number of LPC coefficients.

The magnitude response matches of a) and b) for the vowels are shown in fig 4 19 The close matches indicate that the estimated parameters are reasonable

The area functions of the speech signals were estimated using the autocorrelation method of LPC analysis. A frame size of 250 points is used at the sampling frequency of 16 kHz for this purpose For the signals analysed, this frame size would include both the open and closed cycles of the glottal flow signal Therefore we, feel that the ripples in the DIF waveforms are due to the source - tract interaction specially during the open phase of the glottal flow signal.

4.6 CONCLUSION

7The use of area function shapes is not common in the inverse filtering of speech In the current work, area function shapes estimated from the speech signals are successfully used for the inverse filtering of the signals Both IF and DIF signals were used in the parameterisation of the LF model This makes the parameterisation more reliable

As shown, estimated glottal flow waveforms and area functions match well with the original shapes of both glottal flow and area function The parameters obtained for the natural speech signal are acceptable and produce reasonable wave shapes The speech spectra obtained from the estimated parameters are comparable with the natural speech spectra

Extra care should be taken in the recording of speech signals to obtain linear phase signals; the use of non linear phase signals distorts the glottal flow signals.

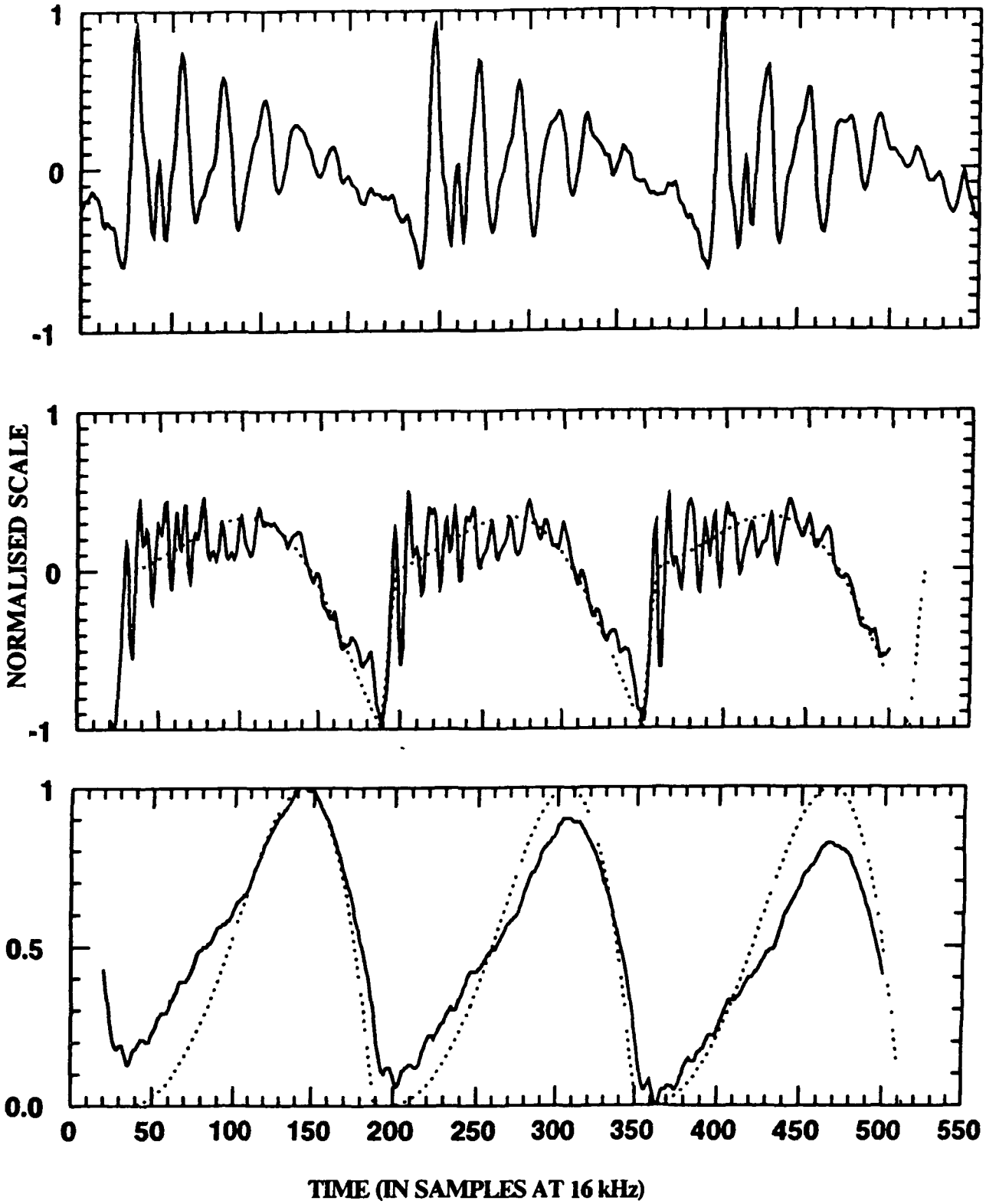


Fig. 4.14 The synchronised matches of the natural vowel /a/

a): Speech waveform b): DIF waveform and the superimposed LF model match

c): IF waveform (glottal flow) and the superimposed LF model match

_____: Original waveform .....: LF model matched waveform

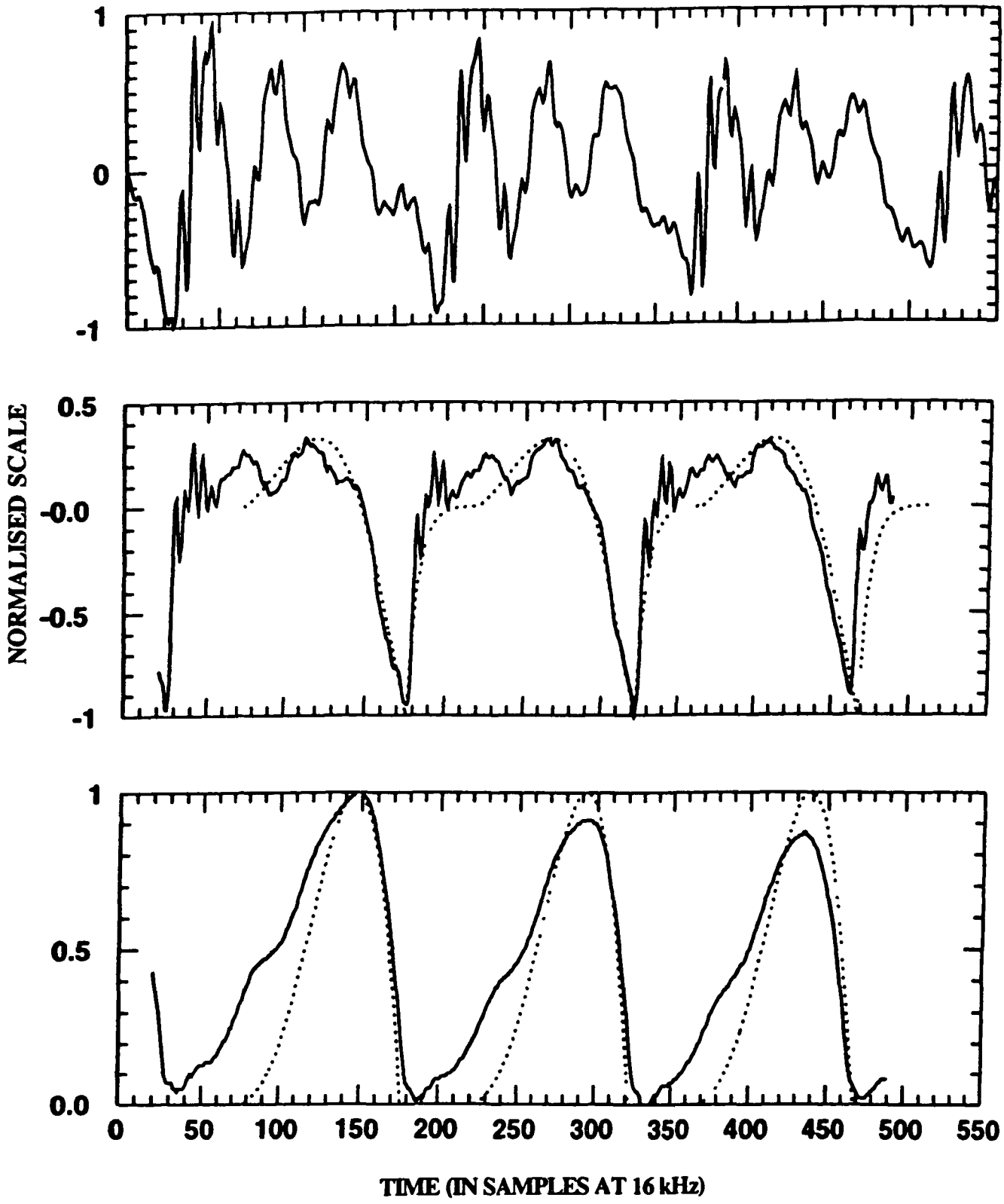


Fig. 4.15 The synchronised matches of the natural vowel */e/*

a): Speech waveform b): DIF waveform and the superimposed LF model match

c) IF waveform (glottal flow) and the superimposed LF model match

_____: Original waveform .....: LF model matched waveform

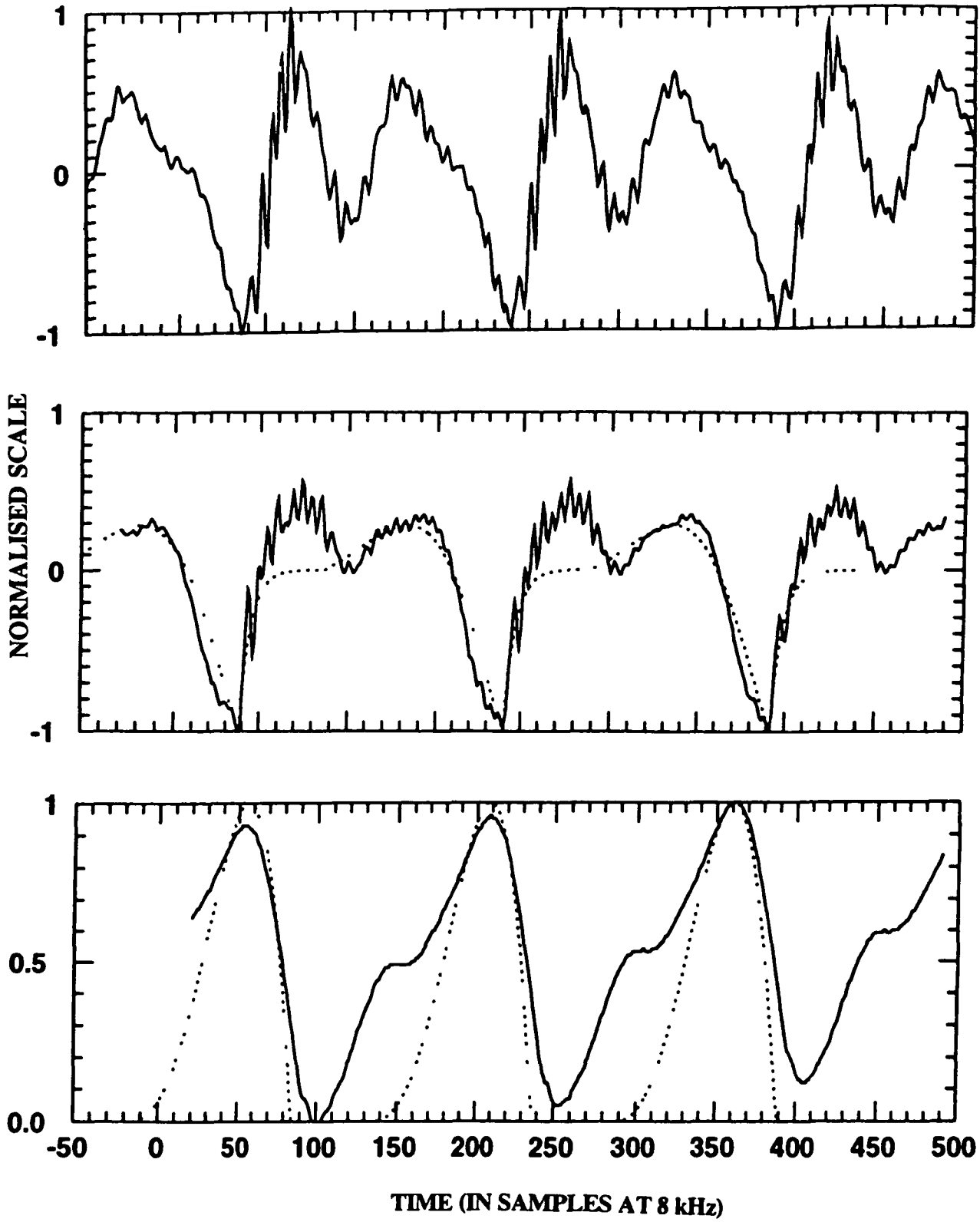


Fig. 4.16 The synchronised matches of the natural vowel /i/

a): Speech waveform b): DIF waveform and the superimposed LF model match

c) IF waveform (glottal flow) and the superimposed LF model match

_____: Original waveform .....: LF model matched waveform

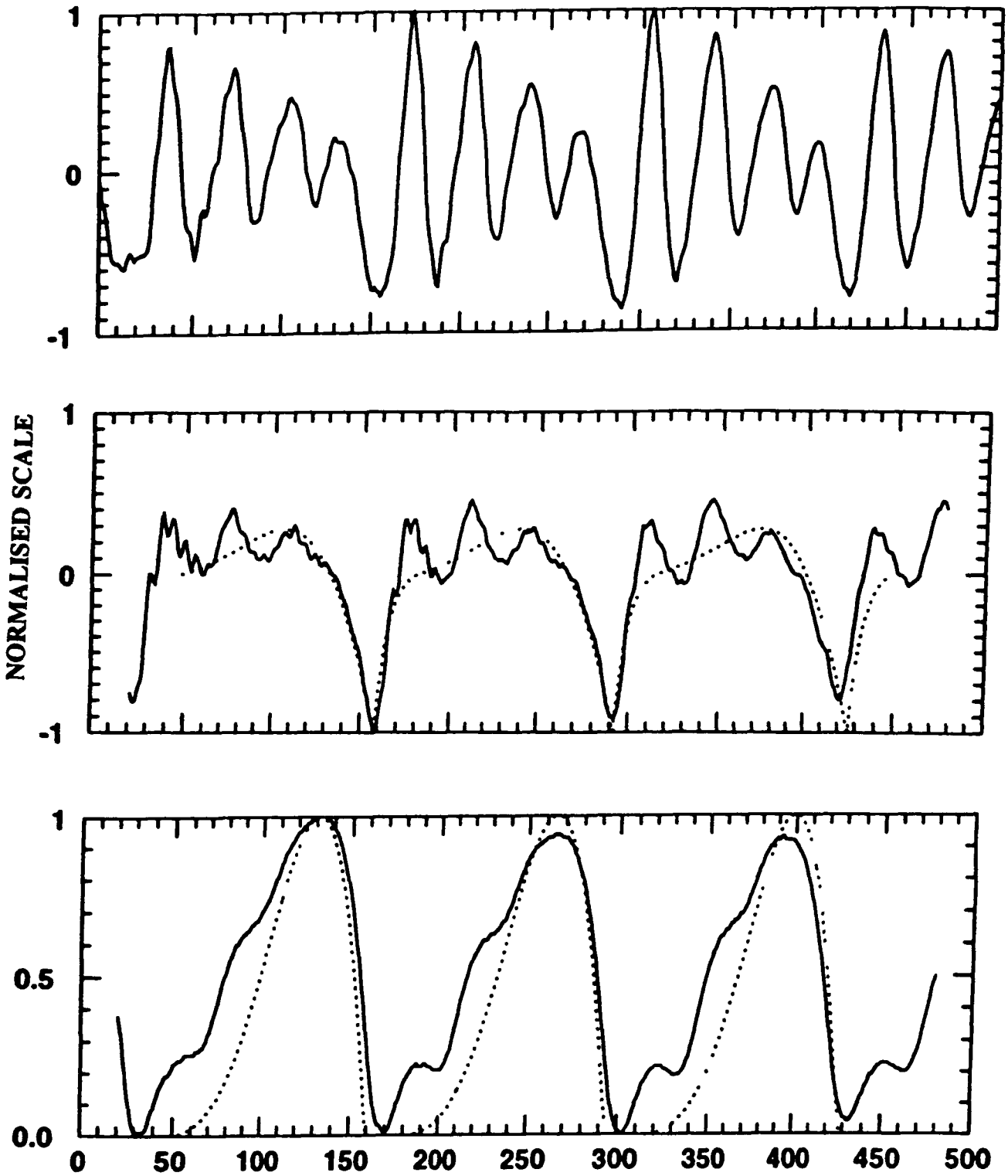


Fig. 4.17 The synchronised matches of the natural vowel */0/*

a): Speech waveform b): DIF waveform and the superimposed LF model match

c) IF waveform (glottal flow) and the superimposed LF model match

_____: Original waveform .....: LF model matched waveform

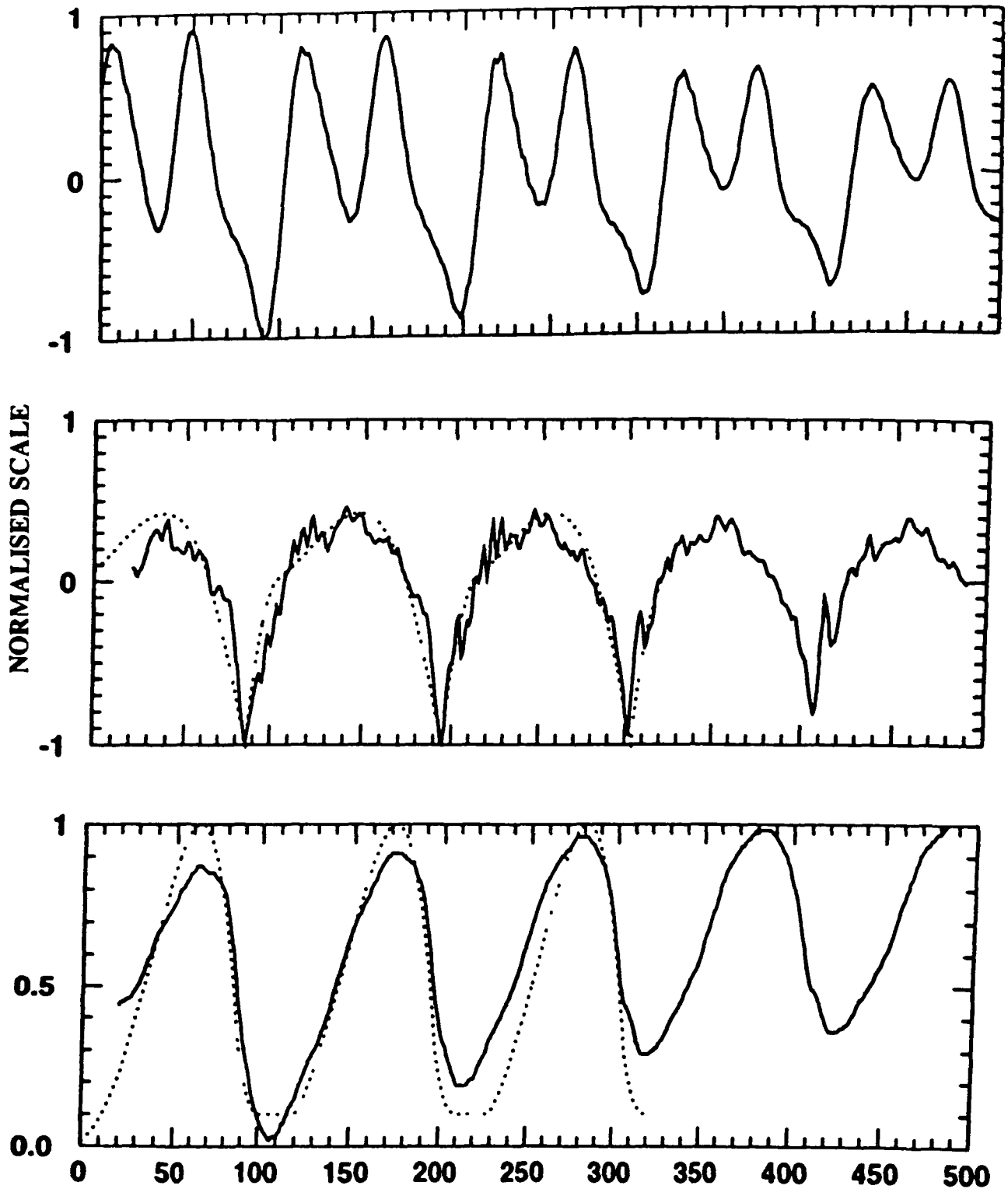


Fig. 4.18 The synchronised matches of the natural vowel /u/

a): Speech waveform b): DIF waveform and the superimposed LF model match

c) IF waveform (glottal flow) and the superimposed LF model match

_____: Original waveform .....: LF model matched waveform

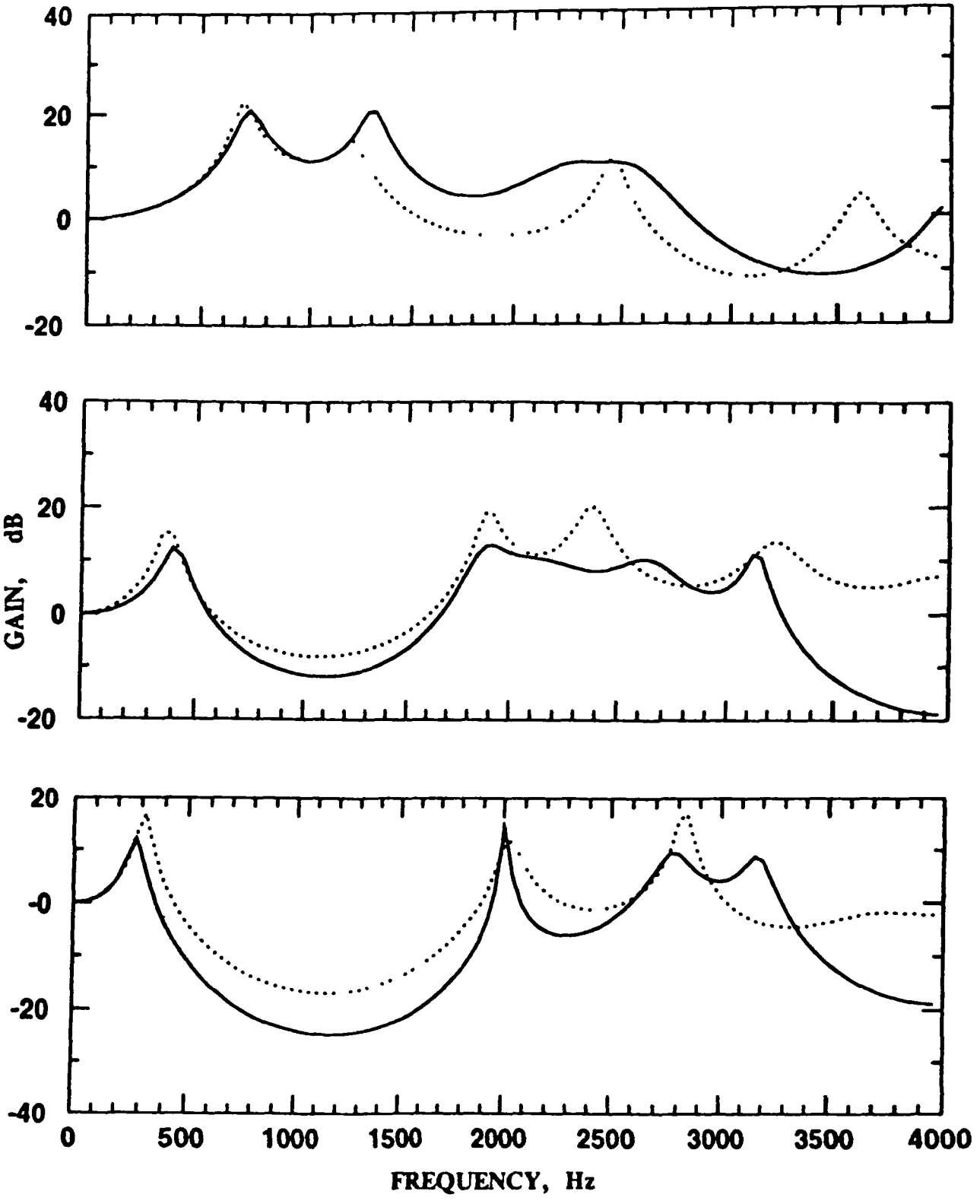


Fig. 4.19 LPC spectra of the natural vowels and the LPC spectra of the vowels synthesised from the estimated area function and the parametrised LF glottal flow.
a) Vowel *la/* b): Vowel *le/* c): Vowel *li/*
_____: Natural vowel .....: Synthetic vowel

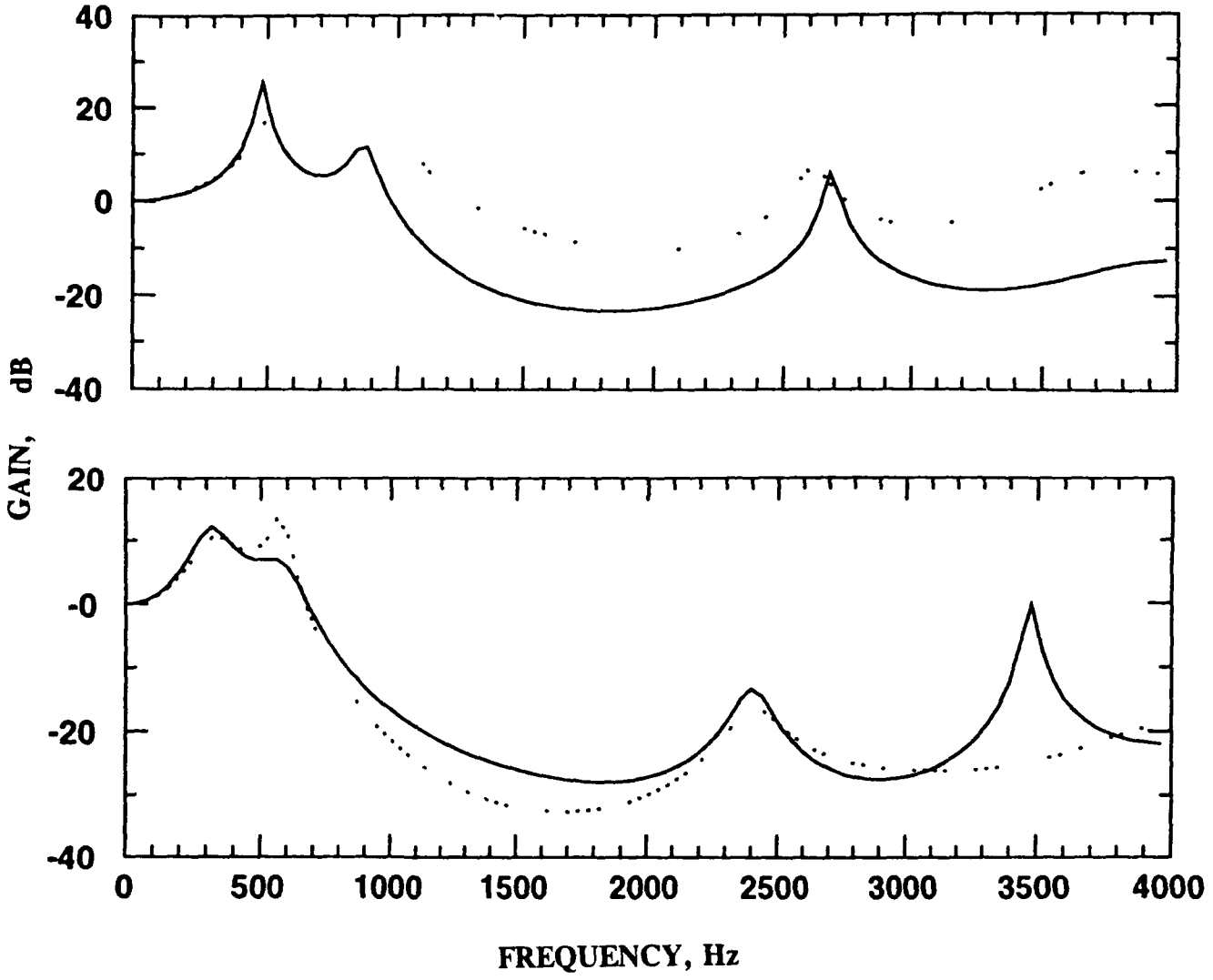


Fig. 4.20 LPC spectra of the natural vowels and the LPC spectra of the vowels synthesised from the estimated area function and the parametrised LF glottal flow.
d) Vowel *//o//* e): Vowel *//u//*
_____: Natural vowel .....: Synthetic vowel

CHAPTER 5

CONCLUSIONS

5.0 CONCLUSIONS

Measurements on human speakers can be used to improve modelling of the system The aim of this work has been to acquire vocal tract configurations and glottal flow signals from speech pressure waveforms and to parameterise the data In addition, a number of relevant models for the vocal tract and glottal flow signal were discussed Important energy loss mechanisms of speech production were addressed and included in the model. The methodology described here produces only humanly attainable shapes for the vocal tract. The data estimated from several synthetic voiced speech waveforms and vowels extracted from natural speech sequences, produced good matches with the original data in both the area function and spectral domain

Purely statistical encoding and decoding methods can produce good approximations to human speech, but offer little insight into the underlying physiological and neural phenomena The strategy adopted to obtain realistic articulatory data was based on a constrained SPM which includes some vital information about the vocal apparatus. The vocal tract shapes were estimated in the acoustic domain, using a cepstral distance measure. The estimated area functions were used in an inverse filter to obtain time domain glottal flow waveforms from voiced speech

The SPM developed and employed (chapter 2) is an enhanced version of a previous model [Scaife, 1989] The speech production system is described by 9 parameters, 5 vocal tract (or SPM) and 4 glottal flow parameters The 5 vocal tract parameters have a close relationship with the natural shapes of the vocal tract. The glottal flow is related to the actual movements of the vocal cords The 4 glottal flow parameters used in the system are also related to the terms (e.g opening phase, closing phase, open phase etc) that used to describe the movement of the vocal folds Energy losses due to vocal tract wall vibration and glottal resistance have significant effects on low formant frequencies of the speech signals Radiation losses affect the high formant frequencies These effects on the system were simulated by parametric models and placed at the glottis and lip end of the SPM [Scaife, 1989] The SPM is similar to the Kelly - Lochbaum model [Kelly and Lochbaum, 1962] i e. the vocal tract is represented by a chain of cylindrical tubes of equal length. The inclusion of the losses captures some features of the real vocal tract and hence improves the modelling.

The vocal tract wall vibrations occur along the vocal tract have their greatest

effect in a region about 4 cm from the glottis [Badin and Fant , 1984] These effects were included at the glottis end of the SPM, by a parallel combination of a single wall vibration model and a glottal model. Placing the wall vibration model at the glottis simplified the modelling without much adverse effect. The effects of placing the wall vibration model at different points along the vocal tract was studied in relation to the formant frequencies (section 2 3.3). A distributive and discrete lossy vocal tract model (see appendix B4) was used to place the wall vibration model at different points along the vocal tract. The wall vibrations, in general, lower the low formant frequencies of the system It was demonstrated in section 2.3 3 that the composite termination of the glottal model produces results which were very similar to those produced by placing the wall model at 4 cm above the glottis.

The SPM in [Scaife, 1989], coupled with a five parameter model of the vocal tract area function [Lin, 1990] was employed to gather data on vocal tract shapes (see chapter 3) and glottal flow signals (see chapter 4) The structure of the synthesis filter and the inverse filter model is shown in fig. 2.2

Initially the five parameters of the vocal tract were estimated by matching the model spectra to the original spectra. A small linked codebook and an optimisation procedure were used for this purpose [Scaife and Kanagaratnam, 1991].

The need for optimisation is eliminated by a use of a larger codebook comprehensively spanning the acoustic space The number of parameters required to describe the spectral shape of the acoustic codebook was reduced by representing each spectrum by cepstral coefficients An articulatory-to-acoustic "linked codebook" of approximately 1600 shapes was generated based on the above SPM and exhaustively searched to estimate the vocal tract parameters (chapter 3). The codebook contains only humanly attainable vocal tract shapes This was achieved by non uniform interpolation between 15 vocal tract area functions estimated from mid-sagittal X-rays of fricatives and vowels. Therefore, although possibly erroneous, the vocal tract shapes estimated using the codebook are generally possible for a human speaker.

The acoustic section of the codebook contains a set of cepstral coefficients for each vocal tract shape The F_1 - F_2 span of the codebook adequately spans a large acoustic space (fig 3.11). This indicates that this codebook can be used in the estimation of voiced speech signals The existence of a close correlation between the

area function parameters (X_c, A_0) and the formants (F_1 ,F_2) was observed (see Table 3.4)

Vocal tract area function data was estimated from several different synthetic speech waveforms and Russian vowels [Fant, 1960] */a/, /e/, /i/, /o/* and */u/* extracted from natural speech The vocal tract area functions estimated from these signals match well with the original root shapes in both the area function and spectral domains The advantages of the use of cepstral coefficients are discussed. It is demonstrated that a weighted cepstral Euclidean measure performs substantially better than an unweighted measure in the estimation of shapes.

A parametrised *area function*, estimated from the speech signal, is used in the construction of an inverse filter model. The use of the area function in the inverse filtering of signals is rare in previous techniques A four parameter model of differentiated glottal flow [Fant *et al* , 1985] is used to parameterise the DIF (Differentiated Inverse Filtered) signals obtained in the inverse filtering process The four parameters are estimated from the DIF and IF (Inverse Filtered) signals using an optimisation algorithm. The optimisation procedure involves matching the area under the DIF signals and the areas under the LF model signals Several methods have been proposed to parameterise the DIF signal using the LF model [Fant *et al* ,1985, Childers and Lee, 1991]. These methods work well for clean DIF signals The algorithm described is robust for DIF signals with several formant ripples in the closed phase of the glottal flow signal The method was tested on synthetic speech generated from the LF glottal flow model waveforms and the X - ray area function data of some voiced sounds. The parameters estimated for the LF model from the DIF signals match well with the original LF model parameters of the glottal flow signal.

Tests on natural vowels indicated that the waveform parameters obtained for the glottal flow waveforms were at least plausible The parameters obtained here can be applied to classify different voice types in natural speech

An overview of the system developed for the estimation of area functions and glottal flow signals is given in fig. 5 2 All of the models and procedures described above were implemented and tested using programs written in "C".

5.1 Directions for further research

This section addresses the following research areas: Reduction of codebook size, expansion of the codebook to cover both voiced and non-voiced sounds, the problem of non-uniqueness in the estimation of area functions and an iterative procedure to obtain a better estimate for area functions.

5.1.1 Optimising the codebook size

The codebook used in the work contains 1635 shapes. The formant frequency span of the codebook showed (fig. 3.11) that certain regions in the codebook comprise a large number of vocal tract shapes. This suggests that the codebook contains a number of redundant vocal tract shapes i.e very similar vocal tract shapes Obviously, eliminating (pruning) [Schroeter, Meyer and Parthasarathy, 1990] the redundant shapes would substantially reduce the codebook size without affecting the quality of the codebook

In the pruning process a vocal tract shape can be discarded if the geometrical distance between the two shapes V_1 and V_2 is smaller than a pre-defined reasonable threshold value. The geometrical distance value should be defined in terms of the articulatory parameters with a relevant weight placed on each parameter Defining a suitable geometrical distance is not a trivial task as the resultant formant frequencies are much more sensitive to changes in A_0 or X_c than in A_c or L_0. In addition, a sensitivity to a particular parameter may be greater in different regions depending on the values of the parameter. For instance the influence of A_0 is on the spectrum is greater for values between 0 cm^2 and 1.5 cm^2 than above 1 5 cm^2 Therefore in the pruning criterion the parameters should be weighted according to their sensitivity An example of a geometrical distance that could be used in our codebook is given below (eqn 5 1) However, further research is required to choose proper values for the constants in eqn 5.1 or a better distance measure is needed in the articulatory domain

$$d(V_1,V_2) = (b.a_1)A_0^2 + (da_2)X_c^2 + a_3 A_c^2 + a_4 dL_0^2 \quad for\ A_0 < 1\ 5 \qquad 5\ 1a$$

$$d(V_1,V_2) = (c\ a_1)A_0^2 + (ea_2)X_c^2 + a_3.A_c^2 + a_4.dL_0^2 \quad for\ A_0 > 1\ 5 \qquad 5\ 1b$$

where V_1, V_2 are any two vectors to be compared from the codebook. a, b, c, e, a_1, a_2,

a_3, a_4, are constants which emphasise the articulatory parameters according to their weights on the formant frequencies.

5.1.2 Clustering the codebook (Vector Quantisation of the codebook)

Some attempts were made to cluster a codebook of 960 shapes (cdbk1), using a k - means [BMDP Statistical Software package, 1983] clustering algorithm. The 960 shapes were initially clustered into 16 clusters in the acoustic domain i.e. cepstral coefficients in the acoustic section of the codebook were clustered. The mean of the cepstral coefficients were then stored at the nodes of the tree (fig. 5.1). The shapes in the linked codebook in each of the 16 clusters were also stored under each node.

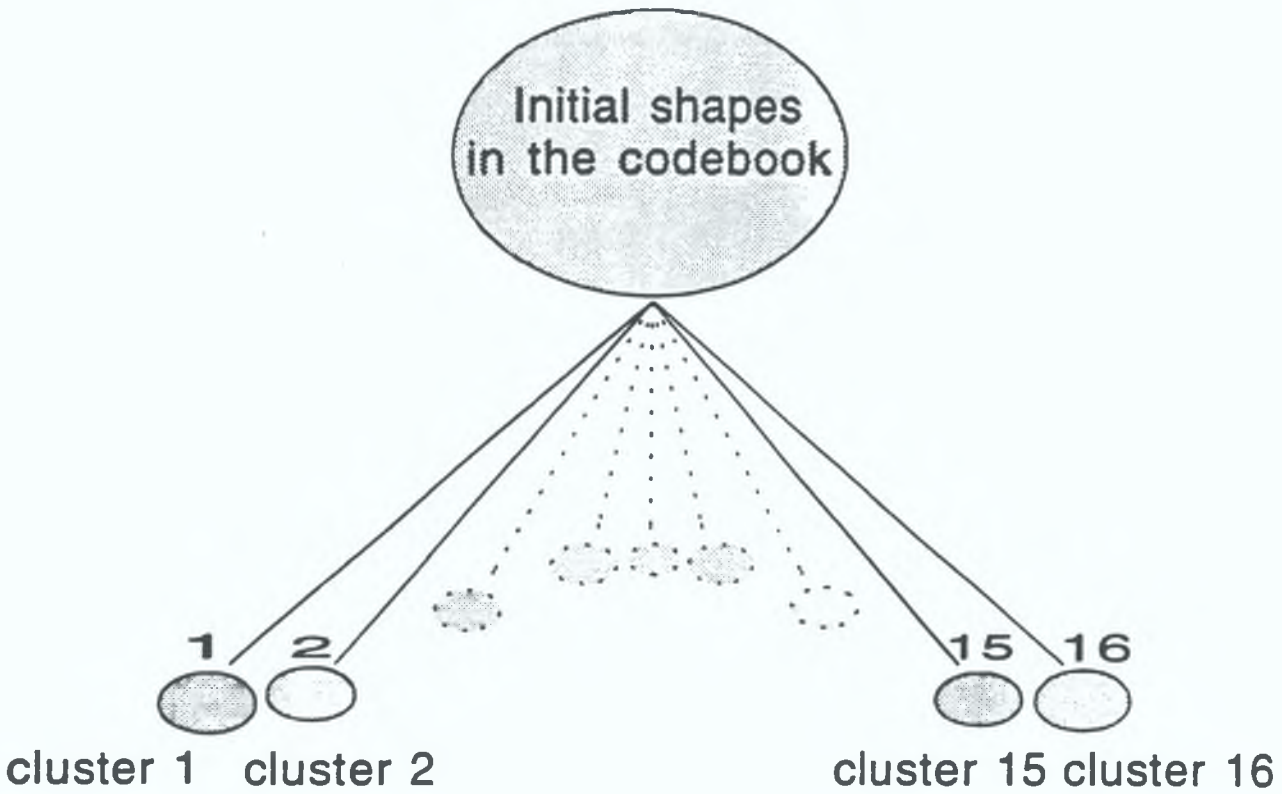


Fig 5.1 A simple tree type clustered codebook (2 - level coding) used in the preliminary study

In the estimation procedure, the cepstral coefficients of the given speech signal was firstly matched at these 16 nodes, using the normal Euclidean cepstral distance

measure. The node which gives the minimum distance is then assumed to contain a close match for the given shape. The search is then performed in that corresponding cluster using the methods described in chapter 3. This method can reduce estimation time of the parameters. The method was tested on synthetic speech. Vocal tract shapes obtained were the same as (or very similar to) those obtained using the exhaustive search. However, a clear theoretical reasoning cannot be given for clustering the cepstral coefficients using the k-means algorithm

Again, a proper distance measure should be chosen. The modified k -means algorithm used in [Sondhi et al., 1987], can be employed to cluster the codebook More research is required in clustering the codebook

5.1.3 Expansion of the codebook

The codebook in the current work was generated by a 5-parameter area function model This model is optimal for voiced sounds However, a large codebook spanning both voiced and the non-voiced sounds is required to analyse natural speech sequences Therefore a combination of the current codebook and another codebook covering the fricatives and consonants could be used for the analysis of continuous speech signals The articulatory models described in [Mermelstein, 1973; Coker, 1976] may be used to expand the codebooks

5.1.4 Non-uniqueness Problem

The problem of non-uniqueness (i.e the same speech sounds can be produced from two different area functions) in the estimation of area function can be, to a certain extent, overcome in the analysis of continuous speech.

Frame by frame analysis using a spectral criterion could find two completely different tract shapes for successive frames. However, it is reasonable to expect that vocal tract shapes cannot change erratically between two successive frames Therefore, the information on the vocal tract shape in the previous frame could be used to accurately infer the vocal tract shape in the current frame This can be done by defining a matching criterion which combines both the spectral matching and knowledge of the articulatory parameters in the previous frame. This criterion could produce a smooth area function variation for successive frames and could possibly eliminate the

non-uniqueness problem. A criterion along this line was used in the analysis of speech using Dynamic Programming [Schroeter, Sondhi, 1989].

5.1.5 An iterative procedure in the estimation of area function

The current work estimated both the area function and the glottal flow signals The area function was estimated from speech by removing the glottal and lip radiation characteristics through a pre emphasis filter. This is an approximation

A better estimate for the area function could be obtained by iterating the estimation process In the iteration, the glottal and lip radiation effects can be removed from speech, using the estimated glottal flow characteristics. This technique is similar to the iterative inverse filtering technique described in [Alku,1992].

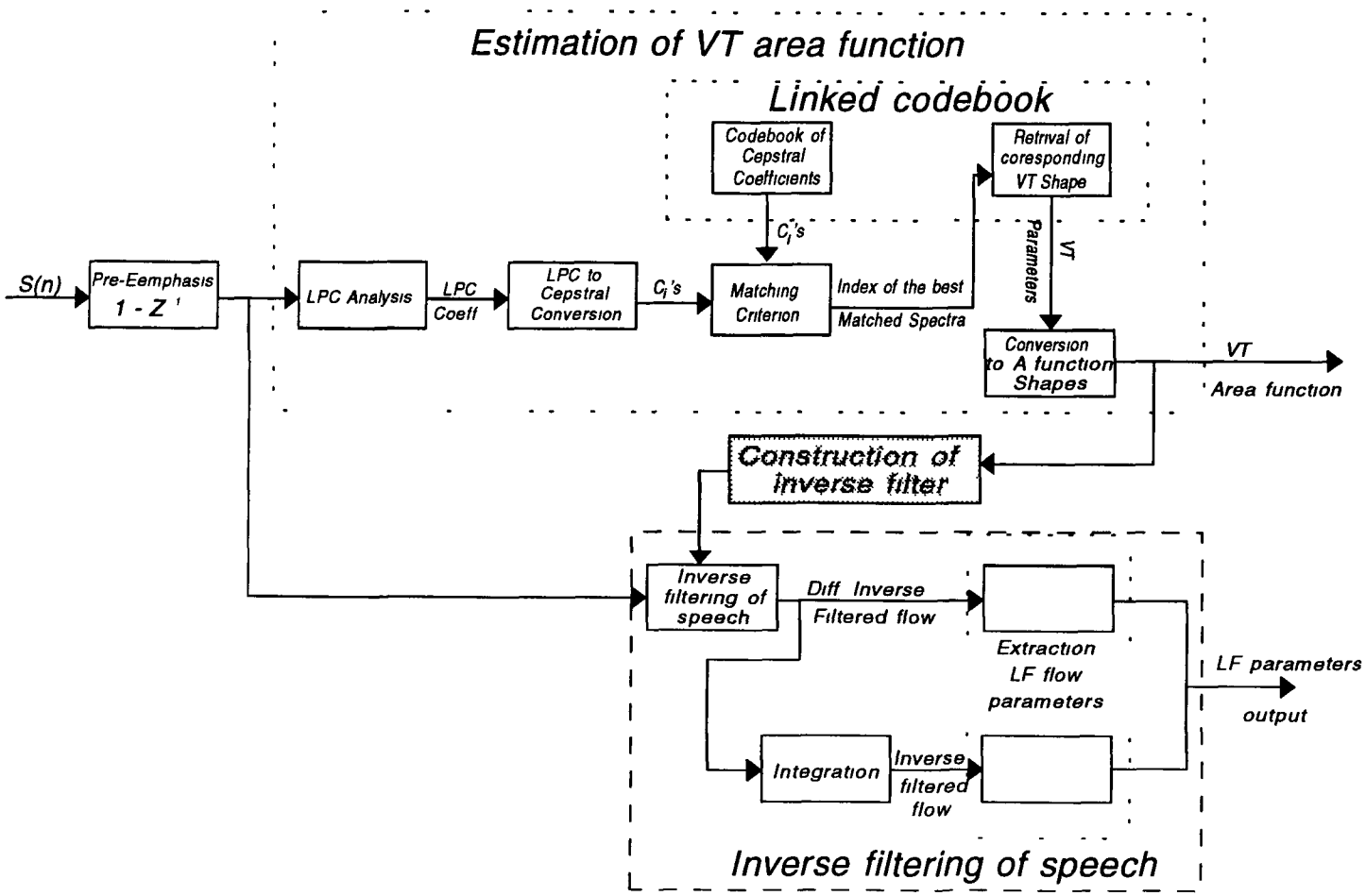


Fig 5 2 Block diagram of the current work, showing the estimation of VT area function and the inverse filtering of speech

APPENDICES

APPENDIX A

A 1.0 SPEECH SOUNDS

Can be classified into three distinct classes.

Voiced sounds: Produced by forcing the air through the glottis with the tension of the vocal cords adjusted so that they vibrate in a relaxation oscillation, thereby producing quasi-periodic of air which excite the vocal tract..

Fricative or Unvoiced sounds: Produced by forming a constriction at some point in the vocal tract, and forcing the air through the constriction at a high enough velocity to produce turbulence.

Plosive sounds: Produced by building up pressure behind a closure in the vocal tract and abruptly releasing it.

A 2.0 PHONEME CATEGORIES

Vowels : Vowels are produced by exciting a fixed vocal tract with quasi-periodic pulses air caused by vibration of the vocal cords. e.g. *uh a e i o u aa ee er uu ar aw*

Diphthong: Gliding monosyllabic speech item that starts at or near the articulatory position of one vowels and moves to or towards near the position of another. These are not classified as individual phonemes [Witten, 1982]. e.g. *eI* (bay) *oU* (boat) *aI* (buy) *aU* (how) *oI* (boy) *ju* (you)

Glide (liquid/semi vowels): Generally characterised by a gliding transition in a vocal tract between adjacent phonemes. These are vowel like sounds and hence are similar to vowels and diphthongs. e.g. *w l r y*

Stop

voiced stop: Produced by building up pressure behind a total constriction somewhere in the oral tract and suddenly releasing the pressure. e.g. *b d g*

unvoiced stop: Similar to their voiced counterparts except, during the period of total

closure of the tract, as the pressure builds up, the vocal cords do not vibrate e g *p t k*

Nasal· Produced with glottal excitation and the vocal tract totally constricted at some point along the oral passage way Air radiates both at the nostrils and mouth e.g *m n η*

Fricative:

unvoiced fricative: Produced by exciting the vocal tract by a steady airflow which becomes turbulent in the region of a constriction in the vocal tract. e.g. *f θ s sh* (/ ʃ /)

voiced fricative : Two excitation sources involved. One by the vibrating vocal cords Other, as with the unvoiced counterpart, by exciting the vocal tract by a steady airflow which becomes turbulent in the region of a constriction in the vocal tract e.g. *v th z zh*

Affricate:

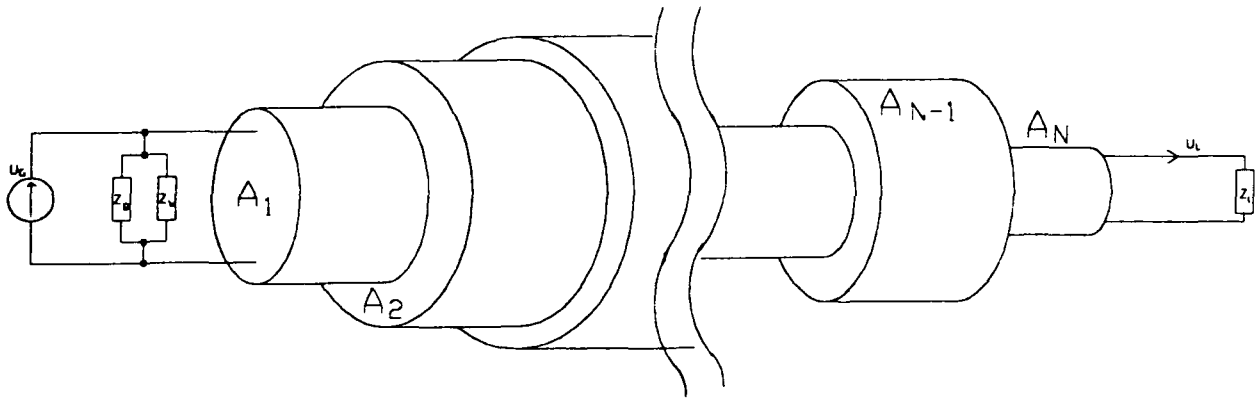
voiced affricate: Modelled by concatenation of the stop */d/* and the fricative */zh/.* e g. *J*

unvoiced affricate· Modelled by concatenation of the stop */t/* and the fricative / ʃ /. e.g *ch*

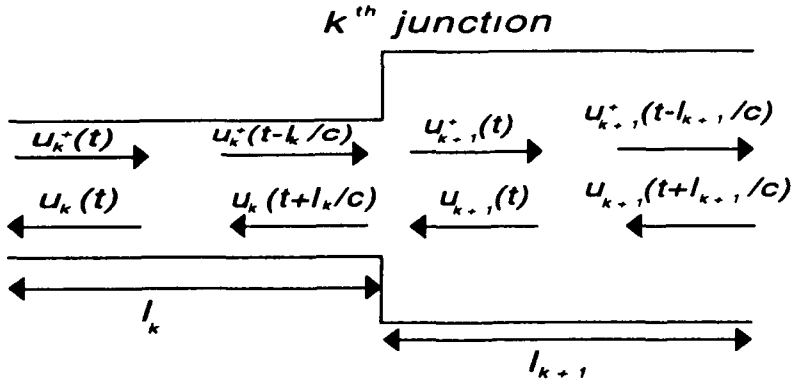
Aspirate:y Produced by a steady air flow without the vocal cord vibrating but with turbulent flow being produced at the glottis. e.g. *h*

APPENDIX B

B 1.0 Derivation of the SPM transfer function



(a)



(b)

Fig. B1 a): Vocal tract constructed of concatenated lossless cylindrical tubes. Wall impedance Z_w is lumped at the glottis with glottal impedance Z_g. Z_l is lip impedance. U_G and U_L are glottal flow and the lip flow. L is the vocal tract length and N is the number of small cylindrical tubes.

b): Wave propagation at k^{th} junction of the lossless tube model illustrated in (a). In the current work, $l_k = l_{k+1} = L/(Nc)$.

The SPM used in the current research was discussed in chapter 2 Fig 2 2 shows the Kelly - Lochbaum lattice filter type model The transfer function of the flow at the lips to the flow at the glottis (eqn. 2.3) is derived here The vocal tract was considered as

concatenation of small cylindrical lossless tubes As mentioned, the wall losses are placed at the glottis.

Sound propagation in each tube is described by the Portnoff's eqn (2 1) Applying this equation to the k^{th} tube (fig B1) of constant cross sectional area, A_k, gives the pressure p_k and the volume velocity u_k in the tube as·

$$-\frac{\delta p}{\delta x} = \frac{\rho}{A_k}\frac{\delta u}{\delta t} \qquad \text{B1 1a}$$

$$-\frac{\delta u}{\delta x} = \frac{A_k}{\rho c^2}\frac{\delta p}{\delta t} \qquad \text{B1 1b}$$

The solution to eqn (B1.1), with reference to fig. B2, has the form:

$$u_k(x,t) = [u_k^{+}(t-x/c) - u_k^{-}(t+x/c)] \qquad \text{B1 2a}$$

$$p_k(x,t) = \frac{\rho c}{A_k}[u_k^{+}(t-x/c) + u_k^{-}(t+x/c)] \qquad \text{B1 2b}$$

where x is distance measured from the left hand end of the k^{th} tube ($0 \le x \le L/N$) L is the vocal tract length and N is the number of small cylindrical tubes u_k^{+} and u_k^{-} are positive and negative going travelling waves. (L/N) is the length of a small cylindrical tube.

The relationship between the travelling waves in adjacent tubes is obtained by applying the principle the of pressure and the volume velocity must be continuous in both time and space everywhere in the system Using this principle at the junction between k^{th} and $(k+1)^{st}$ tubes gives·

$$u_k(\frac{L}{N},t) = u_{k+1}(0,t) \qquad \text{B1.3a}$$

$$p_k(\frac{L}{N},t) = p_{k+1}(0,t) \qquad \text{B1.3b}$$

Substituting eqn. (B1.2) into eqn (B1 3) with $x = (L/N)$ for k^{th} tube and $x = 0$ for $(k+1)^{st}$ tube gives:

$$u_k^+(t-\frac{L}{Nc}) - u_k^-(t+\frac{L}{Nc}) = u_{k+1}^+(t) - u_{k+1}^-(t) \qquad \text{B1 4a}$$

$$\frac{A_{k+1}}{A_k}\left[u_k^+(t-\frac{L}{Nc}) + u_k^-(t+\frac{L}{Nc})\right] = [u_{k+1}^+(t) + u_{k+1}^-(t)] \qquad \text{B1 4b}$$

Substituting the value of $u_k^-(t+L/Nc)$ from eqn (B1 4a) into eqn (B1.4b) gives

$$u_{k+1}^+(t) = \left[\frac{2A_{k+1}}{A_k+A_{k+1}}\right]u_k^+(t-\frac{L}{Nc}) + \left[\frac{A_{k+1}-A_k}{A_k+A_{k+1}}\right]u_{k+1}^-(t)] \qquad \text{B1.5a}$$

Subtracting eqn. (B1.4a) from eqn. (B1 4b) gives

$$u_{k+1}^-(t+\frac{L}{Nc}) = -\left[\frac{A_{k+1}-A_k}{A_{k+1}+A_k}\right]u_k^+(t-\frac{L}{Nc}) + \left[\frac{2A_k}{A_k+A_{k+1}}\right]u_{k+1}^-(t)] \qquad \text{B1.5b}$$

The quantity $\frac{A_{k+1} - A_k}{A_{k+1} + A_k}$, is the amount of travelling waveform reflected back at the junction and therefore called the reflection coefficient r_k Rewriting the eqn. (B1 5) in terms of r_k yields

$$u_{k+1}^+(t) = (1+r_k)u_k^+(t-\frac{L}{Nc}) + r_k u_{k+1}^-(t)] \qquad \text{B1.6a}$$

$$u_{k+1}^-(t+\frac{L}{Nc}) = -r_k u_k^+(t-\frac{L}{Nc}) + (1-r_k)u_{k+1}^-(t)] \qquad \text{B1 6b}$$

This type of equations expressed in terms of reflections coefficients are first used by Kelly - Lochbaum [Rabiner and Schafer, 1978] The eqn (B1.6) is the time domain representation of the waveform equations. The signal flow diagram of the equation is given in [Rabiner and Schafer, 1978]

The equivalent digital domain equation of eqn (B1 6) in terms of z^{-1} can now be written as:

$$U_{k+1}^{+}(z) = (1+r_k)z^{-\frac{1}{2}}U_k^{+}(z) + r_k U_{k+1}^{-}(z) \qquad \text{B1.7a}$$

$$U_k^{-}(z) = -r_k z^{-1}U_k^{+}(z) + (1-r_k)z^{-\frac{1}{2}}U_{k+1}^{-}(z) \qquad \text{B1 7b}$$

where z^{-1} and $z^{-1/2}$ are delay elements in digital domain equivalent to *((-2L) / (Nc))* and *(-L/(Nc))*. Flow in digital domain is expressed by the capital U, whereas in time domain it is expressed by small u

With some rearrangements, eqn. (B1.7) can be written in matrix form as:

$$U_k = Q_k\, U_{k+1} \qquad \text{B3.18}$$

where U_k and Q_k are matrices given by (note that U_k and Q_k are matrices and therefore shown in bolded letters)

$$U_k = \begin{bmatrix} U_k^{+}(z) \\ U_k^{-}(z) \end{bmatrix}$$

and

$$Q_k = \begin{bmatrix} \frac{z^{\frac{1}{2}}}{1+r_k} & \frac{-r_k z^{\frac{1}{2}}}{1+r_k} \\ \frac{-r_k z^{-\frac{1}{2}}}{1+r_k} & \frac{z^{-\frac{1}{2}}}{1+r_k} \end{bmatrix}$$

We have so far derived the equations for wave propagation at kth junction Clearly, except at the boundaries i.e at the lips and at the glottis, the forward and the backward going waves at each junction of the system can be described by the eqn (B1 8).

Therefore by repeatedly, applying eqn. (B1.8), the input to the first tube U_1 can be related to the output at the beginning of last tube (note. not at the lips)U_N by the matrix product

$$U_1 = Q_1 . Q_2 . \; . \; . Q_{N-1} . U_N \qquad \text{B1.9}$$

B 1.1 Boundary condition at the lips

At the lips (at the end of N^{th} tube), pressure and the flow are related by the frequency domain relation of the form (details are in [Rabiner and Schafer, 1978])

$$P_N(l_N, \Omega) = Z_L U_N(l_N, \Omega) \qquad \text{B1 10}$$

Z_L is the radiation impedance at the lips. Rewriting eqn (B1 10) in terms of forward and backward going waves using eqn. (B1.2) in N^{th} tube gives

$$\frac{\rho c}{A_N}\left[u_N^+(t-\frac{L}{Nc}) + u_N^-(t+\frac{L}{Nc})\right] = Z_L\left[u_N^+(t-\frac{L}{Nc}) - u_N^-(t+\frac{L}{Nc})\right] \qquad \text{B1 11}$$

Solving for $u_N^-(t+\frac{L}{Nc})$ gives

$$u_N^-(t+\frac{L}{Nc}) = -r_L u_N^+(t-\frac{L}{Nc}) \qquad \text{B1.12}$$

where

$$r_L = \left[\frac{\frac{\rho c}{A_N} - Z_L}{\frac{\rho c}{A_N} + Z_L}\right]$$

writing eqn. (B1.12) in the digital domain gives (refer [Rabiner and Schafer, 1978])

$$U_N^-(z) = -r_L z^{-1} U_N^+(z) \qquad \text{B1.13}$$

Since there is no reflected wave at the lip end, the output volume velocity at the lips u_L (equivalent to u_{N+1}), using eqn (B.4a) and then eqn (B1.12) is given by

$$u_L = u_N^+(t - \frac{L}{Nc}) - u_N^-(t + \frac{L}{Nc})$$

$$= (1 + r_L) u_N^+(t - \frac{L}{Nc}) \qquad \text{B1 14}$$

In digital domain it is written as

$$U_L(z) = (1 + r_L) z^{-\frac{1}{2}} U_N^+(z) \qquad \text{B1 15}$$

If we now rewrite eqn (B1.13) and eqn (B1 15) in terms of U_L, we will obtain,

$$U_N^-(z) = \frac{-r_L}{1 + r_L} z^{-\frac{1}{2}} U_L(z) \qquad \text{B1.16}$$

$$U_N^+(z) = \frac{1}{1 + r_L} z^{\frac{1}{2}} U_L(z) \qquad \text{B1.17}$$

Eqn. (B1.16) and eqn (B1 17) are now written in matrix form as

$$U_N = \begin{bmatrix} \frac{z^{\frac{1}{2}}}{1 + r_L} \\ \frac{-r_L z^{-\frac{1}{2}}}{1 + r_L} \end{bmatrix} U_L \qquad \text{B1 18}$$

B 1.2 Boundary condition at the glottis

At the glottis, assuming that the excitation source is linearly separable from the vocal tract, using the principle of continuity, flow in the first tube is given by [Rabiner and Schafer, 1978]

$$U_1(0,\Omega) = U_G(\Omega) - P_1(0,\Omega)/Z_G \qquad \text{B1.19}$$

where U_1 and P_1 are the pressure and volume velocity at the input of the first tube, Z_g is the glottal acoustic impedance

Rewriting eqn (B1.19) in terms of forward and backward going waves, we obtain

$$u_1^+(t) - u_1^-(t) = u_G(t) - \frac{\rho c}{A_1}\left[\frac{u_1^+(t) + u_1^-(t)}{Z_G}\right] \qquad \text{B1 20}$$

Solving for $u_1^+(t)$, we obtain

$$u_1^+(t) = \frac{(1 + r_G)}{2}u_G(t) + r_G u_1^-(t) \qquad \text{B1 21}$$

where the reflection coefficient at the glottis r_G is given by

$$r_G = \left[\frac{Z_G - \frac{\rho c}{A_1}}{Z_G + \frac{\rho c}{A_1}}\right]$$

In digital domain eqn (B1.21) with $U_G(z)$ in the right hand side is written as

$$U_G(z) = \frac{2}{(1+r_G)}U_1^+(z) - \frac{2r_G}{(1+r_G)}U_1^-(z) \qquad \text{B1.22}$$

In matrix form it is written as

$$U_G(z) = \left[\frac{2}{(1 + r_G)}, -\frac{2r_G}{(1 + r_G)}\right]U_1 \qquad \text{B1.23}$$

B 1.3 Complete equation of the model

The transfer function of the lattice filter SPM can be now written by substituting the values for U_1 and U_N from eqn. (B1.9) and eqn. (B1 18) into eqn (B1 24) i e

$$U_G(z) = \left[\frac{2}{(1 + r_G)}, -\frac{2r_G}{(1 + r_G)}\right]Q_1 \cdots Q_{N-1}\begin{bmatrix} \frac{z^{\frac{1}{2}}}{1+r_L} \\ \frac{-r_L z^{-\frac{1}{2}}}{1+r_L} \end{bmatrix} U_L(z) \qquad \text{B1.25}$$

where the values of Q_k are given eqn (B1 8)

The transfer function of the model $V(z)$, i e; output at the lips to the input at the glottis is related in the frequency domain as

$$V(z) = \frac{U_L(z)}{U_G(z)} = \frac{0.5(1+r_G)(1+r_L)\prod_{k=1}^{N-1}(1+r_k)z^{\frac{-N}{2}}}{D(z)} \qquad \text{B1 26}$$

where $D(z)$is given by

$$D(z) = [1, -r_G)]Q_1 \cdot \cdot \cdot Q_{N-1}\begin{bmatrix} 1 \\ -r_L z^{-1} \end{bmatrix} \qquad \text{B1.27}$$

B 2.0 Flow chart for implementing the 5 parameter area function model

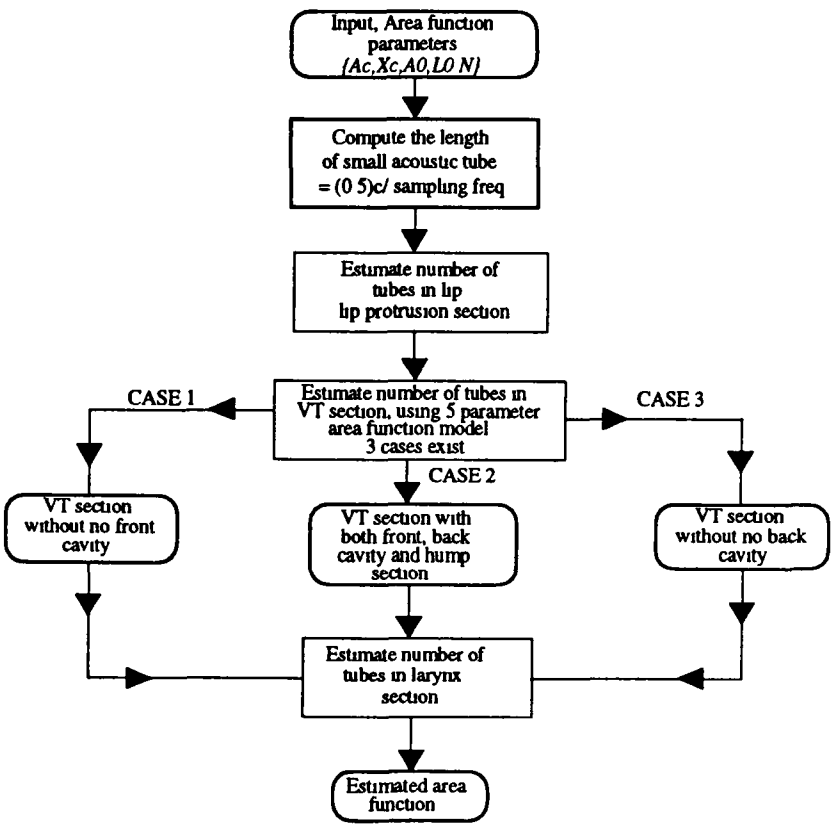


Fig. B2 Flow chart for synthesising area function from the parameters A_c, X_c, A_0, L_0 and N. Area of each tube is calculated by matching the area under each tube segment analog area function.

B 3.0 Derivation of 2 pole - 2 zero filter for composite glottal termination

A 2 pole - 2 zero digital model for the composite glottal impedance of the network, used in the SPM (fig. 2.5 repeated again in fig 3), is derived here [Scaife, 1989] The analog impedance can be simulated in digital domain using bilinear transformation However, the digital model in the current work was obtained by fitting the analog model to a 2 pole - 2 zero digital model at different frequencies

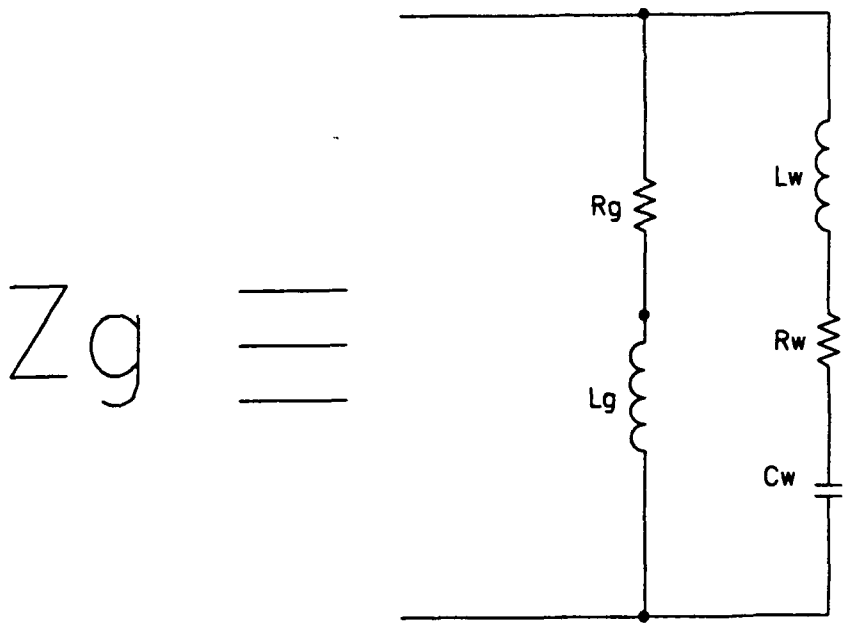


Fig. 2.5 Composite termination at the glottal end, modelled by the parallel combination of wall and glottal impedance

The glottal impedance Z_T of the net work shown in fig B3, in analog frequency domain, is given by eqn B(3 1):

$$Z_T = R_g \left\{ \frac{s^2 + \frac{R_w}{L_w} + \frac{1}{L_w C_w}}{s^2 + \frac{(R_g + R_w)}{L_w} + \frac{1}{L_w C_w}} \right\} \qquad \text{B3 1}$$

The proposed digital model is given by eqn (B3 2) as·

$$\bar{Z}_T = K_0 \frac{(z-a)^2}{(z-b)(z-d)} \qquad \text{B3.2}$$

where a is the zero of the model and b and d are poles of the transfer function At low frequencies the z - domain model can be approximated by substituting $z = e^{j\Omega\tau} = cos(\Omega\tau) + j\, sin(\Omega\tau) \approx 1 + j\,\Omega\tau$ in the transfer function. Substituting $1 + j\,\Omega\tau$ for z in eqn. (3.2) gives:

$$\bar{Z}_T \approx K_0 \frac{(1+j\Omega\tau-(1-e))^2}{(1+j\Omega\tau-(1-u))\ (1+j\Omega\tau-(1-k))}$$

$$= K_0 \frac{(j\Omega\tau+e)^2}{(j\Omega\tau+u)\ (j\Omega\tau+k)} \qquad \text{B3 3}$$

where $1 - e = a$, $1 - u = b$ and $1 - k = d$ and K_0 is a constant.

The values of these constants are found by matching the digital transfer function (eqn (3.3)) with the analog impedance in eqn. (3.1) at 0 Hz, resonance and at frequencies much larger than the resonance frequency

At DC ; $\Omega = 0$

$$Z_T = R_s \qquad \text{and} \qquad \bar{Z}_T = \frac{K_0 e^2}{uk}$$

Therefore $K_0 e^2 = R_s$ B3.4a

At resonance i.e. at $\Omega = 1 / \sqrt{L_w C_w}$

$$Z_T = \frac{R_w R_g}{R_w + R_g} \approx R_w$$

and

$$\bar{Z}_T = \frac{K_0(je+e)^2}{(je+u)(je+k)} \approx \frac{K_0 e^2 (j+1)^2}{jek + jeu} = \frac{2K_0 e}{k}$$

Therefore $\frac{2K_0 e}{k} = R_w$ B3 4b

At frequencies above resonance i.e $\Omega \gg 1 / \sqrt{L_w C_w}$

$$Z_T = \frac{R_g L_w s}{R_g + L_w s} \approx R_g \frac{s}{\left[\frac{R_g}{L_w} + s\right]}$$

and

$$\bar{Z}_T = \frac{K_0(j\Omega\tau+e)^2}{(j\Omega\tau+u)\ (j\Omega\tau+k)} = \frac{K_0(j\Omega\tau)^2}{(j\Omega\tau)\ (j\Omega\tau+k)} = \frac{K_0(j\Omega\tau)}{(j\Omega\tau+k)}$$

Substituting $s = j\Omega\tau$ in Z_T gives

$$Z_T = \left[R_g \frac{s}{\left[\frac{R_g}{L_w} + s\right]}\right]_{s=j\Omega\tau} = \frac{R_g(j\Omega\tau)}{(j\Omega\tau) + \frac{R_g}{L_w}\tau}$$

Now by comparing Z_T with Z_T, we can write

$$K_0 = R_g \qquad \text{and} \qquad k = \frac{\tau R_g}{L_w} \qquad \text{B3.4c}$$

We already know the value for e i.e. $e = 1/\sqrt{L_w C_w}$. Therefore from eqn. (B3.4) values for K_0, k and u can be found. The values of R_g, R_w, C_w and L_w are given as

$R_g = 110\ \Omega$, $\qquad R_w = 8\ \Omega$

$C_w = 0.0008\ \text{cm}^5\ \text{N}^{-1}$ $\qquad L_w = 0.012\ \text{dyne s cm}^{-5}$

B 4.0 A lossy vocal tract SPM (SPM2)

In chapter 2 SPM used in the current research was discussed. In appendix B1 transfer function of the lossless SPM was derived. The losses in the vocal tract were accounted for by placing an equivalent parametric model at the glottis or at the lip end (chapter 2). This simplifies the digital filter modelling. In reality these losses are of a distributed or discrete nature and occur along the vocal tract.

In this section, a SPM (SPM2) is described by placing the effects of distributive and discrete losses along the vocal tract. The lattice filter coefficients for this lossy SPM are derived using *loss factors* D [Liljencrants, 1985]. The reflection coefficients obtained here are similar to those of the lossless vocal tract SPM. With the substitution of D = 0, the reflection coefficients values reduce to those of lossless case. The structure of the synthesis filter model and its inverse filter model, is same as the SPM described in section 2.3. Only, the reflection coefficients are different. The magnitude response of this lossy model are calculated using the program "lossy.C" in appendix B5. Preliminary results obtained show that the formant frequencies and bandwidths are very much comparable with those of the SPM used in the current research.

Calculations involved in this model (SPM2) is high compare to the SPM used in the current research. SPM2 is not fully investigated in the present work. However we feel that SPM2 with a constrained vocal tract model may be used in future research work, to improve the area function estimations.

By considering z-domain models for internal losses e.g. viscous losses, laminar losses, turbulent losses and wall vibration losses, a lossy vocal tract is included in the

Speech Production Model In the implementation, distributive losses are considered as *series loss factors* and discrete losses as *shunt loss factors*. Reflection coefficients are derived relating the forward waves, backward waves and the series and shunt loss factors [Liljencrants, 1985]. Using these equations, lossy synthesis filter and inverse filter models are implemented

B 4.1 Series and Shunt losses in the system

Detailed descriptions and discussions of internal losses in the human speech production system are given in many classical speech text books [Fant, 1960, Flanagan, 1972] Transmission line electrical analog theory is used to simulate the lossy speech production system in the above books Typically, a lossy cylindrical pipe was used for the wave propagation. The electrical equivalent circuit of the lossy pipe was given in terms resistor R, shunt conductance G, per unit capacitance C and per unit inductance L (see [Flanagan, 1972] for details) From this electrical analogy, equivalent acoustic R, G, L and C were derived to represent the losses

In [Liljencrants, 1985], losses were represented as series and shunt z-domain loss factors and were incorporated into the scattering equations to synthesize speech We also, use the same idea to represent the losses as loss factors and incorporate them into the lossy synthesis filter and inverse filter model.

B 4.2 Derivation of the wave equations for the lossy model

A series loss resistance R for distributive loss (fig. B4.1) and shunt conductance G for discrete loss (fig B4.2) were introduced in [Liljencrants, 1985] between two tube junctions. This enables the wave equations to be expressed in terms of a new set of reflection coefficients and loss factors

B 4.3 Series loss:

Series losses such as viscous, laminar and turbulent are accounted for in the wave equations by inserting the loss resistor R at the tube junctions fig B4 1

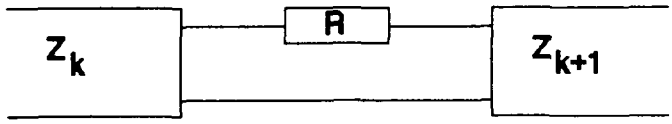


Fig. B4.1 Representation of series loss between two tubes

where Z_k and Z_{k+1} are the acoustic impedances of the k^{th} and $(k+1)^{th}$ tubes respectively

If we consider the above k^{th} tube junction as similar to three tube junction we obtain two reflection coefficients r^+ and r^- as [Liljencrants, 1985]:

$$r^+_k = \frac{(R + Z_{k+1}) - Z_k}{(R + Z_k + Z_{k+1})} \qquad r^-_k = \frac{Z_{k+1} - (R + Z_k)}{(R + Z_k + Z_{k+1})} \qquad \text{B4.1}$$

The forward waves (U^+) and backward waves (U^-) in the tubes are expressed using the normal continuity equations (see [Rabiner and Schafer, 1978]) but in terms of the above reflection coefficients r^+ and r^- as (similar to eqn. (B1.7))·

$$U^+_{k+1} = (1-r^+_k)z^{-1/2}U^+_k - r^-_k U^-_{k+1} \qquad \text{B4.2a}$$

$$U^-_k = r^+_k z^{-1}U^+_k + (1+r^-_k)z^{-1/2}U^-_{k+1} \qquad \text{B4.2b}$$

To express the above set of eqn. (B4.2) in terms of the constant D (loss factor) and r (lossless tube model reflection coefficient), we write r^+ and r^- as·

$$r^+ = \frac{r+D}{1+D} \qquad r^- = \frac{r-D}{1+D} \qquad \text{B4 3}$$

where loss factor D is defined as the ratio of R to the sum of the surrounding line impedances Z_k and Z_{k+1} and r is the reflection coefficient between two tubes(as in appendix B1).

$$D = \frac{R}{Z_k + Z_{k+1}}$$ B4 4

$$r_k = \frac{Z_{k+1} - Z_k}{Z_{k+1} + Z_k} = \frac{\frac{\rho c}{A_{k+1}} - \frac{\rho c}{A_k}}{\frac{\rho c}{A_{k+1}} + \frac{\rho c}{A_k}} = \frac{A_k - A_{k+1}}{A_k + A_{k+1}}$$ B4.5

Using the r^+_k and r_k from eqn. (B4 3) and defining $X = D/(1+D)$ allow equations (B4.2a) and (B4.2b) to be written as :

$$U^+_k = \frac{1}{(1-r_k).(1-X)} . z^{1/2} U^+_{k+1} + \left[\frac{r_k}{(1-r_k)} - \frac{X}{(1-r_k)(1-X)} \right] . z^{1/2} . U^-_{k+1}$$ B4 6a

$$U^-_k = \left[\frac{X}{(1-r_k)(1-X)} + \frac{r_k}{(1-r_k)} \right] . z^{-1/2} . U^+_{k+1} + \left[\frac{1}{(1-r_k)} - \frac{X}{(1-r_k)(1-X)} \right] . z^{-1/2} . U^-_{k+1}$$ B4.6b

These equations can then be expressed in matrix form as (similar to eqn (B1 8))

$$\begin{bmatrix} U^+_k \\ U^-_k \end{bmatrix} = z^{1/2} \begin{bmatrix} \frac{1}{(1-r_k)(1-X)} & \frac{r_k}{(1-r_k)} - \frac{X}{(1-r_k)(1-X)} \\ \left[\frac{r_k}{(1-r_k)} + \frac{X}{(1-r_k)(1-X)} \right] z^{-1} & \left[\frac{1}{(1-r_k)} - \frac{X}{(1-r_k)(1-X)} \right] z^{-1} \end{bmatrix} \begin{bmatrix} U^+_{k+1} \\ U^-_{k+1} \end{bmatrix}$$ B4 7

If the vocal tract is represented by N small cylindrical tubes, by repeatedly applying eqn. (B4 7) (similar to the equations in appendix B1), it can be shown that the input at the glottis U_G can be expressed in terms of the output flow at the lips U_L, by a matrix product of the form (similar to eqn (B1.25)

$$U_G = z^{\frac{-N}{2}} \begin{bmatrix} \frac{2}{(1+r_G)} & \frac{-2r_G}{1+r_G} \end{bmatrix} \prod_{l=1}^{N-1} Q_k \begin{bmatrix} 1 \\ -r_L z^{-1} \end{bmatrix} U_L \qquad \text{B4 8}$$

where

$$Q_k = \begin{bmatrix} \frac{1}{(1-r_k)(1-X_k)} & \frac{r_k}{(1-r_k)} - \frac{X_k}{(1-r_k)(1-X_k)} \\ \frac{r_k}{(1-r_k)} + \frac{X_k}{(1-r_k)(1-X_k)} & \frac{1}{(1-r_k)} - \frac{X_k}{(1-r_k)(1-X_k)} \end{bmatrix} \qquad \text{B4 9}$$

, r_G and r_L are reflection coefficients at the glottis and lips respectively.

Incorporating the series losses in the lossless model modifies the lossless equations derived in the appendix B1. It is noticeable from the equations that an extra term X due to the series losses is included in the wave equation. Having obtained this equation, it can be now implemented in the lattice filter model as shown in fig B4 3

The equations have been derived for the series lossy model considering only one type of distributive loss. However, as was mentioned earlier the losses manifest differently in the vocal tract e.g viscous loss, laminar loss, etc and hence will have different loss factors D. Fortunately these losses are of an additive nature and hence the losses can be combined m the lossy eqn. (B4 4) to give the same equation but with X defined as:

$$X = \sum_{l=1}^{N} \frac{D_l}{1 + D_l} \qquad \text{B4.10}$$

where D_l are the loss factors for different lossy mechanisms.

B 4.4 Proof of eqn. (B4.7) with several series loss mechanisms

Substituting r^+ and r^- from eqn. (B4 3) into eqn (B4.2) and after some manipulations, eqn. (B4 2) can be written as

$$U^{+}_{k+1} = U^{+}_{k} - (U^{+}_{k}.z^{-1/2} + U^{-}_{k+1})\, r_k + (U^{+}_{k}(1-r_k).z^{-1/2} - U^{-}_{k+1}(1+r_k))\, \frac{D}{1+D} \qquad \text{B4 11a}$$

$$U^{-}_{k} = U^{-}_{k+1}\, z^{-1/2} + (U^{+}_{k}.z^{-1} + U^{-}_{k+1}\, z^{-1/2}) r_k + (U^{+}_{k}(1-r_k)\, z^{-1} - U^{-}_{k+1}(1+r_k)\, z^{-1/2}) \frac{D}{1+D}$$

B4.11b

As can be seen clearly from eqn (B4 11) the first two terms arise from the lossless tube model and only the third term is due to the loss factor. If we, therefore include different loss mechanisms in the model, only the last term of eqn (B4.11) will change Fortunately, these losses can be combined to give eqn (B4 7) with X defined as in eqn (B4 10)

B 4.5 Shunt losses

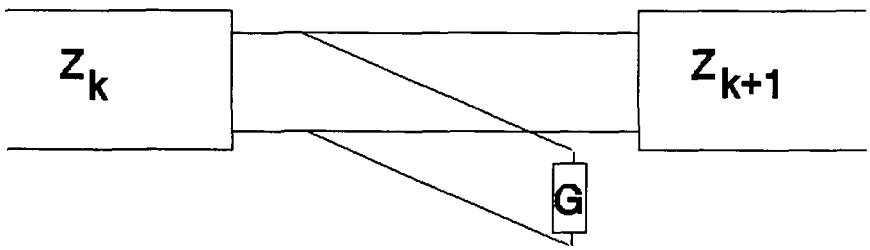


Fig. B4.2 Representation of shunt loss between two tubes

In a similar manner to the series losses, shunt losses are also included in the wave equations In this case, the shunt loss factor E is defined in terms of shunt conductance G and surrounding line impedances Z_k and Z_{k+1} as [Liljencrants, 1985]

$$G = E\,(\frac{1}{Z_k} + \frac{1}{Z_{k+1}}) \qquad \text{B4 12}$$

with the substitution of $Z = \rho.c/A$, the eqn (B4 12) can be rewritten as

$$\rho\, c.G = E\,(A_k + A_{k+1}) = A_Y \qquad \text{B4 13}$$

where area A_Y represents the shunt loss branch of the three way tube junction Unlike

in the series case, energy lost in the shunt case through the branched tube does not reflect back into the tube. Using this fact the wave equations for the shunt loss model can be derived in a similar fashion to those of the series loss as follows

$$\begin{bmatrix} U^{+}_{k} \\ U^{-}_{k} \end{bmatrix} = z^{1/2} \begin{bmatrix} \dfrac{1}{(1-r_k)(1-Y)} & \dfrac{r}{(1-r_k)} + \dfrac{Y}{(1-r_k)(1-Y)} \\ \left[\dfrac{r_k}{(1-r_k)} - \dfrac{Y}{(1-r_k)(1-Y)} \right] z^{-1} & \left[\dfrac{1}{(1-r_k)} - \dfrac{Y}{(1-r_k)(1-Y)} \right] z^{-1} \end{bmatrix} \begin{bmatrix} U^{+}_{k+1} \\ U^{-}_{k+1} \end{bmatrix}$$

B4 14

where

$$Y = \sum_{i=1}^{N} \frac{E_i}{1 + E_i}$$

B4.15

with E_i being different *shunt loss factors* due to discrete losses

B 4.6 Complete model with Series and Shunt Losses

A complete synthesis model for a speech production system is given in this section by including both the series and shunt losses. Since series and shunt losses can be applied additively we can write the wave equations for this complete model as:

$$U^{+}_{k+1} = U^{+}_{k} - (U^{+}_{k} z^{-1/2} + U^{-}_{k+1})\, r_k + (U^{+}_{k}(1-r_k)\, z^{-1/2} - U^{-}_{k+1}(1+r_k))\, X - (U^{+}_{k} + U^{-}_{k+1}).(1-r_k).Y$$

B4.16a

$$U^{-}_{k} = U^{-}_{k+1} z^{-1/2} + (U^{+}_{k} z^{-1} + U^{-}_{k+1} z^{-1/2}).r_k + (U^{+}_{k}(1-r_k)\, z^{-1} - U^{+}_{k+1}(1+r_k)\, z^{-1/2})\, X - (U^{+}_{k} + U^{-}_{k+1}).(1+r_k)\, Y$$

B4.16b

The lattice filter structure of the complete model is the same as the one shown in fig. 2.2 (chapter 2). However due to the extra loss terms in the equations a typical k^{th} junction is implemented as shown in fig B4 3, where the *series and shunt loss factors* X and Y are function of z^{-1}.

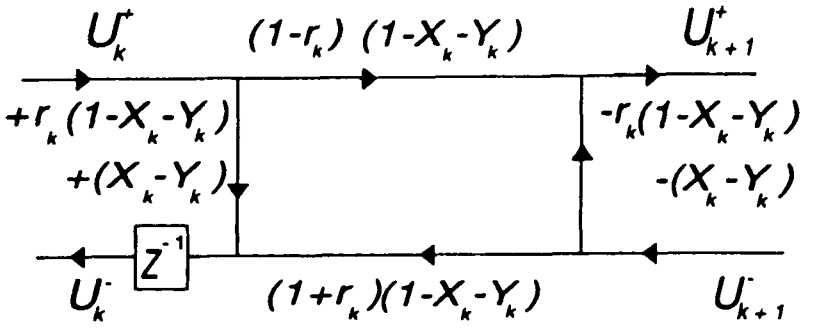


Fig. B4.3 k^{th} junction of the distributed lossy synthesis filter model

The inverse filter model of the system is implemented by rewriting the eqn. (B4.13) with U^+_{k+1} and U^-_{k+1} in the left hand side as given below in eqn (B4 17).

$$U^+_{k+1} = (U^+_k + (r_k.(1-X-Y)-(X-Y)).U^-_k.z^{-1}).\frac{1}{(1-r_k).(1-X-Y)} \qquad \text{B4 17a}$$

$$U^-_{k+1} = (\ (r_k.(1-X-Y)+(X-Y)).U^+_k + (1-2Y)(1-2X)\ U^-_k z^{-1})\ \frac{1}{(1-r_k).(1-X-Y)} \qquad \text{B4.17b}$$

These equations are now used to implement the inverse model as shown in fig B4 4.

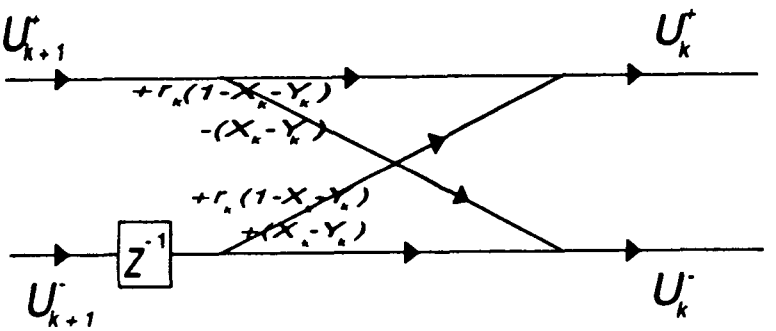


Fig. B4.4 k^{th} junction of the distributed lossy inverse filter model

B 4.7 Summary on SPM2

In SPM2, vocal tract is modelled by including the distributive and discrete losses. A set of reflection coefficients were obtained for the implementation of analysis and

synthesis model. The structure of SPM2 is same as the SPM used in the current work.

SPM2 is used in section 2.3 3 1 to study the effects of placing the wall vibration load at different points along the vocal tract Model was used to study other losses such as viscous and heat losses Digital models used in [Liljencrants, 1985] were employed to model these losses. Inclusion of the losses seem to affect the transfer function in the high frequency range.

In section 2 5, a lossy vocal tract section was included to model the speech production system. A set of reflection coefficients similar to those of lossless vocal tract SPM were obtained This SPM2 produces similar formant and bandwidth values as the SPM in section 2.3 However, this SPM2 is not used for the current research This model may be investigated in a future research.

APPENDIX C

C 1.0 Simplex method

The downhill simplex method [Press *et al* , 1988, pp 289-293] is an algorithm, used for multi-dimensional minimisation of a function Simplex is a geometrical figure consisting, in *n* dimensions, of *n+1* points (vertices) and all their interconnecting lines segments, polygonal faces, etc In two dimensions, a simplex is a triangle In three dimensions, it is a tetrahedron and so on.

The simplex method of minimisation must be started with *n+1* starting points (i.e. n+1 function values) for *n* variables, defining an initial simplex The simplex method then goes through a series of iterative steps.

Firstly, the point of the simplex where function is largest is moved to the opposite face of the simplex (reflections) to a lower point (i.e to a low function value) If it can find a lower function value, the method expands in to another direction to take larger steps. If a lower point is not found through reflections, the method tries for different methods such as reflections and expansion *or* reflection and contraction *or* a contraction along all dimensions of the simplex to find a low point of the function These steps are iterated until the function reaches the lowest value.

If the given function has several minima then the simplex method may converge to one of the local minima points Therefore it is always advisable to start the minimisation procedure with different starting values.

Termination criterion can be difficult in multi dimensional minimisation Therefore, the user has to provide a termination criterion that is suited for the function (see [Press et al., 1988] for further details)

C 2.0 Flowchart for estimating the magnitude response of the SPM

Magnitude response of the SPM (eqn B(1.26)) was implemented using a "C" program The numerator part of the eqn B(1.27) is a constant and can be implemented easily The denommator of eqn. (B1.26) is described by three matrix section

a) 1 X 2 matrix [1, $-r_G$] relating to the glottal section

b) 2 X 2 matrices of (N-1) sections Q_1 . Q_{N-1} relating to the vocal tract

c) 2 X 1 matrix $\begin{bmatrix} 1 \\ -r_L z^{-1} \end{bmatrix}$ relating to the lip radiation section

The flowchart for the implementation is given in Fig C2.

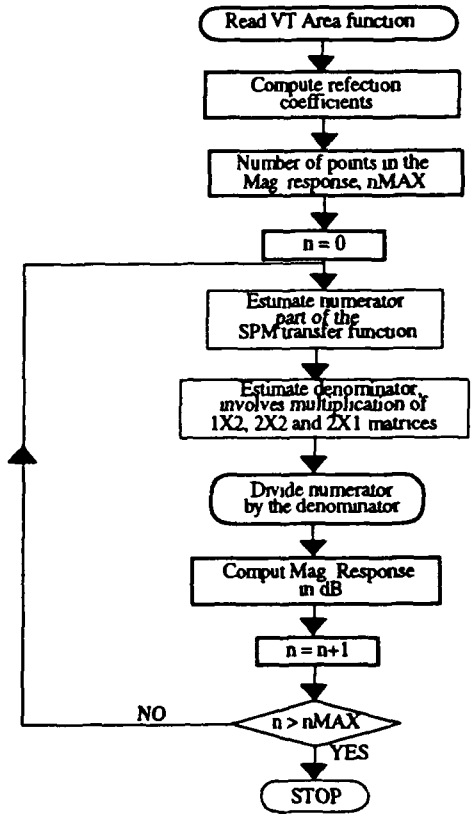


Fig. C2 Flowchart for estimating the magnitude response of the SPM

C 3.0 Cepstrum

The complex cepstrum is defined as the inverse fourier transform of the logarithm of the power spectrum $X(e^{j\omega})$ of a signal $X(n)$ [Rabiner and Schafer, 1978]. i.e.

$$\hat{x}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{X}(e^{j\omega}) e^{j\omega n} d\omega \qquad \text{C3.1}$$

where

$$\hat{X}(n) = \ln|X(e^{j\omega})| + j\arg[X(e^{j\omega})]$$

The real cepstrum is defined as:

$$c(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln|X(e^{j\omega})|\, e^{j\omega n} d\omega \qquad -\infty < n < \infty \qquad \text{C3.2}$$

C 3.1 Relationship between the Linear Predictive coefficients and the complex cepstral coefficients

A simple recursive relationship between the cepstral and the LPC coefficients is derived here using the definition

$$\ln(H(z)) = \sum_{n=1}^{\infty} c_n z^{-n} \qquad \text{C3 3}$$

where c_n - complex cepstral coefficients and

$$H(z) = \frac{G}{1 - \sum_{k=1}^{P} a_k z^{-k}} \qquad \text{C3 4}$$

G - Gain of the LPC model

P - number of LPC coefficients used in the LPC model.

Taking the derivatives in both sides of eqn.(C4.3) with respect to z^{-1} gives:

$$\frac{d[H(z)]}{dz^{-1}} = \frac{d\left[\sum_{n=1}^{\infty} c_n z^{-n}\right]}{dz^{-1}} \qquad \text{C3.5}$$

This eqn (C4.5) is simplified to

$$\sum_{k=1}^{P} k a_k z^{-k+1} = 1 - \sum_{k=1}^{P} a_k z^{-k} \cdot \sum_{n=1}^{\infty} n c_n z^{-n+1} \qquad \text{C3.6}$$

If we now equate the constant terms and the various powers of z^{-1} on the left and the right of the eqn (C4 5) we obtain the recursive relationship between a_n's and c_n's namely.

$$c_1 = a_1$$

$$c_n = \sum_{k=1}^{n-1} \left[1 - \frac{k}{n}\right] a_k c_{n-k} + a_n \qquad 1 < n < P \qquad \text{C3.7}$$

$$c_n = \sum_{k=1}^{n-1} \left[1 - \frac{k}{n}\right] a_k c_{n-k} \qquad n > P$$

The zeroth cepstral coefficient can be found by equating the constant coefficient in eqn.(C3.3), which gives:

$$c_0 = \ln(G) \qquad \text{C3.8}$$

Another very useful property of the minimum phase z transfer function is that the cepstrum is zero for $n<0$ and the complex cepstrum, which is an even function of n is related to the real cepstrum by:

$$c_n = 0 \qquad n < 0$$

$$c_n = r_n \qquad n = 0 \qquad \text{C3.9}$$

$$c_n = 2r_n \qquad n > 0$$

where c_n - complex cepstrum

r_n - real cepstrum.

It is worth mentioning that the cepstral coefficients for an all pole model can be found recursively from the impulse response of an all pole model [Atal, 19]

C 3.2 Equivalency of Log Spectral Distances (LSD) and the Euclidean cepstral distances

It is proved here that the Log Spectral Distances ($L\ S\ D$) between two frames is equal to the Euclidean cepstral distances between those two frames This was used in the

current work, to estimate the vocal tract area function parameters from the real speech signals (chapter 3)

L.S.D between two spectra $A(\omega)$ and $B(\omega)$ is

$$L.S.D = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\ln(A(\omega)) - \ln(B(\omega))\right]^2 d\omega \qquad \text{C3.10}$$

using the definition of real cepstrum

$$\ln(A(\omega)) = \sum_{-\infty}^{\infty} c_a(k)\, e^{-j\omega k} \qquad \text{C3.11}$$

and substituting this into eqn(C3 10) yields

$$L.S.D = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\sum_{-\infty}^{\infty} C_a(k)\, e^{-j\omega k} - \sum_{-\infty}^{\infty} C_b(k)\, e^{-j\omega k}\right]^2 d\omega \qquad \text{C3.12}$$

It is known that cross terms products involving different frequencies of eqn (C3.12), integrate to zero along the unit circle. Only the following terms survive:

$$e^{-j\omega k}(c_a(k) - c_b(k)).e^{j\omega k}(c_a(k) - c_b(k)).d\omega \qquad \text{C3.13}$$

Note that $c_a(k)$'s and $c_b(k)$'s are even functions of k for a minimum phase all pole spectrum and we therefore write

$$L.S.D = \frac{1}{2\pi}\int_{-\pi}^{\pi} d\omega \sum_{-\infty}^{\infty} e^{-j\omega k}[C_a(k) - C_b(k)].\, e^{j\omega k}[C_a(-k) - C_b(-k)] \qquad \text{C3.14}$$

Since $c_a(k) = c_a(-k)$ and $c_b(k) = c_b(-k)$ we can write this as

$$L.S.D = \frac{1}{2\pi}\int_{-\pi}^{\pi} d\omega \sum_{-\infty}^{\infty} [C_a(k) - C_b(k)]^2 \qquad \text{C3.15}$$

Therefore

$$= \sum_{-\infty}^{\infty} [C_a(k) - C_b(k)]^2$$

$$= 2 \sum_{k=1}^{\infty} [C_a(k) - C_b(k)]^2 + [C_a(0) - C_b(0)]^2 \qquad \text{C3 16}$$

$$L.S.D = \frac{1}{2\pi} \int_{-\pi}^{\pi} [\ln(A(\omega)) - \ln(B(\omega))]^2 d\omega$$

$$= 2. \sum_{k=1}^{\infty} [C_a(k) - c_b(k)]^2 + [C_a(0) - c_b(0)]^2 \qquad \text{C3.17}$$

This relationship (eqn. B3.17) shows that *L S.D* is exactly the same as the Euclidean cepstral distance measure for an infinite summation. However, it is not necessary to do the summation over infinite terms as the cepstrum dies down to zero after about 2.P terms. Therefore, 2*P terms are usually used in the cepstral distance measure criterion, where P is the number of LPC coefficients as defined before.

C 4.0 Characteristics of the digital Low Pass Filter used in the research

A low pass filter with 2.5 kHz cut off frequency and a gradual roll off was designed at 16 kHz sampling frequency. The 'Hyper signal' Digital Signal Processing software package was used for the design A 3 - pole IIR Bessel filter design satisfies the above requirement.

The transfer function of the filter *H(z)* is found to be·

$$H(z) = \prod_{i=1}^{2} \frac{(a_{0,i} + a_{1,i} z^{-1} + a_{2,i} z^{-2})}{(1 + b_{1,i} z^{-1} + b_{2,i} z^{-2})} \qquad \text{C4 1}$$

where the numerator coefficients are given by·

$a_{01} = 0\ 356;\ a_{11} = 0\ 356,\ a_{21} = 0\ 0;$ *and* $a_{02} = 0\ 356,\ a_{12} = 0\ 712;\ a_{22} = 0.356,$

and the denominator coefficients are given by

b_{11} = -0.07289, b_{21} = 0.0, and b_{12} = -0 0647, b_{22} = 0 161

The filter has a linear phase response. The magnitude and the phase characteristics of the filter is shown below Fig C4.

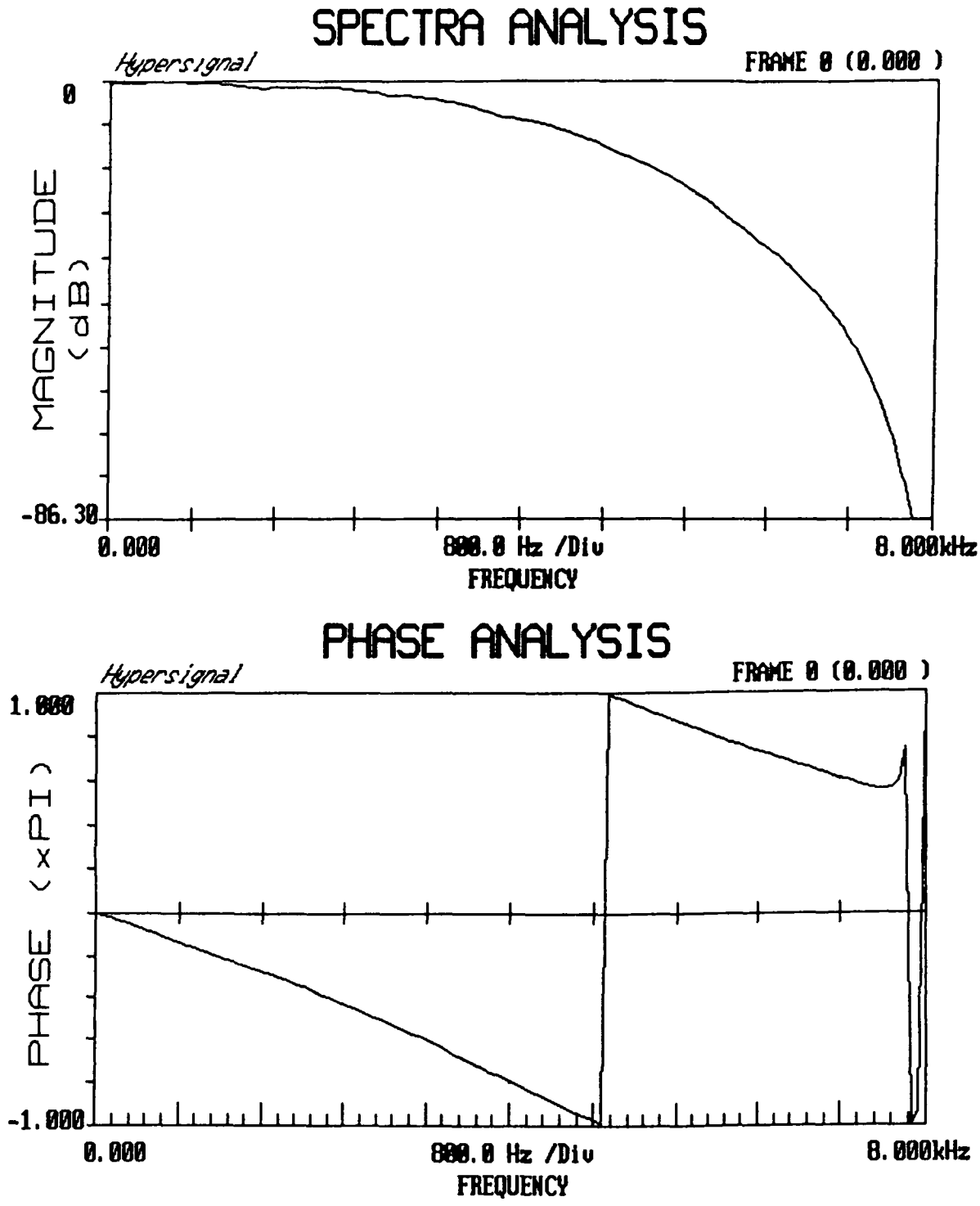


Fig. C4 Characteristics of the low pass filter

a): Magnitude response

b): Phase response

C 5.0 Distortion measures

In order to optimise the model parameters we must define a distortion measure by which model speech can be compared to the real speech. Selection of an optimum distance measure (distortion) between reference and test/model frame is essential for obtaining a good model. Several distortion measures are available for speech analysis, synthesis and recognition tasks. These are discussed below.

Let denote $X(\omega)$ be the short time spectral envelope of the original speech waveform and $S(\omega)$ be the spectral envelope of the estimated/model signal, then the distance between these two frames can be defined by various measures. The $X(\omega)$ is the Digital Fourier Transform (DFT) spectrum. $S(\omega)$ is the spectrum of the LPC model defined by

$$S(e^{j\omega}) = \frac{G}{(1 + a_1 e^{-j\omega} + a_2 e^{-j2\omega} \ldots + a_p e^{-jp\omega})} \qquad \text{C5.1}$$

where G and a_i are the gain and i^{th} LPC prediction coefficient of the p^{th} order LPC model, respectively.

C 5.1 Likelihood ratio

This distortion measure is given by [Ye and Tuffelli, 1987]:

$$D_{lr} = \int_{-\pi}^{+\pi} \left| \frac{X(e^{j\omega})}{S(e^{j\omega})} \right|^2 \frac{d\omega}{2\pi} \qquad \text{C5.2}$$

C 5.2 Itakura - Saito likelihood ratio

This distortion measure is given by [Soong and Sondhi, 1987]:

$$D_{IS} = \int_{-\pi}^{+\pi} \left[\frac{|X(e^{j\omega})|^2}{|S(e^{j\omega})|^2} + \log \frac{|X(e^{j\omega})|^2}{|S(e^{j\omega})|^2} - 1 \right] \frac{d\omega}{2\pi} \qquad \text{C5.3}$$

C 5.3 Itakura ratio (also known as gain-optimised Itakura-Saito distortion measure or log likelihood ratio distortion)

The Itakura - Saito measure is not very appropriate for comparing two given LPC

spectra in recognition system because the measure is sensitive to the LPC gain [Soong and Sondhi, 1987]. Itakura overcomes this problem by defining a new measure, given by [Itakura, 1975]:

$$d_I = \min_{\sigma_B>0} d_{IS}\left[\frac{\sigma_A^2}{|A|^2}, \frac{\sigma_B^2}{|B|^2}\right]$$

$$= \log \int_{-\pi}^{+\pi} \frac{|1+b_1 e^{-j\omega}+b_2 e^{-j2\omega} \ldots +b_p e^{-jp\omega}|^2}{|1+a_1 e^{-j\omega}+a_2 e^{-j2\omega} \ldots +a_p e^{-jp\omega}|^2} \frac{d\omega}{2\pi} \qquad \text{C5.4}$$

where $\frac{\sigma_A^2}{|A|^2}$ and $\frac{\sigma_B^2}{|B|^2}$ are LPC spectra of two given AR models.

This Itakura ratio can be weighted to give more emphasis to spectral peaks than the spectral vallys for speech recognition purposes [Soong and Sondhi, 1987]. The weighted Itakura measure is given by

$$d_{WI} = \log \int_{-\pi}^{+\pi} F(\omega) \frac{|B(\omega)|^2}{|B(\omega)|^2} \frac{d\omega}{2\pi} \qquad \text{C5.4}$$

where $A(\omega)$ and $B(\omega)$ reference and test LPC spectrum respectively. $F(\omega)$ is a weighting function.

C 5.4 Euclidean cepstral distance measure

As discussed in section C 4.3 LSD between two frames of speech is equivalent to Euclidean cepstral distance between two those frames. It is given by

$$L.S.D = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\ln(A(\omega)) - \ln(B(\omega))\right]^2 d\omega$$

$$= 2.\sum_{k=1}^{\infty}[C_a(k) - c_b(k)]^2 + [C_a(0) - c_b(0)]^2 \qquad \text{C5.5}$$

C 5.5 Weighted Euclidean cepstral distance measure

Weighted distortion measure is used to emphasise the important characteristics of the spectra. This is given by:

$$Distance = \sum_{k=1}^{\infty} W_k^2[C_a(k) - c_b(k)]^2 \qquad \text{C5.6}$$

where w_k is a predetermined weighting coefficient. Several cepstral weights have been proposed e.g Root triangular weighted cepstral measure, raised sine etc. [Juang et al., 1986].

C 6.0 Pre - Emphasis of speech signals

There is a -6 dB/octave fall off in voiced speech spectra radiated from the lips. This is due to the -12 dB/oct fall off of the excitation source (at the input) and +6 dB/oct rise the radiation at the lips (at the output). Usually, in non-interactive source-filter processing of speech vocal tract, the lip radiation at the lips and the excitation at the glottis are modelled separately. When the vocal tract (which has in a wide sense 0 dB/oct roll off i.e. flat spectra) is modelled from the speech, it is desirable to use the signal which has a flat spectrum. Therefore, -6dB/oct trend in the speech spectra is equalised by +6dB/oct lift prior to the processing. The -6 dB/oct lift can be achieved digitally by differencing the signal i.e.

$$s'(n) = s(n) - as(n-1) \qquad \text{C6.1}$$

where the constant a takes a value between 0.9 to 1.0, for voiced speech.

The pre-emphasis process can also be expressed by the transfer function of the

form

$$H(z) = 1 - az^{-1} \qquad \text{C6 2}$$

It should be mentioned that this pre-emphasis is not required for unvoiced speech

APPENDIX D

D 1.0 Rosenberg glottal flow model

The glottal flow g(n) modelled by Rosenberg, is given by the following sinusoidal segments as:

$$\begin{aligned} g(n) &= \frac{1}{2}\left[1 - \cos(\pi n/N_1)\right] && 0 \le n \le N_1 \\ &= \cos(\pi(n-N_1)/2N_1)) && N_1 \le n \le N_1+N_2 \\ &= 0 && \textit{otherwise} \end{aligned} \qquad \text{D1.1}$$

where $0 < n < N_1$ is the opening phase, $N_1 < n < N_1 + N_2$ is the closing phase and g(n) = 0 implies the closed phase

D 2.0 'L' model of the glottal flow

The four parameter model (LF model) described in section 4.1 2 was developed in two parts The first part of the LF model from 0 to t_e was described by a three parameter model of the glottal flow [Liljencrants, 1984]. This model given by eqn. 4.2 was initially used to describe the flow over the full fundamental period ($0 < t < T_0$). In this case, flow at T_0 is forced to zero and therefore the return phase which was described by an exponential curve in the LF model, was not allowed

$$\frac{dU_g(t)}{dt} = E(t) = E_0 e^{\alpha t}\sin\omega_g t \qquad 0 < t < T_0 \qquad \text{D2.1}$$

This model can be implemented with a second order digital filter with positive exponent It displays a more gradual rise than the F-model [Fant, 1979] for a given set of constraints of equal values of t_e and E_e/E_t

REFERENCES

Ananthapadmanabha, T.V (1984) "Acoustic analysis of voice source dynamics," STL - QPSR 2-3/1984, pp 1-24

Alku, P (1992) "An automatic inverse filtering method for the analysis of glottal waveforms," Thesis for the degree of Doctor of Technology, Acoustics Laboratory, Helsinki University of Technology.

Atal, B S.(1974)· "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acous. Soc. Am.* 55, pp. 1304-1312

Atal, B.S , Chang, J.J., Matthews, M.V. and Tukey, J W (1978) "Inversion of articulatory - to - acoustic transformation in the vocal tract by a computer - sorting technique," *J Acous Soc Am* 63, pp. 1535-1555.

Atal, B.S. and Hanauer, S.L (1971): "Speech analysis and synthesis by linear prediction of the speech wave," *J Acous. Soc Am* 50, pp 637-655

Badin, P. (1990)· "Area function data of 6 vowels and 9 fricatives as presented in [Fant, 1960]," Obtained by a private correspondence with Dr P Badin of ICP, Grenoble, France

Badin, P. and Fant, G (1984)· "Notes on vocal tract computation," STL - QPSR 2-3/1984, pp 53-108

Campbell, J.P., Welch, V.C and Tremain T E. (1990)· "The New 4800bps Voice Coding Standard," Proc IEEE ICASSP, pp. 64-70.

Coker, C H. (1976). "A model of articulatory dynamics and control," Proc of the IEEE, Vol 64(3), pp 452-460

Coker, C H and Fujimura, O (1966): "A model for specification of vocal tract area function, " *J Acous Soc Am 40,* pp 1271

Childers, D G and Lee, C.K (1991)· "Vocal quality factors· Analysis, synthesis and perception, " *J. Acous Soc. Am* 90(5), pp. 2394-2409

Cranen, B and Boves, L. (1988): "On the measurements of glottal flow," *J Acous Soc. Am* 84(3), pp 888-900

Cummings, K E and Clements, M.A (1992)· "Improvements to and applications of analysis of stressed speech using glottal waveforms," Proc IEEE ICASSP Vol 2 , pp 25-28

Edwards, A D.N (1991) *Speech Synthesis, Technology for Disabled People,* Paul Chapman Publishing Ltd, London

Fant, G (1970) *Acoustic Theory of Speech Production,* The Hague Mouton (1st edition, 1960).

Fant, G. (1972): "Vocal tract wall effects, losses and resonance bandwidths," STL - QPSR 2-3/1972, pp. 28-52.

Fant, G. (1990): "Speech research in perspective," Speech Communication (9), pp. 171-176.

Fant, G., Liljencrants, J. and Lin, Qiguang (1985): "Speech production," STL - QPSR 4/1985, pp. 1-13.

Flanagan, J.L., Ishizaka, K., Shipley, K.L. (1975): "Synthesis of speech from a dynamic model of the vocal cords and vocal tract," *The Bell system technical journal* 54(3), pp. 485-506

Flanagan, J.L., Ishizaka, K., Shipley, K.L. (1980): "Signal models for low bit-rate coding of speech," *J. Acous. Soc. Am. 68*(3), pp. 780-791.

Flanagan, J.L. (1972): *Speech Analysis Synthesis and Perception,* New york, Berlin: Springer Verlag (1st edition 1965).

Furui, S. (1989): *Digital Speech Processing, Synthesis and Recognition,* Marcel Dekker, INC., New york.

Ghitza, O. and Sondhi, M.M. (1993): "Hidden Markov models with templates as non-stationary states: an application to speech recognition," Computer Speech and Language (2), pp. 101-119.

Giachin, E.P., Rosenberg, A.E. and Lee, Chin-Hui (1991): "Word juncture modeling using phonological rules for HMM-based continuous speech recognition," Computer Speech and Language(5), pp. 155-168.

Gupta, S.K. and Schroeter J. (1991): "Low update rate articulatory analysis/synthesis of speech," Proc. IEEE ICASSP, pp. 481-484.

Haagen, J., Nielsen, H. and Hansen, S.D. (1991): "A 2.4 KBPS high quality speech coder," Proc. IEEE ICASSP, pp. 589-592.

Holmes, J.L. (1988): *Speech synthesis and recognition,* Van Nostrand Reinhold(UK) Co. Ltd, Berkshire, England.

Holmberg, E.B., Hillman, R.E. and Perkell, J.S. (1988): "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal and loud voice," *J. Acous. Soc. Am. 84*(2), pp. 511-529.

Ishizaka, K., French, J.C. and Flanagan, J.L. (1975): "Direct determination of vocal tract wall impedance," *IEEE transac. Vol. ASSP-35*, No.7, pp. 947-954.

Juang, B.H., Rabiner, L.R., and Wilpon, J.G. (1987): "On the use of band pass liftering in speech recognition," *IEEE Transac. Vol. ASSP-23*, pp. 370-373.

Karlsson, I. (1988) "Glottal waveform parameters for different speaker types," STL - QPSR 2-3/1988, pp. 61-67.

Kelly, J , Lochbaum, C. (1962): "Speech synthesis," G42, Proc. 4th Int Conf on Acoustics, Copenhagen

Klatt, D.H and Klatt, L C (1990). "Analysis, synthesis and perception of voice quality variations among female and male talkers," *J Acous Soc. Am 87*(2), pp. 820-855.

Krishnamurthy, A.K. and Childers, D G. (1986)· "Two - channel speech analysis," *IEEE Transac. Vol. ASSP-34*, No 4, pp. 730 - 743.

Larar, J.N., Schroeter, J. and Sondhi M M. (1988): "Vector Quantisation of the articulatory space, " *IEEE Transac Vol. ASSP-36*, No. 12, pp 1812-1818.

Laine, U (1982): "Modelling of lip radiation impedance, " Proc IEEE ICASSP, pp 1992-1995.

Lieberman, P. (1977): *Speech Physiology and Acoustic Phonetics,* Macmillan Publishing Co , Inc. New york 10022.

Liljencrants, J (1985)· "Speech synthesis with a reflection-type line analog", Tekn.D Thesis, KTH Stockholm.

Lin, Qiguang. (1990): "Speech Production Theory and Articulatory Speech Synthesis", Doctoral Thesis, Royal Institute of Technology, Stockholm.

Lobo, A P. and Ainsworth, W A. (1989) "Evaluation of a glottal ARMA modelling scheme," Proc. European Conf. on Speech Communication and Technology, Vol. 2, pp 27-30, EUROspeech 89, Paris, Sept , 1989.

Maeda, S. (1979): "An articulatory model of the tongue based on a statistical analysis," *J.Acous. Soc Am* 65, S1, p. 22(A).

Merhav, N. and Epraim, Y. (1991): "Hidden Markov modeling using a dominant state sequence with application to speech recognition," Computer Speech and Language (5), pp 327-339

Mermelstein, P (1973): "Articulatory model for the study of speech production," *J Acous. Soc Am.* 53(4), pp. 1070-1082

Meyer, P., Schroeter, J and Sondhi M M. (1991). "Design and evaluation of optimal cepstral lifters for accessing articulatory codebooks, " *IEEE Transac. Vol. ASSP-39,* No 7, pp 1493-1502.

Muller, J M(1990). "Improving performance of Code Excited LPC -coders by joint optimization," Speech Communication 8 (1990), pp. 363-369.

Papamichalis, P.E (1987): *Practical approaches to speech coding,* Prentice-Hall, inc , Englewood Cliffs, New Jersey.

Parthasarathy, S. and Coker, C.H. (1992): "On automatic estimation of articulatory parameters in a text-to-speech system," Computer Speech and Language (2), pp. 37-76.

Prado, P.P L, Shiva, E H and Childers, D G (1992): "Optimisation of acoustic to articulatory mapping," Proc IEEE ICASSP Vol 2., pp 33-36

Press, W.H., Flannery, B.P., Teukolsky, S.A and Vetterling. W.T. (1988). *Numerical Recipes,* C.U.P Cambridge.

Price, P J (1989): "Male and female voice source characteristics: Inverse filtering results," Speech Communication 8, pp. 261-277.

Quarmby, D.J. and Holmes, J N (1984) "Implementation of a parallel-formant speech synthesiser using a single-chip programmable signal processor," IEE Proceedings, Vol. 131, Pt.F, No. 6, Oct 1984.

Rabiner, L R. and Schafer, R W (1978). *Digital Signal Processing of Speech Signals,* Prentice-Hall, inc., Englewood Cliffs, New Jersey 07632.

Rosenberg, A.E. (1971): "Effect of glottal pulse shape on the quality of natural vowels," *J Acous. Soc Am 49(2),* pp 583-590

Scaife, R. (1989): "Voacl tract area estimation - Extending the Wakita inverse filter," Proc of EUROSPEECH, Vol 2, pp. 648-651.

Scaife, R. and Kanagaratnam, J. (1990): "Articulatory speech synthesis Estimation of data from speech," Colloquium on Digital Signal Processing and Control, The Queen's University of Belfast, U.K., pp. 19/1-19/7

Schroeder, M.R. and Atal, B S. (1985): "Code Excited Linear Prediction (CELP): High quality speech at very low bit rates," Proc IEEE ICASSP, pp. 937-940.

Schroeter, J., Larar, J N and Sondhi, M.M (1987): "Speech parameter estimation using a vocal tract/cord model," Proc IEEE ICASSP, pp 308-311.

Schroeter, J., Meyer, P., Parthasarathy, S. (1990): "Evaluation of improved articulatory codebooks and codebook access distance measures," Proc. IEEE ICASSP, pp 393-396

Schroeter, J , Sondhi, M.M (1989) "Dynamic programming search of articulatory codebooks," Proc IEEE ICASSP, pp. 588-591.2

Shirai, K. and Honda, M. (1977) "Estimation of articulatory motion," U S. - Japan Joint Seminar on Dynamic aspects of speech production, Tokyo University Press, pp.279-302

Shirai, K. and Kobayashi, T. (1986)· "Estimating articulatory motions from speech wave," Speech Communication 5, pp 159-170

Soong, F.K and Sondhi, M M (1987). "A frequency-weighted Itakura spectral distortion measure and its application to speech recognition in noise," Proc. IEEE ICASSP, pp. 625-628

Sondhi, M.M (1975): "Measurement of the glottal flow," *J. Acous Soc. Am. 57(1)*, pp. 228-232.

Sondhi, M M (1979): "Estimation of vocal tract areas: The need for acoustical measurements," *IEEE Transac Vol ASSP-27,* No 3, pp. 268-273

Sondhi, M.M and Resnick, J R (1983): "The inverse problem for the vocal tract· Numerical methods, Acoustical experiments, and speech synthesis," *J. Acous. Soc. Am 73(3)*, pp 985-1003.

Sorokin, V.N. (1992) "Determination of vocal tract shape for vowels," Speech Communication 11, pp. 71-85.

Stevens, K N., Kasowki, S and Fant, G. (1953) "An electrical analog of the vocal tract," *J Acous Soc Am 25*, pp 734-742.

Stevens, K N and House, A.S (1955). "Development of a quantitative description of vowel articulation," *J. Acous. Soc. Am. 27*, pp 484-493.

Veeneman, D.E and BeMent, S.L (1985). "Automatic inverse filtering from speech and electro glotto graphic signals," *IEEE Transac. Vol. ASSP-33,* No 2, pp 369-376.

Wakita, H. (1973). "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans AU, Vol 21*(5), pp 417-427

Wakita, H. and Fant, G. (1978). "Toward a better vocal tract model," STL - QPSR 1/1978, pp 9-29

Wilpon, J.G. and Rabiner, L.R. (1985)· "A modified K-means clustering algorithm for use in isolated word recognition," *IEEE Transac Vol. ASSP-33*, pp. 587-594

Witten, I.H. (1982): *Principles of computer speech,* Academic press inc. (London) Ltd, London

Wong, D Y , Marcel, J D. and Gray, A.H (1979) "Least squares glottal inverse filtering from the acoustic speech waveform,"*IEEE Transac Vol ASSP-27*, No 4, pp. 350-355

Wong, D.Y., Juang, Biing-Hwang and Gray, A.H (1982): "An 800 bit/s vector quantization LPC vocoder," *IEEE Transac. Vol. ASSP-30*, No 5, pp. 770-780.

Ye, H. and Tuffelli, D. (1987). "Deterministic characteristics of LPC distances. An inconsistency with perceptual evidence," Proc IEEE ICASSP, pp. 1-42.