A MODIFIED BLOCK ADAPTIVE PREDICTIVE CODER FOR SPEECH PROCESSING

A THESIS SUBMITTED BY
TESSAMMA THOMAS
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
UNDER THE FACULTY OF TECHNOLOGY

DEPARTMENT OF ELECTRONICS
COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY
KOCHI - 682 022, INDIA.

1993

CERTIFICATE

This is to certify that the thesis entitled "**A MODIFIED BLOCK ADAPTIVE PREDICTIVE CODER FOR SPEECH PROCESSING**", is a bona fide record of the original work carried out by **Ms. TESSAMMA THOMAS** under my supervision in the Department of Electronics, Cochin University of Science and Technology. The results embodied in this thesis or part of it have not been presented for any other degree.

Dr. C. S. Sridhar
(Supervising Teacher)
Professor
Department of Electronics
Cochin University of Science
and Technology

Kochi 682022

August 24, 1993

DECLARATION

I hereby declare that the work presented in this thesis entitled "**A MODIFIED BLOCK ADAPTIVE PREDICTIVE CODER FOR SPEECH PROCESSING**" is a bona fide record of the original work done by me under the supervision of Dr.C.S. Sridhar in the Department of Electronics, Cochin University of Science and Technology, and no part thereof has been presented for any other degree.

Tessamma Thomas

Tessamma Thomas

Kochi 682022

August 24, 1993

ACKNOWLEDGEMENT

I express my deep gratitude to my guide Prof. C.S. Sridhar for his invaluable guidance and support throughout this work.

I am grateful to Prof. K.G. Nair, Head of the Department of Electronics, for his timely help and inspiration.

I wish to thank Dr. Babu P. Anto, Lecturer, NERIST, Arunachal Pradesh and Dr.N.K. Narayanan, Lecturer, Govt. Brennen College, Tellicherry, for their whole-hearted co-operation and effective and fruitful discussions throughout my work.

I am thankful to all the members of the Faculty, Laboratory and non-teaching staff of the Department of Electronics for their whole-hearted co-operation and help during the course of this work.

I also wish to thank Dr.C.K. Aanandan and few of the research scholars and students, who have helped me during the final stage of my work.

Sincere help and co-operation of Mr.K.P. Sasidharan in bringing this thesis in its present form are also acknowledged.

Tessamma Thomas.

ABSTRACT

This thesis investigated the potential use of Linear Predictive Coding in speech communication applications. A Modified Block Adaptive Predictive Coder is developed, which reduces the computational burden and complexity without sacrificing the speech quality, as compared to the conventional adaptive predictive coding (APC) system. For this, changes in the evaluation methods have been evolved. This method is as different from the usual APC system in that the difference between the true and the predicted value is not transmitted. This allows the replacement of the high order predictor in the transmitter section of a predictive coding system, by a simple delay unit, which makes the transmitter quite simple. Also, the block length used in the processing of the speech signal is adjusted relative to the pitch period of the signal being processed rather than choosing a constant length as hitherto done by other researchers. The efficiency of the newly proposed coder has been supported with results of computer simulation using real speech data.

Three methods for voiced/unvoiced/silent/transition classification have been presented. The first one is based on energy, zerocrossing rate and the periodicity of the waveform. The second method uses normalised correlation coefficient as the main parameter, while the third method utilizes a pitch-dependent correlation factor. The third algorithm which gives the minimum error probability has been chosen in a later chapter to design the modified coder.

The thesis also presents a comparative study between the autocorrelation and the covariance methods used in the evaluation of the predictor parameters. It has been proved that the autocorrelation method is superior to the covariance method with respect to the filter stability and also in an SNR sense, though the increase in gain is only small. The Modified Block Adaptive Coder applies a switching from pitch prediction to spectrum prediction when the speech segment changes from a voiced or transition region to an unvoiced region. The experiments conducted in coding, transmission and simulation, used speech samples from Malayalam and English phrases. Proposal for a speaker recognition system and a phoneme identification system has also been outlined towards the end of the thesis.

CONTENTS

Chapter 1

INTRODUCTION

1.1 Background

Speech is the principal means of human communication. "Speech is civilisation itself. The word, even the most contradictious word, preserves contact - it is silence which isolates" - Thomas Mann, The Magic Mountain, 1924 [1]. It is a fascinating human attribute, which can be analysed, synthesised and recognized. It can be compressed, stored and also enhanced by digital signal processing techniques.

Right in 1780, Professor Christian Kratzentein had designed a vox humana , capable of producing the vowel sounds, from a set of tubes of different shapes [1]. In the later years, coming to the actual representation, storage and transmission of speech data, it was understood that digital processing techniques are superior to their analogue counterparts. Digital signals are less sensitive to noise and can reliably be transmitted over noisy channels. They are easy to store and regenerate, to encrypt and error-protect, and to multiplex, mix and packetize [2]. Starting from the earliest electronic speech synthesizer, the 'Vocoder' of Homer Dudley in 1939, we have a myriad type of digital speech communication systems. Digital speech technology has already been applied to telephony systems which can access the computer data bases and receive relevant information from stored or synthetic speech [1]. The talking calculator, reading machines for the blind, etc. are some other applications of the computer voice-response systems [3,4,5,6]. A more flexible method is the interactive speech recognition, where Artificial Intelligence is used for Automatic Speech Recognition and Speech Response (synthesis). Now-a-days, Artificial Neural Networks are being used to improve speech recognition and understanding. Digital services such as the transmission of data, fascimile, vision and videotex are beginning to proliferate on a global basis, with the help of a variety of optical cables, satellites and radio systems [1].

In digital speech processing, the speech signal is sampled, quantized and then processed by appropriate processing techniques, depending on the nature of applications. Finally, the speech signal is reconstructed from its digital form using a decoder.

Digital speech coding algorithms can be broadly classified into three basic categories: waveform coding, vocoding (parametric coding) and hybrid coding. Waveform encoders follow the original signal waveform faithfully, by directly digitizing the speech signals. They require bit rates in the range 16 kbit/s to 64 kbit/s.

In parametric coding, the speech signal is represented as the output of a speech production model, which is excited by a source. The coder analyses the input speech samples and estimates the vocal tract parameters, which are related to the individual speech sounds, and the excitation parameters, which are related to their source. These parameters are transmitted to the decoder which retrieves them and synthesizes the speech signal. Vocoders operate at very low bit rates (less than 4.8 kbit/s) but gives only synthetic quality speech and hence is inadequate for general purpose voice communication.

Hybrid coders combine the features of waveform and parametric coders and operate at bit rates between the two. Many hybrid coders employ an analysis--by-synthesis procedure, to derive the codec parameters.

Based on the above techniques, we have the simplest and the oldest method, the Pulse Code Modulation (PCM) operating at 56-64 kbit/s, to the Adaptive Differential Pulse Code Modulation (ADPCM) giving good quality speech at 24-28 kbit/s, and the Adaptive Predictive Coding (APC) and the Adaptive Transform Coding (ATC) techniques operating at 16 to 32 kbit/s to give good quality speech [7]. Some complex ATC coders operate at bit rates of the order of 4-8 kbits/s.

1.2 Motivation

Encoding of the speech signal and recognition of the encoded signal are two important aspects of speech processing. In speech encoding, we develop a suitable method to encode and transmit a signal, while in speech recognition we try to identify the synthesized speech obtained from the encoded parameters.

For efficient coding and transmission of speech signals, the channel capacity required to transmit a given signal, with a given fidelity should be minimised. When low bit-rate codecs are considered, they become highly complex, and also complex algorithms requiring heavy computation, will become essential, if the quality of the reconstructed signal is to be maintained high. The low bit-rate vocoder outputs are usually poor in quality, though they are intelligible. Modern processing techniques have improved the quality of the reconstructed speech to a better level of acceptability. When these coders are used in practical applications, a compromise has to be achieved between quality, robustness, delay, complexity and low bit rate. Adaptive Predictive Coding (APC) analysis is a potential area in the low bit rate coding, which can deliver good quality speech.

In the APC method, using both spectrum and pitch predictors, an estimate of the samples is made at the encoder and the residual is transmitted to the decoder along with the predictor parameters. In the modern version of the APC - example, Code Excited LPC (CELPC), regular pulse excited LPC (RPELPC), multipulse LPC (MPLPC), self excited vocoder (SEV), etc. - a code index of the best suited residual vector, or the appropriate pulse positions and amplitudes need only be transmitted. But the amount of computations needed is enormous. It is reported that [8], in a

CELP system, the number of multiples/adds is 80 MFLOPs, while in SEV, it is 4 MFLOPs. In SEV, the decoder is more complex with the introduction of an additional pitch predictor. The maximum segmental signal-to-noise ratio (SNRSEG) for a CELPC is around 13.8 dB (an SNR of 18 dB) at a bit rate of 2.8 kbps and with quality almost same as the original [9]. An SNRSEG value of about 10 dB is obtained at 4.8 kbps for an SEV [8], while an SNR of 14.5 dB to 18.5 dB is obtained at 9.6 kbps to 16 kbps using a variable bit rate APC [10].

The present modified coder is an attempt to obtain an almost good quality speech reconstruction, with a reduction in complexity and computational burden, and increase in data transmission, as compared to the above mentioned systems. In the modified block adaptive coder (MBAC), the residual signal is not transmitted at all. Only a coded version of the first few sample values is transmitted along with the predictor parameters. Hence not much computation is needed other than that required for evaluating the predictor parameters. Also, the high-order predictor can be removed from the transmitter side, which in turn reduces the coder complexity. To reduce the computational burden still further (as compared to an APC), changes in the evaluation methods are introduced. The SNRSEG value obtainable is around 10 to 12 dB, at a transmission rate of 6.2 kbps to 11 kbps.

1.3 Brief Overview of the Work

A brief review of the previous works in the field of speech recognition, for the efficient transmission of speech signals, is presented in chapter 2. Special stress is given to the Linear and Adaptive Predictive Coding techniques. The objective as well as the subjective tests that are useful for the determination of the coder performance are also included in this chapter. Of those, the SNRSEG measurement is used in this work, as the criterion of the coder performance.

Chapter 3 describes the basic theory of linear predictive coding. The basis and the need for considering the long-term prediction, is also explained. In Linear Predictive Coding [11,12,13,14], the vocal tract is represented by a linear time-varying digital filter. A set of parameters is extracted from the speech signal, to specify the filter transfer function, which will give the best match to the signal to be coded. The minimum mean square error criterion is used. The modified coder is presented in the latter part of the chapter.

Chapter 4 presents the details of the speech data used for the simulation study and also the actual simulation work of the modified coder. Voiced/unvoiced/silence/transition classification is to be done first. Three methods have been developed. In the first method, parameters like zero crossing rate (ZCR), energy, periodicity of autocorrelation functions (ACF), and S_{rms} to S_{mean} ratio, are to be evaluated on a short-term basis. All possible cases of overlapping between the different regions have been taken into account in the method. The approach has shown to give a high recognition score. In the second method, ZCR, energy and energy–ZCR product are used to first eliminate the silent region. Then, based on the value of the normalised correlation coefficient for lag 1, voiced/unvoiced/transition classification can be done. In the third method, the detection of the silent region is as in the above method. Periodicity of the ACF's and a factor β , determining the correlation between the samples, from one pitch period to the next, are made use of in the final classification. For the above detection process, blocks of 160 normalised samples are used.

The actual simulation process of the modified coder is presented next. Once the block is detected to be voiced, the value of M, the number of samples in one pitch period of the signal, is found by noting the position of the maximum value of the ACF, R(J) for lags J above 15. The block length is then fixed as N = 4M and the predictor parameters are evaluated. These predictor parameters and the first few sample values are suitably encoded and transmitted to the receiver. Prediction is done at the receiver, based on the earlier predicted values. The predictor is updated by transmitting the parameters afresh every block. For transition segments N = 2M. For unvoiced samples prediction is based on the nearby samples only. No processing is needed for the silent region.

The actual simulation results of the modified block adaptive coder is presented in chapter 5. An average SNRSEG value of 8 to 12 dB is obtained on the whole. The results of the trial on the applicability of this modified coder to phonemes in our regional language - Malayalam - which are not normally present in English, are also presented. The sound

/ṇ/ – ‘ENN’, with the nasal /n/ following a vowel gives the highest gain, around 30 dB.

A close examination of the results arrived at the different stages of the simulation work of the modified coder, revealed that phoneme identification as well as speaker identification are possible, using certain sets of the coder parameters. An adaptive knowledge-based speaker recognition system and an adaptive phoneme identification method are developed. The work done in this direction is presented in chapter 6.

Chapter 7 is the concluding chapter, wherein, the observations and the inferences already brought out in the earlier chapters are summarised. An attempt to reduce the complexity of the coder and the computational burden, has been attained. A good quality speech reconstruction has been attained at a transmission rate of 11.9 kbit/s. The suggestions for further work are also presented.

Chapter 2

REVIEW OF THE PAST WORKS IN THE FIELD

2.1 Introduction

The pursuit of artificial speech synthesis and recognition began long back in the seventeen eighties [1]. But the first electronic synthesis of speech was achieved only in 1936. The earliest attempt at voice recognition was the voice-operated phonographic alphabet writing machine brought forth by J. Flowers, in 1916 [1]. Attempts in the direction of electronic speech recognition took place in 1950's, which were mainly concentrated on the recognition of spoken digits [1]. An accuracy of nearly 95% is recorded from these early systems [15]. These devices used analogue circuits to perform the spectral analysis by filtering and decision logic, to make the pattern match between the uttered word and the reference word.

During the decade of 1960-70, it became practicable to represent information--bearing waveforms digitally, and to do signal processing on these digital representations. In April 1965, Cooley and Tukey put forth an algorithm for computing the Discrete Fourier Transform, which gave a tremendous impetus to this emerging field [16].

2.2 Review of Speech Waveform Coders

In this section, the results obtained by different researchers, in the field of waveform coding is presented. Different persons use different speech samples and different criteria for judging the performance of their coders.

The qualitative efficiency of an encoder is characterized by the terms- toll, communication and synthetic quality [7]. Toll quality is the quality comparable to that of an analogue speech signal, band-limited to 200-3200 Hz and having a signal-to-noise ratio of 30 dB and less than 2.3% harmonic distortion [7]. It can also be considered as the speech quality as in a long distance telephone call on the analogue PSTN [1]. CCITT recommendations currently include 64 kbit/s PCM and 32 kbit/s ADPCM as standards for comparison [1]. The term, communication quality [7], is used to connote detectable distortion, but its degradation in intelligibility is only very little (from toll quality). Synthetic quality, is the lowest in the hierarchy of speech quality levels, where there will be a substantial loss of naturalness and robustness with respect to speakers and speaking environments.

Some of the objective and subjective tests used in measuring the performance of a coder, is presented in section 2.4.

A comparative study of DPCM-AQB and log-PCM speech coders, over a wide range of bits/sample is presented by Cumminskey, Jayant and Flanagan [17]. A 6-bit log-PCM system, though superior to a 4-bit ADPCM in a signal-to-noise ratio (SNR) sense, is ranked inferior from a subjective point of view. Similarly a 3-bit ADPCM system is superior to a 5-bit log PCM under subjective ranking, but is inferior in an SNR sense. Since, values of

SNR and segmental SNR (SNRSEG) [2] (explained in section 2.4) are closer in DPCM, than in log PCM (SNRSEG values are 4 to 6 dB less than SNR values), SNRSEG tends to be a better indicator than SNR for quality measurement [2]. Thus, low–complexity DPCM can provide communication quality speech at 24 to 32 kb/s.

In the paper on speech waveform coding, Jayant [18] has studied ADPCM, log-PCM and ADM coders. ADPCM at a bit rate of 16 kbit/s gives the best performance with an SNR around 11 dB. In a later paper [19], Jayant compares a regular ADPCM using a three tap fixed predictor, with a pitch–adaptive DPCM. Both coders used adaptive quantization. The average SNR value of four utterances from two male and two female speakers were 11.5 dB for ADPCM at a transmission rate of 16 kbit/s and 15.25 dB for pitch–adaptive DPCM at a rate of 17 kbit/s.

Noll [20] has performed a good comparative study of the various waveform encoding schemes, like log PCM, Adaptive PCM, and DPCM with various combinations of fixed/adaptive predictor, and fixed/adaptive quantizer. It is reported [20] that a DPCM system using fixed prediction and adaptive quantisation displayed a sharp increase in SNR (of the order of 5 to 10 dB) over the PCM and DPCM systems. Using a 12th order adaptive predictor and an adaptive quantizer, the SNR of an ADPCM system was around 17 dB at 16 kb/s, 27 dB at 32 kb/s and 35 dB at 40 kb/s.

An ADPCM system in which the predictor is updated at each sample instant, using gradient techniques, is presented by C.S.Xydeas et al [21]. The system, though very complex, performed better than a standard ADPCM system, with an SNRSEG [2], of 13.5 dB at a rate of 2 bits/sample.

A relative performance study of Sub-Band Coding (SBC) of speech, with standard ADPCM at 16 kb/s and ADM at 9.6 kb/s has been presented by Crochiere et al [22]. At 16 kb/s, though the SNR for SBC (11.1 dB) and ADPCM (10.9 dB) was comparable, 94% of the listeners preferred SBC to ADPCM. As compared to the 8.2 dB attained by ADM at 10.3 kb/s, the SBC achieved an SNR of 9.9 dB at 9.6 kb/s. Using Quadrature-Mirror Filter (QMF) partitioning principle, SBC encoders attained MOS scores of 4.3, 3.9 and 3.1 (on a scale of 0 to 5), at bit rates of 32, 24 and 16 kb/s [23]. Supplementing an SBC with a fourth order spectrum predictor and pitch predictor, the MOS score goes to 3.5 at 16 kb/s and 4.0 at 24 kb/s [24].

Crochiere et al in their paper on Tandem connections of Wideband and Narrowband speech communication systems [25], have presented the performance of a tandem connection of a conventional Linear Predictive Coder (LPC) operating at 2.4 kb/s and a Continuously Variable Slope Delta modulator (CVSD) operating at a bit rate of 16 kb/s. The LPC synthesized speech (using a predictor of order 12), is the input to the CVSD system. It is reported [25] that when rectangular (broadened) LPC excitation source was used, an improvement of 1-2 dB in the SNR value was observed over the usual CVSD systcm, in the slope overload region. But the subjective quality of speech did not improve. It is also notod that the performance of a wideband--to-narrowband link is worse than a 2.4 kb/s LPC synthesizer [26].

Crochiere et al [22, 27], have designed low bit rate sub-band speech coders, where the selection of the sub-bands (from 200 to 3200 Hz) was done based on the perceptual data contained in the articulation index (AI). The quality of the speech produced at 7.2 kb/s was nearly same as that of an 18 kb/s ADM speech [27]. Later, they have presented a variable-band coding scheme [28], where the centre-frequency of the upper two bands is varied in accordance with the dynamic movement of the resonances F2 and F3 of the vocal tract. It is reported that the quality of the 7.2 kb/s fixed--band coder is only slightly better than a 4.8 kbps variable--band coder.

Amano et al [29] have designed a TC-MQ (Time Domain Compression ADPCM-MQ) speech codec, at 8 kbps, using time domain compression on an ADPCM with a multi-quantizer. The SNRSEG obtained for short Japanese sentences were 13--16 dB. The MOS-score was 2.68, on a scale of 0 to 4, as compared to the score of 2.93 of a 16 kbps ADPCM-MQ.

Zelinski and Noll [30] compared the performance of a TC-log (Transform Coding) speech system with a TC-AQF and ATC-AQF systems. An increase in the SNR value, of about 4 dB was obtained for a TC-AQF system (22.7 dB) and about 7.5 dB for an ATC-AQF (25.8 dB) system, over the TC--log (18.3 dB) system. Comparing with the DPCM--AQF--APF system with an

SNR of 22.5 dB, TC--AQF has a very close SNR value, while ATC--AQF was 3 dB better [2]. Subjectively, the ATC shows a MOS of 4.1, on a scale of 1 to 5, at a bit rate of 24 kbps, 3.8 at 16 kbps and 2.4 at 9.6 kbps [31]. The SNRSEG value achieved by an ATC varied from 12.3 dB to 14.9 dB for various speakers [32].

Tree and Trellis encoding of speech was studied by Anderson and Bodie [33]. They considered a search algorithm with fixed number of paths at each level, throughout the code tree. The performance obtained on a speech signal of 2 sec duration was 21 dB at 2 bits/sample and 12 dB at 1 bit/sample. Stewart et al [34] have reported that their Tree and Trellis speech coders obtained an SNR of 13.5 dB "outside" and 16 dB "inside" the training sequence, at a bit rate of 2 bits/sample and 8.7 dB "outside" and 12.2 dB "inside" when transmission rate of 1 bit/sample is considered. (The term "inside the training sequence" is used when the data considered are the same as that used for the encoder design and "outside the training sequence", otherwise).

Fehn and Noll [35] studied the performance of different tree and trellis encoding schemes at 1 bit/sample and got on an average an SNRSEG value of 12 dB.

Marcellin et al [36] have investigated the effect of Trellis coded quantisation (TCQ) on a predictive speech coder, using different combinations of fixed/adaptive prediction and fixed/adaptive residual encoding. It is noted that for a 16 kbps fully adaptive speech coder, the SNRSEG obtained was in the range of 17.5 to 20.2 dB. The SNRSEG gain achieved by a Predictive TCQ system over the scalar DPCM was 1.24 dB to 2.69 dB for the 4 state trellis and 2.86 to 4.69 dB for the 256 state trellis [36]. Gibson and Haschke [37] studied the performance of eight fully adaptive DPCM-based code generators, at 16 kbps, using exhaustive searching. It is reported that the maximum SNR obtained was 15 to 18 dB when the searching depth was 5.

Vector quantisation of speech, using search algorithms, have been presented by several other researchers also [38,39,40,41,42]. They have obtained an SNR value around 13.5 dB and 12.7 dB for "inside" and "outside" the training sequence at 2 bits/sample and 9.7 dB and 8.8 dB respectively, at 1 bit/sample.

Recently, Chouly and Sari [77] have outlined a family of six-dimensional(6-D) trellis coded modulation (TCM) schemes, which involves a 2-step partitioning of the constituent QAM signal alphabet. With infinite constellations without shaping, the asymptotic gain is 3 dB for the 2-state code, 4 dB for the 4 and 8-state codes, and 5 dB for the 16 and 32-state codes which involve a smaller alphabet expansion.

A comparative study of the different waveform coders described above, is given in Table 2.1.

Considering the overall performance of the waveform coders

Table 2.1 Comparative Study of the Various Waveform Coders

	Authors	Systems and their features	Performance Summary	Bit Rate
	(1)	(2)	(3)	(4)
1	Cummiskey et al (1973) [17]	Log-PCM and DPCM-AQB speech coders	A 6-bit log-PCM system is superior to a 4-bit ADPCM in an SNR sense, but is inferior from a subjective point of view. Similarly, 3-bit ADPCM is superior to 5-bit log-PCM under subjective ranking, though inferior in an SNR sense.	48 to 24 kbps
2	Jayant 1974, {18]	ADPCM, log-PCM and ADM systems	ADPCM gives the best performance with an SNR of 11 dB	16 kbps
3	Jayant 1977, [19]	An ADPCM using fixed 3-tap predictor with a pitch-adaptive DPCM	SNR value is 11.5 dB for ADPCM and 15.25 dB for pitch adaptive DPCM	16 kbps for ADPCM and 17 kbps for pitch adaptive DPCM
4	Noll 1975, [20]	Log PCM, APCM and DPCM with fixed/adaptive predictor and fixed/adaptive quantizer	DPCM system with fixed prediction and adaptive quantisation was the best. Using a 12th order adaptive predictor and adaptive quantiser, SNR value of a DPCM system varied from 17 dB to 35 dB.	16 to 40 kbps

	(1)	(2)	(3)	(4)
5	Xydeas et al 1982, [21]	ADPCM in which the predictor is updated at each sample instant using gradient technique	Better than a standard ADPCM· SNRSEG value is 13.5 dB	16 kbps
6	Crochiere et al 1975, [22]	Compares sub-band coder (SBC) with standard ADPCM at 16 kbps and ADM at 9.6 kbps.	SNR Value of SBC and ADPCM are comparable, but SBC is preferred to ADPCM subjectively. SNR value of ADM is slightly less than that of SBC.	16 kbps
7	Daumer 1982, [23]	SBC using quadrature-mirror filter (QMF) partitioning principle.	MOS score between 3.1 and 4.3, on a scale of 0 to 5.	16 to 32 kbps
8	Honda et al 1982, [24]	SBC using a 4th order spectrum predictor and pitch predictor	MOS score between 3.5 to 4.0	16 to 24 kbps
9	Crochiere et al 1977, [25]	Tandem connection of a narrow-band coder (LPC at 2.4 kbps) and wide band coder (CVSD at 16 kbps)	Rectangular LPC excitation shows an improvement of 1-2 dB in SNR over the usual CVSD; but subjectively quality of speech did not improve.	
10	Rabiner et al 1977, [26]	Tandem connection of a wide-band to-narrow-band link	Performance was worse than a 2.4 kbps LPC synthesizer.	
11	Crochiere 1977, [27]	SBC-Selection of sub-bands based on the perceptual data contained in the articulation index	Quality of the SBC speech at 7.2 kbps is nearly same as that of an 18 kbps ADM speech	7.2 kbps

	(1)	(2)	(3)	(4)
12	Crochiere & Sambur 1977, [28]	Variable-band coding- centre frequency of the upper two bands is varied in accordance with the dynamic movement of the resonances F2 and F3 of the vocal tract	Quality of 7.2 kbps fixed-band coder is only slightly better than a 4.8 kbps variable-band coder	4.8 kbps
13	Amano et al 1988, [29]	ADPCM-MQ- Using Time domain compression on ADPCM with multi-quantizer	SNRSEG value is 13-16 dB. MOS score is 2.68, on a scale of 0 to 4 as compared to 2.93 of a 16 kbps ADPCM-MQ.	8 kbps
14	Zelinski & Noll 1977, [30]	Compared Transform Coding (TC)-log system with a TC-AQF and ATC-AQF	SNR value for TC-log is 18.3 dB while that for ATC-AQF is 25.8 dB and TC-AQF system is 22.7 dB.	
15	Jayant & Noll 1984, [2]	Compared DPCM-AQF-APF with TC-AQF and ATC-AQF.	DPCM-AQF-APF and TC-AQF have SNR value around 22.5 dB, while ATC-AQF is 3 dB better.	
16	Netravali & Limb 1980, [31]	ATC system at various bit rates	ATC shows a MOS of 4.1, on a scale of 1 to 5, at 24 kbps, 3.8 at 16 kbps and 2.4 at 9.6 kbps	24 to 9.6 kbps
17	Zelinski & Noll 1979, [32]	ATC system for various speakers	SNRSEG varies from 12.3 dB to to 14.9 dB.	
18	Anderson & Bodie 1975, [33]	Tree and Trellis encoder with a search algorithm having fixed number of paths at each level.	SNR of 21 dB at 2 bits/sample and 12 dB at 1 bit/sample	8 to 16 kbps
19.	Stewart et al 1982, [34]	Tree and Trellis speech coders at different bit rates	SNR of 13.5 dB to 16 dB at 2 bits/sample and 8.7 dB to 12.2 dB at 1 bit/sample.	8 to 16 kbps

	(1)	(2)	(3)	(4)
20	Fehn & Noll 1982, [35]	Different Tree and Trellis encoding schemes	Average SNRSEG of 12 dB	8 kbps
21	Marcellin et al 1988, [36]	Studied effect of Trellis Coded Quantisation (TCQ) on a predictive coder, using different combinations of fixed/adaptive prediction and fixed/adaptive residual encoding.	For a fully adaptive speech coder, SNRSEG value is 17.5 dB to 20.2 dB	16 kbps
22	Gibbson & Haschke 1988, [37]	Fully adaptive DPCM based coders using exhaustive searching	Maximum SNR is 15 dB to 18 dB, at a searching depth of 5.	
23	Chouly & Sari 1992, [77]	6-D trellis coded modulation (TCM) schemes, using 2-step partitioning of the constituent QAM signal alphabet	Asymptotic gain is 3 dB for 2-state code and 4 dB for 4 and 8-state codes and 5 dB for 16 and 32-state codes.	

presented above, it can be noted that the maximum signal-to-noise ratio attainable by an ADPCM system is about 12 to 17 dB at 16 kbps. Highly complex schemes like ATC can achieve an SNR value of 21 to 25 dB at 2 bits/sample and 10-12 dB at 1 bit/sample.

2.3 Review on the Present State of Art of Adaptive Predictive Coders

This section discusses the current state of art of Adaptive Predictive coders. They are low-bit rate, highly complex, hybrid coders, which are robust to background noise [1, 43]. They usually work at a transmission rate of 9.6 to 2.4 kbps.

The fact that adjacent speech samples have high correlation among themselves have induced researchers to apply Linear Predictive analysis techniques to speech processing. Atal and Hanauer [14] performed a listening test on the speech reconstructed using predictors of order p between 2 and 18. It was noted that there was no significant differences in the quality of speech for p above 12. Here, the analysis segment length was one pitch period for voiced sections and 10 msec for unvoiced regions. A bit rate of the order of 7.2 to 2.4 kbps is achieved.

In adaptive predictive coding (APC) [1,8,44,45], pitch prediction and noise shaping are included and the residual samples are quantized and transmitted. A bit rate greater than 16 kbps is needed for good quality speech. Atal and his colleagues investigated an APC [44] system using an eighth-order adaptive spectrum predictor and first order pitch predictor and a one-bit adaptive quantizer. Subjective testing showed that the speech reconstructed using their system was superior to a log-PCM speech with 5 bits/sample. The equivalent gain of the system was around 25 dB. A bit rate of around 10 kb/s was needed for transmitting the binary difference signal and the predictor parameters, when the sampling was done at 6.67 kb/s [44]. Later, they studied the performance of an APC system with noise feedback coding (APC-NF) [2,45,46]. Using an entropy-coded 3-level quantizer, the system gave an average SNR of 21 dB [2]. This is higher than the SNR of 13 dB of an open-loop DPCM [47] or D^*PCM coder [2,48], but lower than the SNR of 23 dB in DPCM, with all the three coders operating at 19.2 kb/s. According to Daumer [23], a 16 kb/s APC-NF system provided a MOS score of 4.0.

A variable rate embedded-code ADPCM system, using Time Domain Harmonic Scaling (TDHS) algorithms is investigated by Copperi [49]. The TDHS algorithms change the speech rate by discarding or repeating short pieces of the waveform, having a length at least equal to a pitch period. Here, the sampling frequency of the input signal has been halved by the use of the TDHS algorithm. Subjective tests taken from 20 listeners showed that the TDHS-ADPCM coder at 9.6, 12.8 and 16 kb/s are equivalent to robust conventional ADPCM coders at 24, 32 and 48 kb/s respectively, for all probabilities of error less than 0.5% [49]. In another paper, Copperi [50] presents two schemes for providing highly intelligible acceptable quality speech at 4.8 kbps, using TDHS algorithm and ADPCM systems. In one scheme, speech signals sampled at 7200 Hz, are frequency divided by a factor of 3, by using a two-step compression algorithm, and is fed to a 2 bit/sample ADPCM. In the next case, signal is sampled at 6400 Hz. Pitch values are also transmitted. It is shown that the output of a TDHS-ADPCM, with a 4-level quantizer closely matches the original one. The maximum SNR obtained was 13.81 dB, for female voice with an error probability equal to 0.0 [50].

Bertorello and Copperi [51] have designed a 4.8/9.6 kbps base-band LPC coder, using split-band and with vector quantisation in both the vocal tract modelling and residual representation. This system is slightly inferior to a 9.6 kbps coder using ADPCM and TDHS algorithm. But a 4 kbps base-band coder with vector quantisation is very superior to a 4.8 kbps LPC and channel vocoders, in preserving both intelligibility and naturalness.

A variability on the APC are the residual excited linear predictive coding (RELP), the multipulse LPC (MPLPC), the regular pulse excited LPC (RPELPC) and the code excited LPC (CELPC) [52]. These coders require still more complex algorithms but the performance is better.

In RELP coding [53], the prediction error is low-pass filtered and down-sampled, so that only fewer samples need be transmitted. At the decoder, by using non-linear distortion techniques, a full-band signal is obtained. The speech quality deteriorated, at bit rates below 8 kbps.

Bruce Fette et al [54] have implemented a high quality RELP coder at 4.8 kbps. They have coded the LPC representation of the spectrum, the energy and a frequency domain representation of the residual, within the base-band of 100-1000 Hz. The quantisation noise of the residual coding is passed through the LPC synthesizer and synthesized into speech, just as the actual residual excitation is synthesised. By coding only the base band frequencies, the effective number of bits per sample is raised by a factor of 6.4 [54].

In the MPLPC of Atal and Remde [55], a series of non-uniformly spaced pulses, typically eight per period [48], of different amplitudes, is used to excite the filter. No distinction is made between voiced and unvoiced speech, with respect to the excitation signal. The pulse positions and amplitudes, yielding the lowest error over a block, form the optimum excitation signal. Even for a very small block size and only a few pulses per block, the computational load is too much. Hence, sub-optimal methods to find the pulse positions and amplitudes one at a time, have been developed [55,56,57].

Singhal and Atal [58] developed a MPLPC, using long-term prediction and perceptual weighting. This reduces the number of pulses required per pitch period to obtain the same speech quality. It has been reported by Ozawa and Araseki [59] that MPLPC codecs operating at bit rates from 8 kb/s to 16 kb/s, produced good quality speech, as equivalent to those from higher bit rate codecs. Ono and Ozawa [60] have developed a good synthetic quality speech, using pitch prediction MPLPC, at 2.4 kb/s.

In the regular pulse excited (RPE) LPC coder [61], excitation signal pulses are spaced uniformly. The quantized pulse amplitudes and an index indicating the best excitation vectors, must be transmitted to the decoder. The speech quality obtained using RPE and MPLPC coders, are similar, at the same bit rates.

The code excited LPC (CELP) or the stochastic coder [62] is a modification of the MPLPC. Here, an 'innovation' sequence, consisting of M samples (typically 32 samples [1]), of white gaussian noise forms the excitation function. At the encoder, from a code book of many such sequences, the optimum one is selected by filtering each of these sequences, by passing it through a pitch filter and then an LPC vocal tract filter, and choosing the one which produces the minimum weighted mean squarred error. An index number to identify the selected sequence is transmitted to the decoder, where the block of M reconstructed speech samples is obtained by filtering, as at the encoder. CELP codecs operate at 4 to 8 kbps, but computation time is enormous, about 100 times more than real time on a Cray 1 Computer [62].

Copperi [9], has reported that, using a rule-based speech analysis to CELP coding, he obtained an average SNRSEG of 13.8 dB at 2.8 kbit/s. The coded speech quality was almost same as the original one.

Some improvements on the CELP coder is presented by Kroon and Atal [63] and Kleijn et al [64]. It is reported that, a two-stage vector quantization, using an adaptive and a stochastic codebook, provided a maximum SNRSEG of 12.1 dB.

Davidson and Gersho [65] have tried a multistage vector quantisation of the input speech vector, in a vector excited coder (VXC). The SNRSEG obtained using a 2-stage codebook was 14.8 dB (or an SNR of 16.7 dB) and for a 4-stage code book was 15.0 dB (or an SNR of 16.8 dB). Hernandez et al [66] have extended vector excitation techniques over to speech coding in the transform domain and have developed an efficient Vector Adaptive Transform Coder (VATC). The perceived quality of the VATC speech is slightly inferior to that of the CELP coded speech, though the SNR values are almost same.

Recently, Cuperman et al [78] have outlined a low-delay vector excitation coding (LD-VXC) algorithm at 16 kb/s, which provides high quality speech with less than 2 ms of coding delay and is robust to transmission errors. The algorithm combines techniques such as vector quantization, analysis-by-synthesis, perceptual weighting, together with backward adaptive linear predictive encoding and a long-term predictor employing backward adaptive tracking. Using a LD-VXC, with a 20th order lattice predictor, the subjective speech quality obtained was comparable to a 7-bit PCM with a MOS of about 4.0 [78]. The SNR value obtained was 18.48 dB.

Similar to the RELP coder of Bruce Fette et al [54], Kondoz and Evans [67] have applied CELP coding to the base-band residual (CELP-BB). Vector quantization is used to code the decimated baseband residual. At the decoder, received base-band is up-sampled by inserting zero-valued samples after each sample and then filtered through the LPC synthesis filter, to produce continuous good quality speech. An SNRSEG of 8.5 dB was achieved at 7 kbps [67]. Informal listening tests proved that a CELP-BB, a CELP and a vector quantized transform coder at 7 kbps, have a quality comparable to the original speech [67].

Rose and Barnwell [8] have presented a "near toll-quality" self-excited vocoder (SEV) at 4.8 kbps. In this SEV, no excitation signal is transmitted to the decoder, after initialization. The current source excitation is derived from the past history of the excitation signal itself, using two long-term predictors instead of the single one at the encoder. The initialization was done by using a Gaussian excitation sequence for the first 100 ms of a 3-s utterance and then setting the input to zero. Comparing the performance of an SEV, MPLPC and CELPC, all working at 4.8 kbps, it was seen that [8, 68], the objective performance of a CELPC (SNRSEG = 10.45 dB) is slightly better than that of an SEV (SNRSEG = 9.93 dB). The performance of a non-homogeneous (NH) coder (SNRSEG = 11.22 dB) is better than the above two. But subjective tests [Paired Acceptability Rating Method-PARM (explained in section 2.3)] showed SEV to be the best, with a score of 57.3 on a scale of 0 to 100, followed by CELP, with 55.6 and NH with 52.9. The computation needed in SEV is only 4 MFLOPs compared to the 80 MFLOPs in ordinary CELPC [8].

Very recently, Dedes et al [10] developed an APC, which can change bit rate on a packet-by-packet basis. Using more than one quantizer to quantize the residue, an almost toll quality speech was produced at rates between 16 kb/s and 9.6 kb/s [10]. Dedes et al have obtained an SNR value between 14.5 and 18.55 dB, a log-likelihood ratio (LLR) of 0.0850 to 0.2071 and a log-area ratio (LAR) of 0.0513 to 0.0833 at 16 kbps, and 10.3 to 14.65 dB, 0.0748 to 0.1202 and 0.1618 to 0.3578 respectively at 9.6 kbps. This SNR value is 10 dB more than the gain of a conventional APC of Atal [45] (The SNR value obtained for the APC of Atal, using the above same sentences, was only 2.60 to 3.60 dB [10], though Atal [45] has claimed a gain of 12 dB).

A comparative study of the various types of predictive coders explained above is presented in Table 2.2.

To summarise, the best performance quoted of an hybrid coder is that of the CELP coder, which gives a SNRSEG value of 13.8 dB at 2.8 kbps and a quality almost same as the original. But this is achieved at the expense of an enormous amount of computational burden. The SEV and the NH coder show an objective performance almost equivalent to that of the CELP coder, at 4.8 kbps, with a reduction in computation, by an order of magnitude.

2.4 Measurement of Coder Performance

It is a difficult task to assess the performance of different coders on a single scale, as they differ in their basic operations. As yet, it is not completely understood how the human ear interprets the sound signals that

Table 2.2 A Comparative Study of the Various Predictive Coding Systems

	Authors	Systems and their features	Performance summary	Bit Rate
	(1)	(2)	(3)	(4)
1	Atal & Hanauer 1971, [14]	Linear predictive coder- analysis segment length was one pitch period for voiced sections and 10 msecs for unvoiced regions. Considered predictors of order p between 2 and 18.	No significant difference in the quality of speech for p above 12	7.2 to 2.4 kbps
2	Atal & Schroeder 1970, [44]	Adaptive predictive coder (APC)- using 8th order spectrum prediction, 1st order pitch prediction and, 1-bit adaptive quantizer for residual transmission. Adaptive every 5 msecs.	Subjectively superior to a log-PCM system with 5 bits/sample	10 kb/s
3	Atal 1982, [45]	APC-NF: APC with noise-feedback and using entropy-coded 3-level quantizer	Average SNR of 21 dB, as compared to the 13 dB of open-loop DPCM and 23 dB of a standard DPCM	19.2 kb/s
4	Daumer 1982, [23]	APC-NF: APC using noise-feedback	MOS score of 4.0, on a scale of of 1 to 5	16 kb/s

	(1)	(2)	(3)	(4)
5	Copperi 1982, [49]	ADPCM using Time domain Harmonic scaling (TDHS) algorithm– changes speech rate by discarding or repeating short pieces of the waveform	TDHS–ADPCM at 9.6, 12.8 and 16 kb/s are respectively equivalent to conventional ADPCM at 24, 32 and 48 kbps, for all error probabilities less than 0.5%.	
6	Copperi 1982, [50]	TDHS–ADPCM– using compression algorithm and 4–level quantizer	Maximum SNR is 13.81 dB. The speech output closely matches the original one.	4.8 kb/s
7	Bertorello & Copperi 1983, [51]	Baseband LPC– using split band and vector quantisation (VQ)	Slightly inferior to a 9.6 kb/s TDHS–ADPCM, but a 4 kb/s base–band coder with VQ is very superior to a 4.8 kb/s LPC and channel vocoders.	4.8 to 9.6 kb/s
8	Dankberg & Wong 1979, [53]	RELPC– LPC system in which the residual is low pass filtered and down sampled.	Speech quality below 8 kbps is very poor.	8 kb/s
9	Fette et al 1988, [54]	RELP coder– using LPC representation of spectrum, energy and a frequency domain representation of the residual within the band 100–1000 Hz.	High quality coder output	4.8 kb/s
10	Atal & Remde 1982, [55]	MPLPC– LPC system using non-uniformly spaced pulses of different amplitudes, to excite the filter	High quality speech output	
11	Singhal & Atal 1984, [58]	MPLPC– using long term prediction and perceptual weighting	Same high quality output comparable to the standard MPLPC.	

	(1)	(2)	(3)	(4)
12	Ozawa & Araseki 1986, [59]	MPLPC	Good quality speech, as equivalent to those from higher bit rate codecs.	8 to 16 kb/s
13	Ono & Ozawa 1988, [60]	MPLPC- using pitch prediction	Good synthetic quality speech	2.4 kb/s
14	Deprette & Karoon 1985, [61]	RPELPC- LPC system in which the excitation pulses are uniformly spaced.	Quality same as MPLPC	
15	Schroeder & Atal 1985, [62]	CELPC- LPC in which, an index corresponding to the optimum innovation sequence, of white Guassian noise is transmitted as the excitation signal.	Good quality speech almost close to original	4 to 8 kb/s
16	Copperi 1988, [9]	CELPC- using a rule-based speech analysis	Quality almost same as the original. SNRSEG is 13.8 dB	2.8 kb/s
17	Kroon & Atal 1988, [63]	CELPC- shaping the code book excitation function	Quality better than the standard CELP coder.	
18	Kleijn et al 1988, [64]	CELPC- 2-stage vector quantisation, using an adaptive codebook and a stochastic code book, is made use of.	Maximum SNRSEG is 12.1 dB	4.8 to 8.0 kb/s
19	Kondoz & Evans 1988, [67]	CELP-BB (CELP coding of base-band residual)- vector quantisation of the decimated base band residual	SNRSEG is 8.5 dB and SNR value is 7.9 dB. Quality of speech output is close to original.	7 kb/s

	(1)	(2)	(3)	(4)
20	Davidson & Gersho 1988, [65]	Vector excited coder (VXC)– using multistage vector quantisation of the input speech vector	SNRSEG of 14.8 dB (or SNR of of 16.7 dB) for a 2–stage code book and 15.0 dB (or SNR of 16.7 dB) for a 4–stage code book.	
21	Cuperman et al 1992, [78]	Low–delay vector excitation coding (LD–VXC)– using vector quantization, perceptual weighting and adaptive long–term prediction	With a 20th order lattice predictor, the subjective quality was comparable to a 7–bit PCM with a MOS of about 4.0. SNR value obtained was 18.48 dB	16 kb/s
22	Hernandez et al 1988, [66]	Vector adaptive transform coder (VATC)– using vector excitation techniques in the transform domain	Perceived quality of speech output slightly less than CELP coded speech, but SNR values are almost same.	
23	Rose and Barnwell 1990, [8]	Self–excited vocoder (SEV)– No excitation signal is transmitted after initialization	Near toll–quality speech is obtained. Comparing an SEV with with a CELPC and a non–homogeneous (NH) coder, objective performance of SEV is slightly lower than NHC and CELPC, but subjectively SEV is the best.	4.8 kb/s
24	Dedes et al 1992, [10]	APC– changing bit rate on a packet–by–packet basis. Used noise shaping and more than one quantizer to quantize the residue.	At 16 kb/s, SNRSEG is 14.5 dB to 18.55 dB, log likelihood ratio (LLR) is .085 to .207 and log area ratio (LAR) is .0513 to .0833. At 9.6 kb/s, the corresponding values are between 10.3 and 14.65 dB, .162 and .3578, and .075 and .120.	9.6 to 16 kb/s

reach its cochlea. Hence a mathematical expression to quantify the "speech quality" has not been completely possible. However, based on the present knowledge of speech perception, many performance measurements have been evolved. They are mainly classified as: objective measures, using mathematical expressions and subjective measures based on listening tests. Objective quality measurements like SNR and SNRSEG give a good indication of the subjective rating of a coder at high bit rates, but at low bit rates, where high complexity coders are considered, they do not correlate with subjective quality [10].

2.4.1 Objective Measurements

There are mainly three categories of speech coding systems. Of these, the waveform coders, try to preserve the shape of the speech waveforms, while the vocoders try to mimick the speech sounds. The third type, the hybrid coder, have characteristics common to both. To suit these distinct types of coders, two types of objective measures are used. The spectral distance measures are better suited to vocoders and hybrid coders, while the SNR and related measures are highly suited to waveform coders.

2.4.1.1 Signal-to-Noise Ratio (SNR)

It is the ratio of the input signal variance to the reconstruction error variance. If x(n) is the coder input at sampling instant n, and y(n) is the corresponding coder output, then the reconstruction error is given by

$$r(n) = y(n) - x(n) \qquad (2.1)$$

The original signal variance is computed as

$$E_x^2 = \frac{1}{N}\sum_{n=1}^{N}[x(n) - \frac{1}{N}\sum_{n=1}^{N} x(n)]^2 \tag{2.2}$$

The error signal variance is

$$E_r^2 = \frac{1}{N}\sum_{n=1}^{N}[r(n) - \frac{1}{N}\sum_{n=1}^{N} r(n)]^2 \tag{2.3}$$

where, N is the number of samples in the interval for which the SNR is calculated.

$$\text{Hence, SNR} = \frac{E_x^2}{E_r^2} \tag{2.4}$$

It is expressed in dB as

$$\text{SNR(dB)} = 10 \log_{10}(\text{SNR}) \tag{2.5}$$

2.4.1.2 Segmental Signal-to-Noise Ratio (SNRSEG)

Speech signals are, by nature, non-stationary, and the same amount of noise has different perceptual values depending on the ambient signal level. To take these facts into consideration, a segmental SNR (SNRSEG) measure is evolved. The SNRSEG [69] is based on dynamic time-log-weighting, so that very high SNR values of the well-coded large-signal segments do not camouflage the coder performance with the weak segments. To compute the SNRSEG, divide the speech signal into segments of short duration, and evaluate the SNR(m) dB, where m = 1,2,....M corresponds to the block number. Then, the segmental SNR is given by

$$\text{SNRSEG (dB)} = \frac{1}{M} \sum_{m=1}^{M} \text{SNR(m) dB} \tag{2.6}$$

2.4.1.3 Articulation Index (AI)

Noise in certain frequency bands is less harmful than that in other bands of an input signal. In speech, in the region from 100 to 1000 Hz, error perception in a speech coding system increases as a function of increasing frequency. Hence, a frequency-weighted SNR index called the articulation index (AI) was used in early speech work [2]. To compute the AI, the speech signal is subdivided into 20 sub-bands, and the signal-to-noise ratio SNR_i(dB) for each band "i" is calculated. Limiting the SNR to a maximum value of 30 dB, the AI is calculated as

$$\text{AI} = \frac{1}{20} \sum_{i=1}^{20} [\min(\text{SNR}_i, 30)/30] \tag{2.7}$$

2.4.1.4 Itakura-Saito's Log-Likelihood Ratio

The human ear is not very sensitive to the short-term phase [43] and hence in vocoders, only the magnitude of the speech spectrum is usually preserved. So, the vocoder output waveform might be quite different from the original speech, but still it will be quite intelligible, and sound the same. Hence, distance metrics, sensitive to spectral differences are to be used to measure the fidelity of the vocoder outputs. These measures are often used for the LPC method.

Let a_1 be the predictor coefficient vector and R_1 the autocorrelation matrix of the input speech. Let a_2 and R_2 be the corresponding quantities of the coder output speech. Then the likelihood ratio is defined as [10]:

$$d_{LR} = \frac{a_2^T R_1 a_2}{a_1^T R_1 a_1}$$

$$= \frac{a_1^T R_2 a_1}{a_2^T R_2 a_2} \qquad (2.8)$$

The log likelihood ratio is the logarithm of the above expression, and is written as

$$d_{LLR} = 10 \log_{10} d_{LR}$$

2.4.1.5 Log-Area Ratio

The Euclidean distance metric is defined as [10]:

$$d_{LAR} = [\frac{1}{p} \sum_{i=1}^{p} (LAR_{1i}-LAR_{2i})^2]^{1/2} \qquad (2.9)$$

where, the subscripts 1 and 2 correspond to the input and output speech respectively and p is the order of the Autoregressive model.

The log-area ratio LAR_i is given by

$$LAR_i = \log_{10} \frac{1+k_i}{1-k_i} \tag{2.10}$$

where k_i are the PARCOR coefficients.

2.4.2 Subjective Measurements

In the perception of any communication signal, the main mechanism for noise measurement is the human perception mechanism. Hence, perceptual and subjective testing procedures, for determining the quality and intelligibility, should supplement the objective measurements, while determining the efficiency of a coder. A good quality output is highly intelligible, but the converse is not true.

2.4.2.1 Quality Tests

The assessment of low bit rate speech codecs is more difficult than those of the waveform codecs used at higher bit rates, since the distortions produced by low bit rate codecs are very diverse in nature and these degradations will be assessed by different persons in a varied manner.

The Mean Opinion Score (MOS) is a systematic approach to determine the speech quality. In this test, an ensemble of well-trained listeners are asked to classify a stimulus (coder output) on an N-point quality scale, for signal quality or impairment. Or else, a particular number or value can be associated with each category, say,

Number scores	Quality scale
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

The MOS value, is a pooled average judgement for the ensemble of listeners. A 64 kbps PCM speech output has a representative MOS value of 4.53 and a standard deviation of 0.57 [23].

Another speech quality measurement is the Modulated Noise Reference Unit (MNRU) [1,2], as recommended by the CCITT. Here, the speech which is passed through the codec is compared with speech to which distortion has been added by a "modulated noise reference unit". The MNRU is calibrated in terms of signal-to-correlated-noise ratio (Q). If one quantization distortion unit (1 qdu) is defined as the distortion from one commercial A-law or μ-law PCM codec, then the qdu and the Q values can be related as [1]

$$\text{qdu} = 10^{(37-Q)/15} \tag{2.11}$$

The subjective equivalent distortion given by the above formula, has been found to give reliable results for the 64 kbps PCM and 32 kbps ADPCM [1,2].

2.4.2.2 Intelligibility Tests

In intelligibility tests, the information-bearing contrast features of the decoded word is tested.

In the Diagnostic Rhyme Test (DRT) of Voiers [70], a set of 96 rhyming pairs are considered. The subject is presented with a coded word and is asked to recognize the possible stimulus from a given pair, by noting the extent or the presence of each of the six attributes of consonant phonemes, like voicing in "vast versus fast" and nasality in "moot versus boot" [2]. The coded word is intelligible if the response from the listener is correct.

The DRT score is given by

$$P = \frac{R-W}{T} \; 100\% \qquad (2.12)$$

where,

R = number of right answers

W = number of wrong answers

and T = total number of items involved.

The typical DRT score is between 75 and 95, and for a "good" system, the DRT score must be 90.

In the Modified Rhyme Test (MRT), the listener is presented with one coded word from a closed rhyming set of six words (example, meat, feat, heat, seat, beat, neat), differing in the beginning consonant, and is asked to select the word.

From table 2.2, it can be noted that all the researchers on predictive coders [2, 10, 23, 27, 44, 45, 46, 50-68], have used SNRSEG as the objective measure of their coder performance, with mere listening tests to supplement the quality determination. Also, some researchers [66-68] have shown that the SNRSEG tests and the quality tests (by mere listening) are showing almost same coder performance. Only two researchers have used concrete subjective tests in addition to the objective SNRSEG test. Daumer [23] has used MOS value and Rose and Barnwell [8] have used a Paired Acceptability Rating Method (PARM) to determine the subjective performance of their coder. It is also reported that [2], the SNRSEG shows more closeness to the subjective rating of the speech coders than the usual SNR. Hence, from among the performance measurements explained in section 2.4, in this thesis, the objective measure, based on the segmental signal-to-noise ratio is used.

Chapter 3

PREDICTIVE CODING METHOD FOR THE ANALYSIS AND SYNTHESIS OF SPEECH SIGNALS

3.1 Introduction

The aim of efficient coders, is to use minimum channel capacity to transmit a signal, within a given time, with a specified fidelity. For this, the redundancy of the transmitted signal should be reduced. Predictive coding [44] is one technique to reduce redundancy. In this method, redundancy is removed by subtracting from the signal that part which can be predicted from its previous samples. The difference signal itself, or a related version of it, is transmitted.

In linear predictive coding (LPC) [11,14], the speech waveform is represented as the output of a linear time-varying filter which is excited by an appropriate excitation signal. Or in other words, the current sample can be approximated as a linear combination of its past values.

3.2 The Human Speech Production Mechanism

The human speech production apparatus consists of the vocal tract, terminated at one end by the vocal cords and at the other end by the lips. The nasal tract can be connected or disconnected to the vocal tract, by the movement of the velum. The shape of the vocal tract is determined by the position of the lips, jaw, tongue and velum. The human

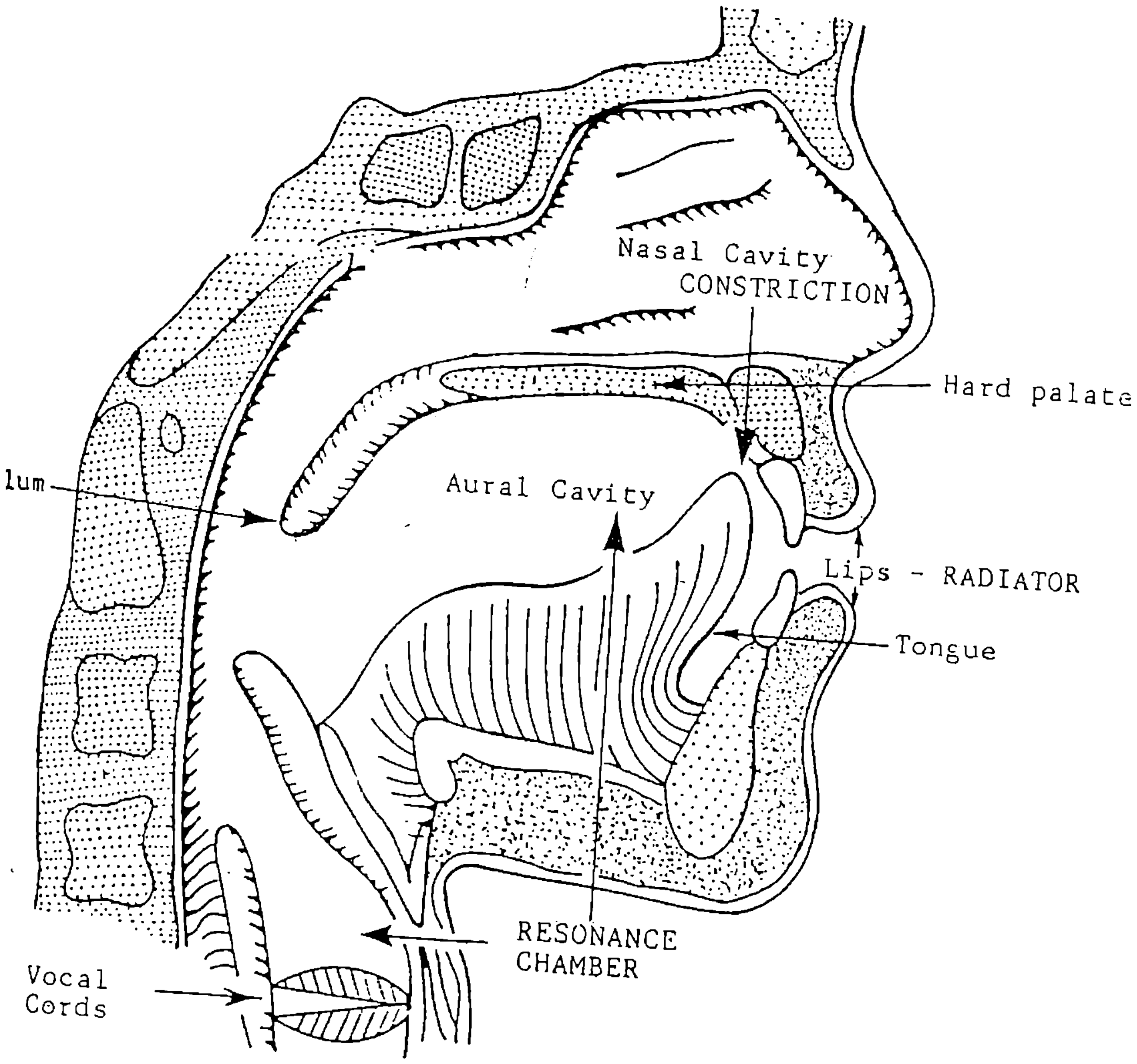


Fig.3.1 Euman Vocal System

vocal system is shown in Fig.3.1.

The air we inhale in, is pushed out of the lungs, through the trachea, past the vocal cords and into the pharynx, and finally comes out through the mouth or nose or both, and is perceived as speech [72,73]. If the air, on its way out, causes the vocal cord to vibrate in a periodic manner, and this vibration is transmitted on to the vocal tract, then a voiced sound is produced. If the source of vibration is a turbulent one, produced by forming a constriction or complete closure within the tract, and then the sudden release of pressure, then we perceive an unvoiced sound. These sources create a wide-band excitation of the vocal tract, which acting as a linear time-varying filter, imposes its transmission properties on the frequency spectra of the sources. Only the first three formants are important in determining the sound that is heard [16, 44].

3.3 The LPC Speech Production Model

Speech production models usually treat the vocal tract and the air entering the vocal tract (that is, the "excitation") separately [52]. In the LPC analysis [11], the time-varying linear digital filter, representing the vocal tract, represents the effects of the lip radiation, the glottal pulse shape and the nasal cavity coupling (wherever required). A set of parameters are extracted from the speech signal, to specify the filter transfer function, which gives the best match to the speech signal that is being coded. It is noted that, in the spectral envelope of the short-term speech signal, there are a number of peaks at frequencies closely related to the formant frequencies; that is, the resonant frequencies of the vocal tract. Similarly, it is the spectral peaks (that is, the resonances represented by the poles of the filter) and not the spectral troughs (that is, the antiresonances represented by the zeroes) that are most significant to speech perception [43, 52]. Hence, an all-pole filter of order p, in the range 10 to 20, is a good model, relating to the way in which the speech is produced and perceived [52].

A digital model for speech signals is shown in fig.3.2. For voiced speech sounds, the filter is excited by a quasi-periodic pulse train, in which the spacing between the impulses corresponds to the fundamental period of the glottal excitation. For unvoiced speech, the filter is excited by a random number generator that produces flat spectrum noise. The amplitude control regulates the intensity of the input to the filter. The filter parameters determine the spectral characteristics of the particular sound, for each type of excitation.

3.4 A Predictive Coding System

The block diagram of a predictive coding system [44] is shown in fig.3.3. The input signal S(t), is sampled at the Nyquist rate, to produce the samples $\{S_n\}$. At the transmitter, the predictor makes an estimate $\hat{S}_n$ of the present value S_n of the signal, based on the past samples of the reconstructed signal, r_{n-1}, r_{n-2}, ...·The difference between the actual and the predicted value, $d_n = S_n - \hat{S}_n$, is quantised, encoded and transmitted to the receiver. At the receiver, the transmitted difference signal is decoded

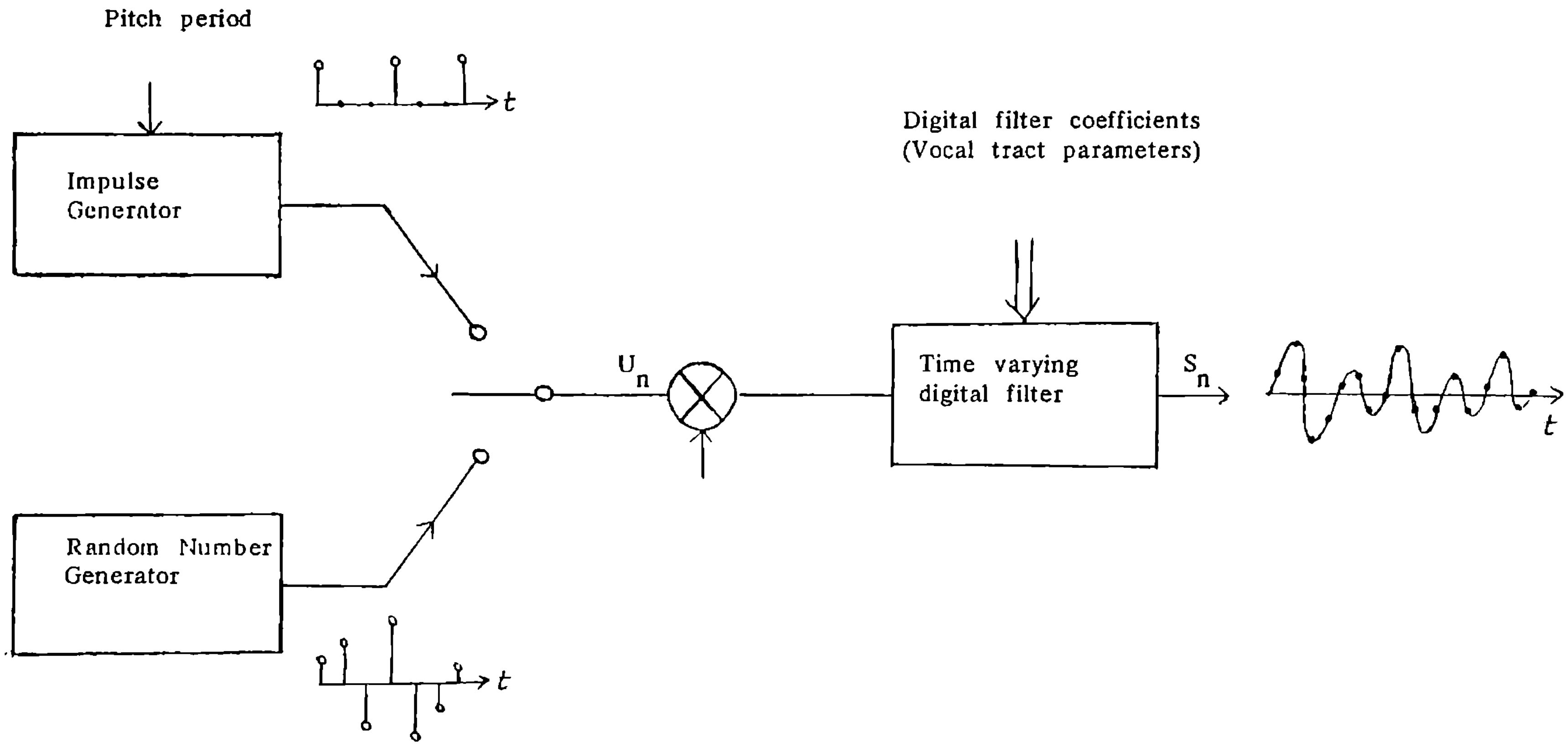

Pitch period
Impulse Generator
t
Random Number Generator
t
U_n
Digital filter coefficients (Vocal tract parameters)
Time varying digital filter
S_n
t

Fig.3.2 Digital Model for Speech Signals

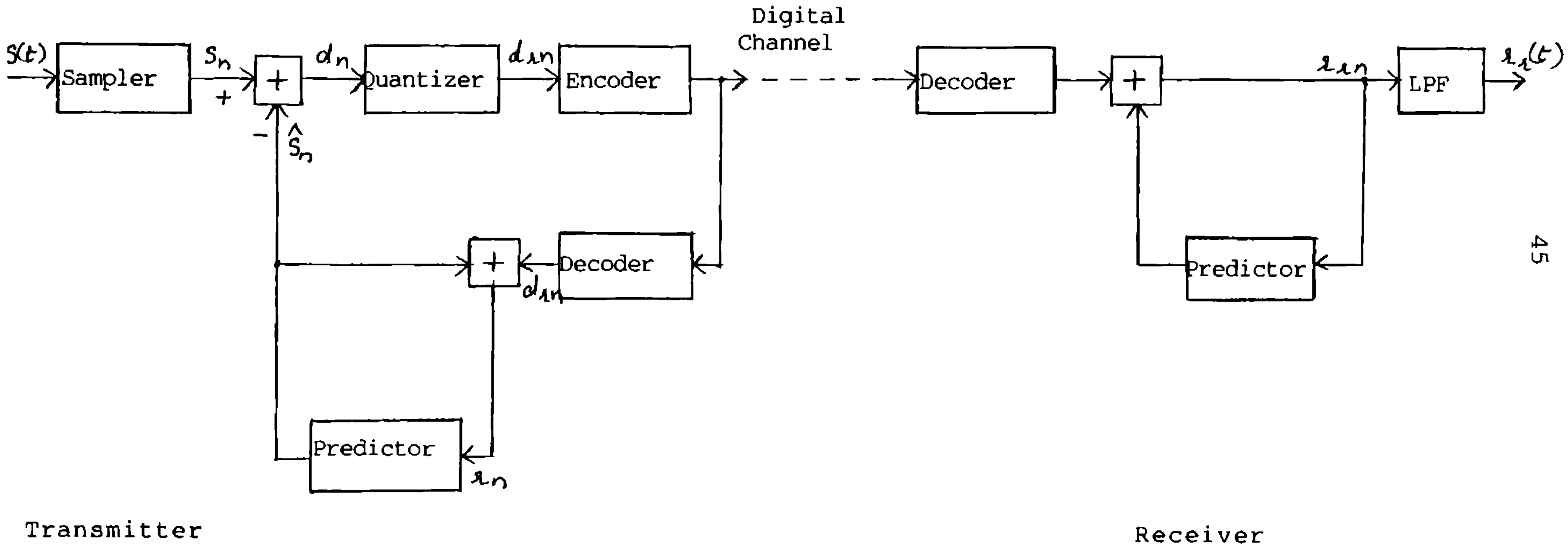


Fig.3.3 Block diagram of a Predictive Coding System [44]

and added to the predicted value of the signal to form the reconstructed samples r_m, which are then low pass filtered to produce the output signal $r_r(t)$. The predictor used at the receiver, is made identical to that at the transmitter by transmitting the predictor parameters also. To follow the non-stationary nature of speech signals, the predictor is updated at regular intervals.

In the original LPC system of Atal and Hanauer [14], only the predictor parameters are transmitted to the receiver. The signal is reconstructed using these parameters. The predictor is updated every pitch period for voiced sounds and every 10 msec for unvoiced sounds.

3.5 Signal-to-Quantizing Noise Ratio

For the system in fig.3.3, let

P_s be the mean square value (MSV) of the input signal samples S_n ,

P_d be the MSV of the difference signal samples d_n,

P_q be the MSV of the quantizing noise in the decoded difference signal d_m and

P_e be the MSV of the quantizing noise in the reconstructed signal r_m.

Then, the signal-to-quantising noise ratio of the reconstructed signal is given by

$$\mathrm{SNR} = \frac{P_s}{P_e} = \frac{P_s}{P_d} \cdot \frac{P_d}{P_e} \tag{3.1}$$

In the absence of digital channel transmission errors, $P_e = P_q$.

$$\therefore \quad \mathrm{SNR} = \frac{P_s}{P_d} \cdot \frac{P_d}{P_q} \tag{3.2}$$

For speech signals, P_s/P_d is nearly 100 [44], and hence from equation (3.2), it can be seen that, using predictive coding, an improvement of about 20 dB in the SNR value can be expected, over a PCM system, using identical quantising levels.

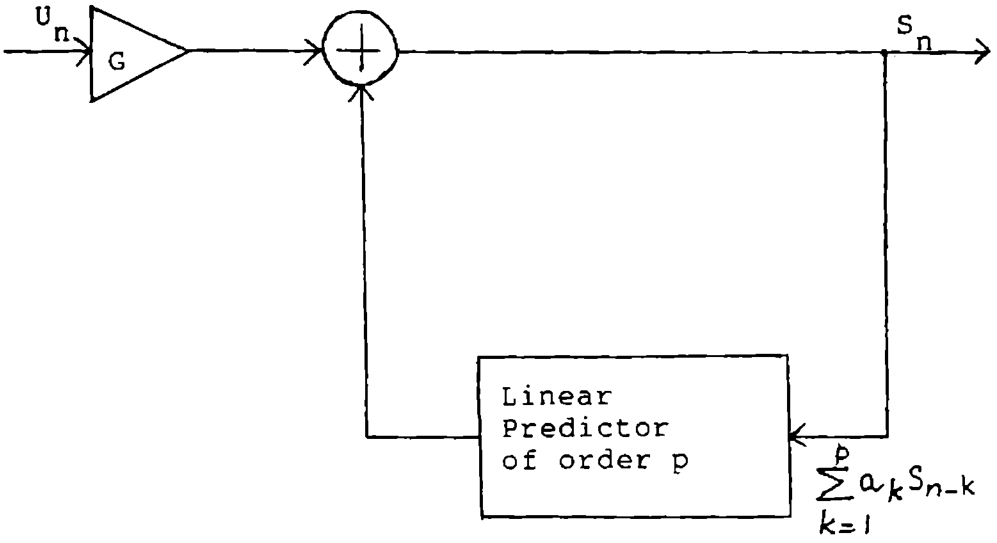
3.6 The Linear Predictive Coder - Using Spectrum Prediction

3.6.1 Theory

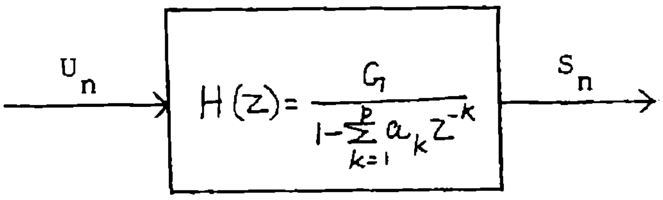
In an all-pole (auto-regressive AR) model, the nth sample S_n is represented by [2,3,11,14]

$$S_n = \sum_{k=1}^{p} a_k S_{n-k} + GU_n \tag{3.3}$$

where,



(a) Time Domain



(b) Frequency Domain

Fig.3.4 The Auto Regressive Model of a Linear Predictor

U_n is the excitation,

G is the gain factor,

a_k's are the filter coefficients, and

p is the order of the filter.

p should be 12 for voiced speech but need be only 6 for unvoiced sounds [14]. Equation (3.3) can be implemented using the AR model of fig.3.4.

If the input U_n is totally unknown, then the estimated value of S_n is

$$\hat{S}_n = \sum_{k=1}^{p} a_k S_{n-k} \quad . \tag{3.4}$$

3.6.2 Parameter Estimation

The filter (predictor) parameters are determined by minimising the total or mean squared error between the actual and the predicted value, with respect to each of the parameters, in the time interval over which the predictor is to be optimum.

The total squared error E is given as

$$E = \sum_{n} E_n^2 = \sum_{n} (S_n - \hat{S}_n)^2 \quad . \tag{3.5}$$

Setting $\partial E/\partial a_i = 0$, leads to the equation,

$$\sum_{k=1}^{p} a_k \sum_{n} S_{n-k}S_{n-i} = \sum_{n} S_nS_{n-i}, \quad 1 \leq i \leq p \tag{3.6}$$

The above equations (eqn. 3.6), called the normal equations, form a set of p equations in p unknowns. They can be solved for the predictor coefficients, a_k, $1 \leq k \leq p$, which minimise the error E. The minimum total squared error E_p, is obtained by substituting eqn. (3.6) in eqn. (3.5) and is given by

$$E_p = \sum_{n} S_n^2 + \sum_{k=1}^{p} a_k \sum_{n} S_nS_{n-k} \tag{3.7}$$

Depending on the range of summation over n in the above equations, two distinct cases arise.

3.6.2.1 Autocorrelation Method

Here n is assumed to have infinite duration and eqn. (3.6) reduces to the form,

$$\sum_{k=1}^{p} a_k R(i-k) = R(i), \qquad 1 \leq i \leq p \tag{3.8}$$

where ,

$$R(i) = \sum_{n=-\infty}^{\infty} S_nS_{n+i}, \tag{3.9}$$

is the autocorrelation function of the signal $\{S_n\}$.

The coefficients R(i-k) form an autocorrelation matrix. But in practice, a window function w_n is used to limit the signal to some interval,

$0 \leq n \leq N-1$. Then equation (3.9) reduces to the form:

$$R(i) = \sum_{n=0}^{N-1-i} S_n' S_{n+i}', \quad i \geq 0 \qquad (3.10)$$

where,

$$S_n' = \begin{cases} S_n W_n, & 0 \leq n \leq N-1 \\ 0, & \text{otherwise} \end{cases} \qquad (3.11)$$

The shape of the window can be altered.

3.6.2.2 Covariance Method

In this method, only a finite segment of speech is considered, such that $0 \leq n \leq N-1$. Equation (3.6) now reduces to

$$\sum_{k=1}^{p} a_k \varphi_{ki} = \varphi_{oi}, \qquad 1 \leq i \leq p \qquad (3.12)$$

where,

$$\varphi_{ik} = \sum_{n=0}^{N-1} s_{n-i} s_{n-k} \qquad (3.13)$$

is the covariance of the signal $\{S_n\}$.

The coefficients φ_{ik} form a covariance matrix. The covariance matrix is symmetric; but unlike the autocorrelation matrix, the terms along each diagonal are not equal. Compared to the autocorrelation coefficients, for the computation of the covariance coefficients, p samples preceeding the current window are also required.

The method for computing the gain factor G, is given in Appendix I.

Comparing the autocorrelation method and the covariance method, the following points can be noted. The covariance method is quite general and can be used without any restriction. But the resulting filter may not be stable, which is not a severe problem. In the autocorrelation method, the filter is guaranteed to be stable, but due to the windowing of the time signal, parameter inaccuracy arises. This creates problem if the signal is a portion of an impulse response. Since the autocorrelation matrix is symmetric Toeplitz, the number of matrix elements that are to be evaluated is very less compared to the covariance method (that is, only p instead of p(p+1)/2).

Speech signals are non-stationary in nature. It is seen that, speech waveform is nearly periodic during voiced regions, and hence the current signal value can be predicted based on the signal value exactly one period earlier. But the period of the speech signal varies with time and therefore, the predictor must be updated periodically. These varying coefficients must also be transmitted to the receiver. This does not consume excessive channel capacity because the coefficients tolerate coarse quantisation and slow updating [2].

3.7 The Linear Predictive Coder - Using Pitch Prediction

The two main causes of redundancy in speech are: (i) quasi--periodicity during voiced segments and (ii) lack of flatness of the short-time spectral envelope [44]. It is the pitch excitation that produces the quasi-periodicity in the amplitude-time waveform and also the fine structure in the short-time log-spectrum of voiced speech segments [2]. The near sample based predictor (spectrum predictor) does not consider this fine structure; they exploit only redundancies in the spectral envelope. Hence the prediction error sequence is not white, but is structured in accordance with the speech periodicity. To remove this periodic structure, a second stage of prediction, called the pitch prediction, exploiting the correlations between the speech sample being coded and a sample, or a set of samples one pitch period away is considered.

3.7.1 Theory

Considering a voiced segment of speech, the nth sample can be expressed as [44]

$$S_n = \sum_{k=1}^{p} a_k S_{n-k} + U_n \tag{3.14}$$

where

a_k's are the predictor coefficients,

p is twice the number of formants of the vocal tract, in the frequency range of interest, and

U_n is the input (excitation) to the filter.

U_n can be written as [44]:

$$U_n = \beta\, U_{n-M} \tag{3.15}$$

where M is the period of the excitation signal and is the pitch period of the voiced speech segment. β is the "pitch gain" [2] which takes account of the variations in the amplitude of the signal from one pitch period to the next. Neglecting the variations in the coefficients a_k from one pitch period to the next, we obtain:

$$S_n - \beta\, S_{n-M} = \sum_{k=1}^{p} a_k (S_{n-k} - \beta\, S_{n-k-M}) + U_n - \beta\, U_{n-M}$$

or

$$S_n = \beta\, S_{n-M} + \sum_{k=1}^{p} a_k (S_{n-k} - \beta\, S_{n-k-M}) \tag{3.16}$$

The above equation determines the structure of the linear predictor, using both spectrum and pitch prediction.

The prediction equation (3.16) can be implemented by the predictor configuration given in fig.(3.5).

$P_1(Z) = \beta\, Z^{-M}$ removes the quasi-periodic nature of the speech signal and $P_2(Z) = \sum_{k=1}^{p} a_k Z^{-k}$ removes the formant information from the spectral envelope. The first predictor is just a gain and delay arrangement, while the second one forms a linear combination of the past p values of the difference between the actual signal value and the value

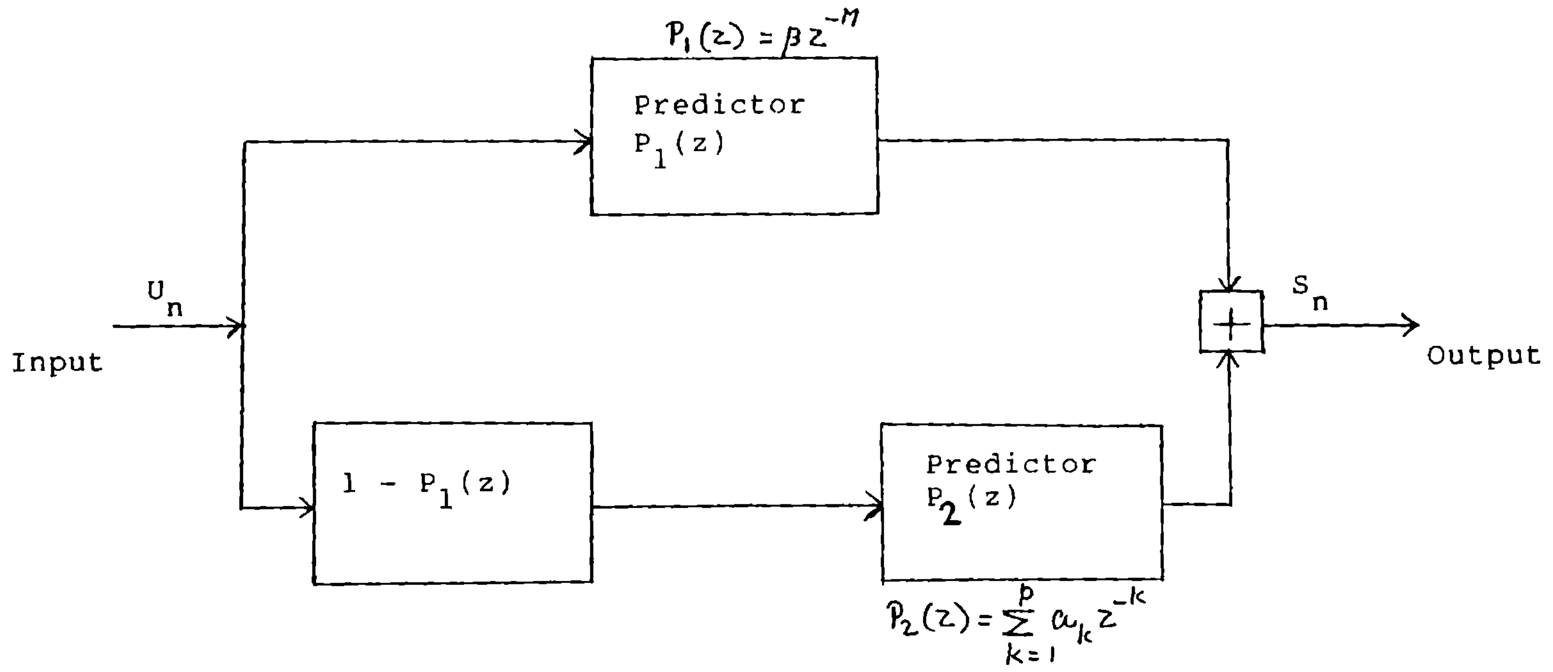


Fig.3.5 Block diagram of the Linear Predictor, including both Spectrum and Pitch Predictors [2, 44].

predicted by $P_1(Z)$.

The predictor parameters - M, β and a_k's - should be readjusted periodically to match the varying characteristics of the input speech signal.

For unvoiced sounds, U_n represents a noise-like excitation. For practical purposes, neglecting the effect of zeroes, equation (3.16) can represent the prediction for unvoiced sounds too, if β is assumed to be zero.

3.7.2 Parameter Estimation

The predictor parameters are determined by minimising the mean square error between the true and the predicted values of the speech samples, with respect to each of the parameters, in the interval over which the predictor is to be optimum.

The details of the parameter estimation method are given in Appendix I. The relevant equations are noted below. The predicted value of the nth sample is given by [2,44].

$$\hat{S}_n = \beta S_{n-M} + \sum_{k=1}^{p} a_k(S_{n-k} - \beta S_{n-k-M}) \tag{3.17}$$

The correlation parameter β is expressed as

$$\beta = \langle S_n S_{n-M} \rangle_{av} \Big/ \langle S^2_{n-M} \rangle_{av} \tag{3.18}$$

where, $\langle \ \rangle_{av}$ denotes the averaging over all the samples in the intervaL The optimum value of M is obtained by locating the position of the maximum of the normalized correlation coefficient ρ given by:

$$\rho\ (M) = \langle S_n S_{n-M} \rangle_{av} \Big/ \left\{ \langle S^2_n \rangle_{av} \ \langle S^2_{n-M} \rangle_{av} \right\}^{1/2}, \quad M > 0 \tag{3.19}$$

Using these values of M and β , the predictor coefficients a_k's are obtained by solving for the vector “a" from the matrix equation.

$$\phi\, a = \psi \tag{3.20}$$

where, ϕ is a p by p matrix with its ijth element given by

$$\varphi_{ij} = \langle v_{n-i}\ v_{n-j} \rangle_{av} \tag{3.21}$$

and

$$v_n = S_n - \beta\, S_{n-M} \tag{3.22}$$

ψ is a p by 1 vector, with its jth component given as

$$\psi_j = \langle v_n v_{n-j} \rangle_{av} \tag{3.23}$$

3.8 Development of the Modified Coder

The main aim of this work is to reduce the computational burden and the complexity of the coder, and to reduce the transmission rate, for a moderate SNR value of the reconstructed signal. The modified block adaptive coder (MBAC) is a modified version of the adaptive predictive coder of Atal and Schroeder [44].

To start with, samples from the word "JOE", sampled at 8 KHz and encoded to 12 bits were used. 64 samples in the voiced region were considered and the algorithm of equation (3.17) was applied. The value of M was determined by locating the position of the first maximum of the normalised correlation coefficient ρ (M), for $M > 0$, and the values of β and hence a_k's were evaluated. The SNR value was computed for values of p from 2 to 12. It was noted that the SNR value was almost same above $p = 4$. That is, the order of the filter required for prediction need only be equal to the number of formants in the frequency range of interest and not twice that number, as hitherto dictated [44]. But the value of SNR was very low (around 1 dB).

On scrutiny, it was noted that the value of M was less than p and hence it did not correspond to the actual pitch period of the signal, but only to the position of the nearest maximum correlated sample. Also, for values of M less than p, the terms inside the bracket on the right hand side of equation (3.17), cancel each other, causing high error.

Considering M as the position of the second maximum in the normalised correlation function, a value of M equal to 38 was obtained and the SNR value increased to 10 dB, when the first (p+M) sample values alone were considered known. When the difference signal was also used, as in an APC system, the SNR value went upto 14 dB. This gave an insight to direct the research in the line, as to try to predict the sample values based on the earlier predicted values alone, without transmitting the difference signal. If the prediction is good enough, the error values will be very low and it would be possible to get a good representation of the signal, without transmitting the error samples, $\{d_n\}$. This will provide a good saving in the transmission rate. With these ideas, the modified coder was developed.

3.9 Steps Towards a Modified Predictive Coder

The various modifications that have been done on the adaptive predictive coder (APC) are explained below.

3.9.1 Reduction in the Computational Load

Preliminary simulation studies on the APC, revealed that the computational load involved in the evaluation of the predictor parameters can be reduced by making certain changes in the evaluation methods. The changes that evolved, are explained below.

3.9.1.1 Pitch Period Determination

The first step in the implementation of the predictor algorithm, is to divide the input data into different blocks, and check whether the block belongs to voiced, unvoiced, silent or transition region. Once the block under consideration is detected to be voiced, the first parameter to be computed is M, which represents the number of samples within a pitch period of the signal being processed. Or in other words, M gives a measure of the pitch period.

Many researchers have put forth different methods for the determination of pitch period. In the inverse filtering approach [14] and the PARCOR approach [14], pitch is determined by noting the peak positions in the linear prediction error signal. For this, the predictor coefficients are to be first evaluated and then the error signal (that is, input to the inverse filter 1/H(z)) is to be retrieved. The SIFT (simplified inverse filter tracking) algorithm [74] is another method. Here, the amount of processing required is reduced by pre-filtering the speech signal to about 1 KHz, down sampling and performing an inverse filtering of the resultant signal. This signal is then autocorrelated and a peak-picking method is employed as in the above two approaches. The parallel processing method (PPROC) of Rabiner et al [75] is yet another method of pitch period determination, which is very tedious.

Atal and Schroeder [44] have considered the periodicity of the normalised correlation coefficients, for evaluating the pitch-period. Sondhi [76] has used autocorrelation coefficients (ACF) of the centre-clipped signal samples. Amano et al [29] have reported that normalised autocovariance function is a better parameter. But the author has found that by just checking the periodicity of the peak positions of the ACF's, formed from the direct samples, we obtain exactly the same results as those got from the methods of Sondhi and Atal and Schroeder. Care was taken not to use overlapping windows while segmenting the data samples. This was done to reduce the chances and hence the effect of including the different regions in the same block, while performing the detection of the relevant region in the block.

The results of the work done in this direction are shown in Table 3.1. For each block under consideration, the normalised correlation coefficients ρ (J), the ACF's R(J) of the original samples and the ACF's R_c(J) of the centre-clipped samples [76] were evaluated. In the centre-clipping method of Sondhi [76], within each block, the maximum absolute value A_0 of the samples is found every 5 ms (here, every 40 samples, as the sampling frequency used is 8 KHz), and all portions of the signal, between $+0.3A_0$ and $-0.3A_0$ were removed, to obtain the clipped samples. ACF's of such 160 centre-clipped samples were computed.

The correlation coefficients are given by the following equations.

$$\rho(J) = \sum_{n=J+1}^{N} s_n s_{n-J} \Big/ \left\{ \left(\sum_{n=1}^{N} s_n^2 \right) \cdot \left(\sum_{n=J+1}^{N} s_{n-J}^2 \right) \right\}^{\frac{1}{2}} \qquad (3.24)$$

$$R(J) = \sum_{n=J+1}^{N} S_n S_{n-J} \qquad (3.25)$$

$$R_c(J) = \sum_{n=J+1}^{N} S'_n S'_{n-J} \qquad (3.26)$$

where, S'_n is the clipped samples and N = block length in samples = 160. (The reason for choosing 160 samples is explained in section 4.2.1).

Under natural speaking conditions, the minimum pitch period is 2 msec [43]. For a sampling rate of 8 KHz, this value corresponds to 16. Hence the value of M was determined by locating the maximum value of the correlation coefficients, for $J > 15$. To check for the periodicity of the correlation coefficients and also to check the consecutive pitch periods of the signal being processed, the positions of the first, second and third maxima were found with respect to all the functions—$\rho(J)$, R(J) and $R_c(J)$. The number of samples within adjacent peak positions were noted as M1, M2 and M3 (as shown in Fig.3.6). It can be seen from Table 3.1 that the values of M1, M2 and M3 obtained from all the above three evaluation methods are identical. While using these above values for the detection of the voiced/unvoiced/silent/transition regions in the speech signal (explained in chapter 4), the voiced regions were found to be detected correctly.

Computational Savings

Considering the computational savings achieved in this case, it

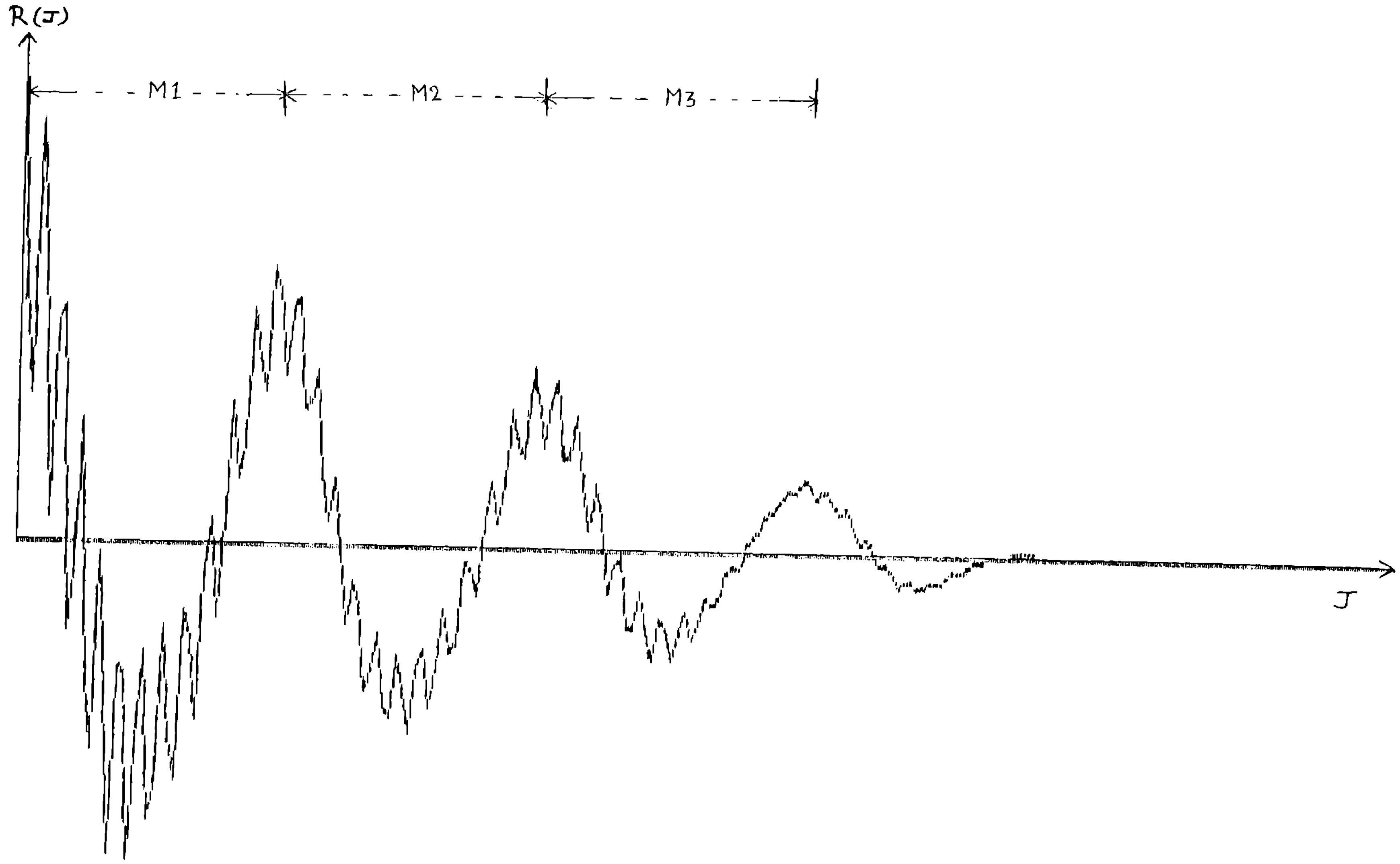


Fig.3.6   R(J) versus J plot to determine pitch positions.

Table 3.1

Pitch number of samples obtained from the different correlation functions, for voiced speech segments

Data files	Segments	No. of samples within peak positions, obtained from								
		$R_c(J)$			ρ (J)			R(J)		
		M1	M2	M3	M1	M2	M3	M1	M2	M3
F1	1	32	33	32	32	32	33	32	32	33
	2	32	32	32	32	32	32	32	32	32
	3	35	35	35	35	35	35	35	35	35
	4	35	34	35	35	34	35	35	34	35
	5	36	36	36	36	36	36	36	36	36
M1	1	38	39	38	38	39	38	38	39	38
	2	41	42	41	41	42	41	41	42	41
	3	39	38	39	39	38	39	39	38	39
	4	51	50	51	51	50	52	51	50	--
	5	54	54	--	54	54	--	54	54	--
B3	1	46	45	46	46	45	46	46	45	46
	2	38	38	38	38	38	38	38	38	38
	3	49	49	49	49	49	49	49	49	49
	4	47	47	47	47	47	47	47	47	47
	5	51	50	--	51	50	52	51	50	51

Note: F1, M1, B3 are different speech files, specified in Tables 4.1 and 4.2. Number of samples in a block = N=160.

can be noted that, compared to the normalised correlation coefficient method, the use of the ACF's saves (2N+1) multiplications, (2N-2) additions, 1 square rooting and 1 division, for each value of J, where N is the total number of samples in the block under consideration. This saves a good amount of computational burden. The centre-clipping method needs an additional amount of computation, at the beginning of each block, which comes on an order of N additions or subtractions (computational values arrived at are shown in Appendix II).

3.9.1.2 Covariance/Autocorrelation Matrix Evaluation

The next step in the direction of reducing the computational burden, was to see if the covariance matrix could be reduced to an autocorrelation matrix, in the evaluation of the predictor coefficients.

In the distant-sample based prediction method (pitch prediction method), the matrix elements are computed from $V_n = S_n - \beta S_{n-M}$. Since S_{n-M} is the sample most correlated with S_n and β is approximately equal to 1, the value of $\{V_n\}$ is very small, and theoretically it makes no difference if the covariance matrix is replaced by the autocorrelation matrix. But it reduces much of the computational work.

The distant sample based algorithm (of equation 3.17) was tried on both voiced and transition segments, using both the autocorrelation method and the covariance method, for various values of p, the order of the predictor, and the average SNR and SNRSEG values were computed. The effect of the spectrum predictor alone (of equation 3.4) was also studied.

It has been noted that the pitch period of a voiced sound remains constant for 3 to 4 consecutive pitch positions (as seen from the values of M1, M2, M3 in Table 3.1). Hence, the block length was taken as 4M, since it is always better to choose a block length relative to the pitch period of the signal being processed, rather than choosing a constant block length of 5 or 10 msecs, as hitherto done. It was assumed that the predicted value of the first (P+M) samples is the same as that of the original samples. For n > (P+M), the earlier predicted values were used. The difference signal is considered unknown. The above same processes were done using a block length equal to 3M also. The results of the work done in this direction are shown in Tables 3.2 and 3.3.

From Table 3.2(a), it can be noted that, using pitch prediction, for voiced segments, the value of SNRSEG varies from 4.82 dB to 12.95 dB (or SNR varies from 4.98 to 16.1 dB), for the various data files and various speakers, as the order of the predictor p, is increased from 4 to 12. Also, when covariance method was used for the evaluation of the predictor coefficients, actually the SNRSEG value decreased, though only marginally by about 0.2 dB to 0.3 dB. In some cases, the predictor became highly unstable and the predictor coefficients went to ± 5 and above, and the algorithm did not work (It has been reported [2] that the covariance method gives better and steadier SNR values than the autocorrelation method). Table 3.2(b) shows the result of spectrum prediction on voiced

Table 3.2(a)
Comparison of the SNR values obtained by using pitch prediction on voiced segments

Data files	P	Autocorrelation method				Covariance method	
		N = 3M		N = 4M		N = 4M	
		Av.SNR	Av.SNRSEG	Av.SNR	Av.SNRSEG	Av.SNR	Av.SNRSEG
F1	4	12.818	11.129	9.937	9.100	9.495	8.662
	6	13.537	11.854	10.130	9.283	8.776	9.594
	8	13.817	12.159	10.241	9.662	8.940	9.558
	10	13.822	13.820	10.327	9.859	8.949	9.565
	12	14.114	12.513	10.438	9.506	8.800	9.500
M1	4	6.066	5.160	4.981	4.822	5.007	4.838
	6	6.086	5.168	5.190	5.004	5.037	4.869
	8	6.012	5.186	5.356	5.218	5.129	4.998
	10	6.245	5.301	5.957	5.856	5.511	5.430
	12	6.651	5.613	8.019	7.932	5.310	5.350
B3	4	16.563	15.142	14.900	12.570	9.560	8.546
	6	16.836	15.444	15.827	12.712	15.580	12.430
	8	17.052	15.703	16.101	12.839	15.646	12.470
	10	17.005	15.631	15.918	12.884	15.779	12.656
	12	16.981	15.907	15.527	12.949	10.751	9.602

Table 3.2 (b)

Comparison of the SNR values obtained by using Spectrum Prediction on Voiced Speech Segment

Speech file	P	Autocorrelation method		Covariance method	
		N = 40		N = 40	
		Av.SNR	Av.SNRSEG	Av.SNR	Av.SNRSEG
F1	4	0.783	0.7139	0.790	-.421
	6	1.521	1.398	1.503	1.770
	8	2.328	2.078	--	--
	10	3.474	2.936	1.451	1.447
	12	5.028	3.657	1.730	2.025
M1	4	0.707	0.670	Obtained only lower values wherever the algorithm worked. In many cases, the system does not work, as the prediction coefficient values shoot upto ±5 and above.	
	6	1.133	1.086		
	8	2.073	1.862		
	10	2.249	1.997		
	12	3.115	2.550		
B3	4	1.021	0.929		
	6	1.553	1.414		
	8	4.628	3.577		
	10	5.575	4.570		
	12	7.753	6.122		

Table 3.3(a)

Comparison of the SNR values obtained by using pitch prediction on transition segments

Speech files	P	Autocorrelation method				Covariance method	
		N = 3M		N = 2M		N = 2M	
		Av.SNR	Av.SNRSEG	Av.SNR	Av.SNRSEG	Av.SNR	Av.SNRSEG
F1	4	9.697	9.420	11.546	11.124	11.041	10.858
	6	9.940	9.731	11.962	11.583	11.032	10.949
	8	10.389	10.261	12.786	12.378	12.749	12.515
	10	10.354	10.250	13.145	12.586	12.750	11.709
	12	10.442	10.341	15.086	14.445	12.900	11.960
M1	4	4.810	5.701	5.900	6.500	7.813	6.779
	6	6.123	6.231	7.501	8.001	7.659	6.914
	8	7.935	7.898	8.598	8.396	7.721	6.771
	10	8.712	8.599	9.386	9.130	—	—
	12	9.351	9.112	10.114	9.732	8.489	7.252
B3	4	11.807	10.568	11.232	10.560	11.791	10.267
	6	11.881	10.473	13.594	12.423	12.751	14.154
	8	12.150	10.532	14.139	12.982	12.344	12.521
	10	12.352	10.824	14.316	13.192	12.150	12.980
	12	12.351	10.910	15.203	13.789	9.601	12.010

Table 3.3(b)

Comparison of the SNR values obtained by using spectrum prediction on transition segments

Speech files	P	Autocorrelation method	
		N = 40	N = 40
		Av. SNR	Av. SNRSEG
F1	4	0.6109	0.4924
	6	2.0380	1.7838
	8	3.1924	2.6155
	10	3.6936	3.1625
	12	4.8313	4.3043
M1	4	2.4915	2.1622
	6	3.4687	2.9835
	8	5.4610	4.4753
	10	6.8182	5.5411
	12	7.2321	5.9598
B3	4	2.4855	2.1806
	6	3.3194	2.7500
	8	4.4572	3.8858
	10	6.3400	5.3730
	12	7.0958	5.9886

Table 3.4

Comparison of the SNR values obtained by using spectrum prediction on unvoiced segments

Speech files	P	Autocorrelation method		Covariance method	
		N = 40		N = 40	
		Av. SNR.	Av.SNRSEG	AV. SNR	Av.SNRSEG
F1	4	0,7089	0.6269	-9.8790	0.2491
	6	1.1016	1.0242	-9.5520	0.3880
	8	1.5320	1.4598	-9.0792	1.0011
	10	2.5568	2.3733	0.7093	0.9601
	12	3.2550	2.9366	0.9112	0.6732
M1	4	1.2890	1.0768	--	--
	6	2.2217	1.6753	-6.4130	0.2430
	8	5.1197	3.2377	1.1606	0.6468
	10	5.9155	4.0600	1.8100	0.9501
	12	8.6048	5.1051	1.0601	1.0610
B3	4	0.5020	0.4937	--	--
	6	0.9608	0.9030	0.4650	0.3601
	8	1.5041	1.4042	1.0515	0.2172
	10	1.9021	1.8213	1.0671	0.2869
	12	2.2594	2.1624	1.0122	0.5241

segments. Comparing Tables (3.2a) and (3.2b), it can be seen that the introduction of a pitch predictor increases the SNRSEG value by around 8 to 9 dB.

Similar results are obtained for the transition segments also (Table 3.3). For unvoiced regions, the correlation factor β tends to zero and the pitch prediction algorithm is reduced to the spectrum prediction algorithm. Here also (Table 3.4), the autocorrelation method turned out to be better, in terms of both SNR value and stability of the filter.

In the transition regions, the changes in the sample values is abrupt and hence, for transition segments, the processing block length is chosen to be N = 2M, while for voiced segments it is 4M and for unvoiced regions, it is taken as 40 (that is, 5 msecs duration).

It was further noted that, for unvoiced segments, SNR increases as p increases and p = 12 is optimum, in an SNR sense. For voiced regions, the increase in SNR for values of p equal to, and above 4, was only marginal (about 0.2 dB). Hence p = 4 is chosen as optimum. Similarly, p = 8 seems to be optimum for transition regions. Thus, a switching between the two algorithms - pitch prediction and spectrum prediction - is to be done when the speech segments change from the voiced or transition regions to the unvoiced region. Silent region needs no processing.

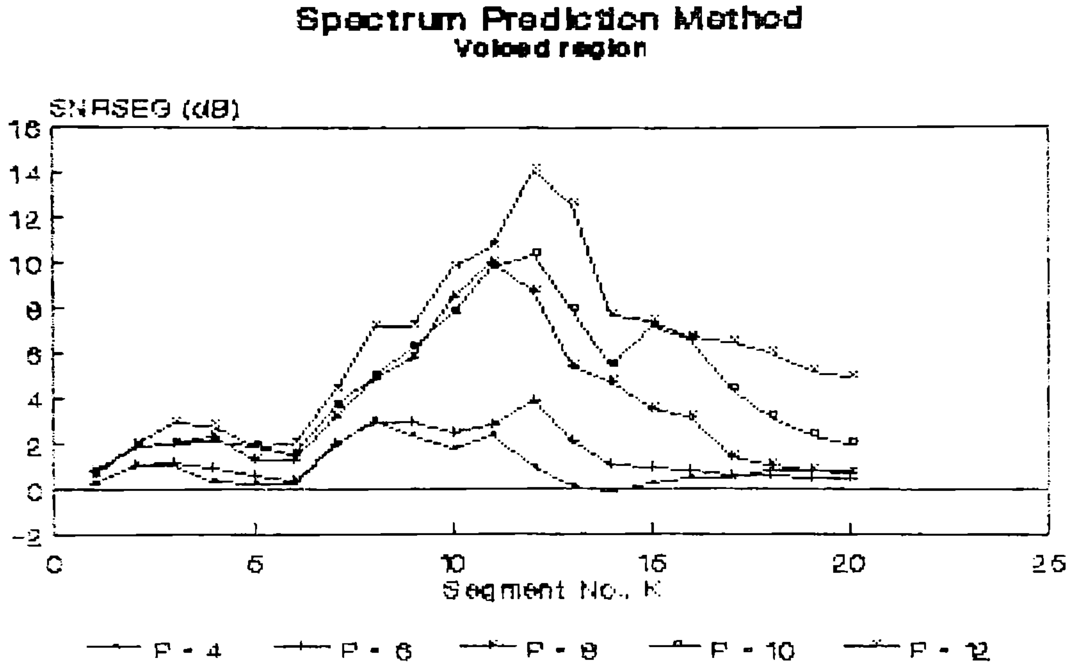


Fig.3.7 (a)

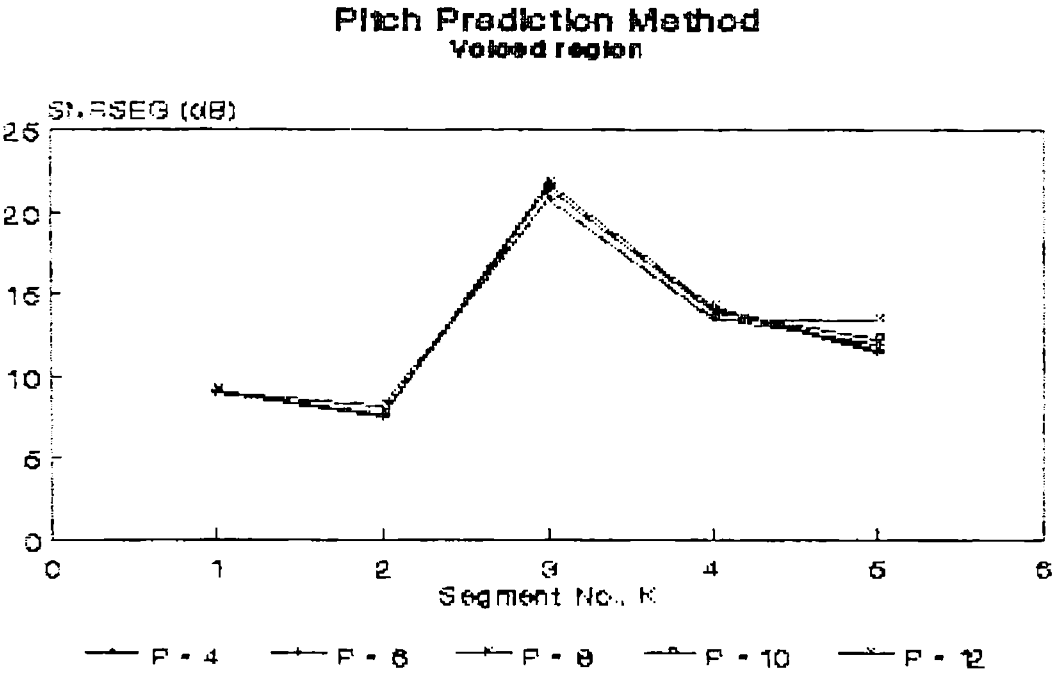


Fig.3.7(b)

Fig.3.7(a–j) Performance of a Predictive Coder with respect to various parameters in different regions of the speech signal.

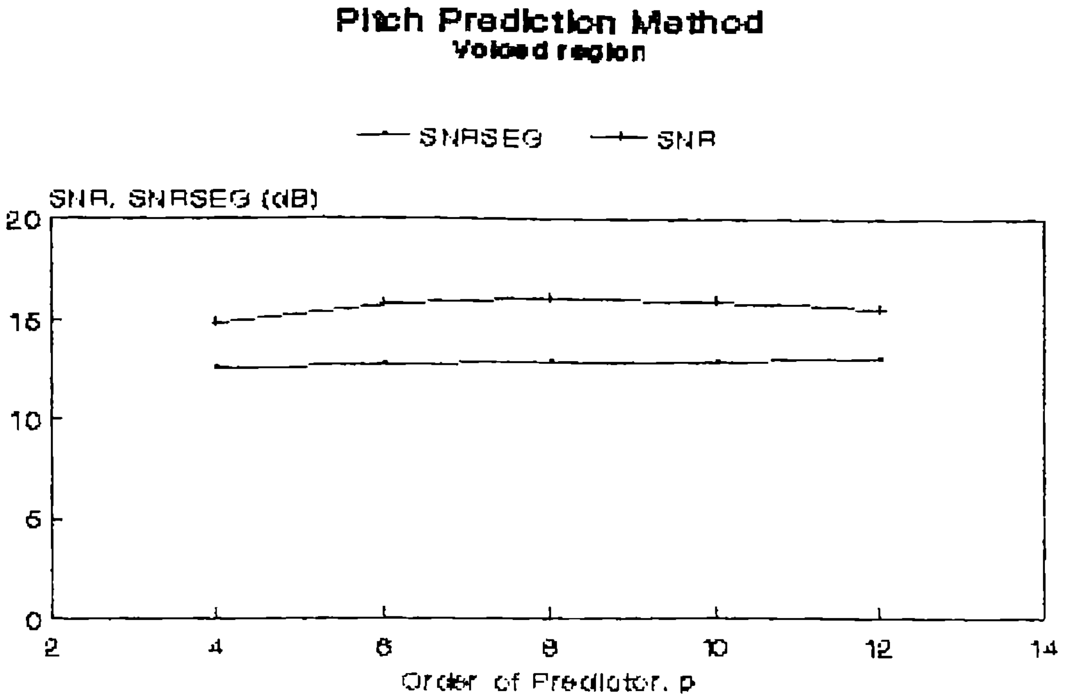


Fig.3.7 (c)

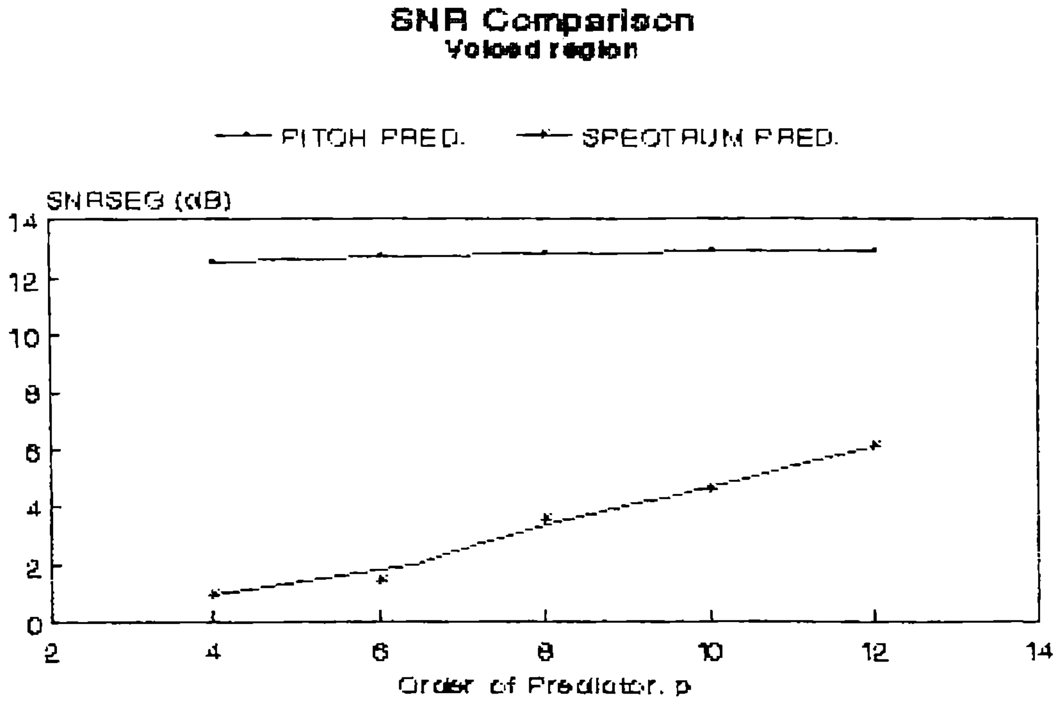


Fig.3.7 (d)

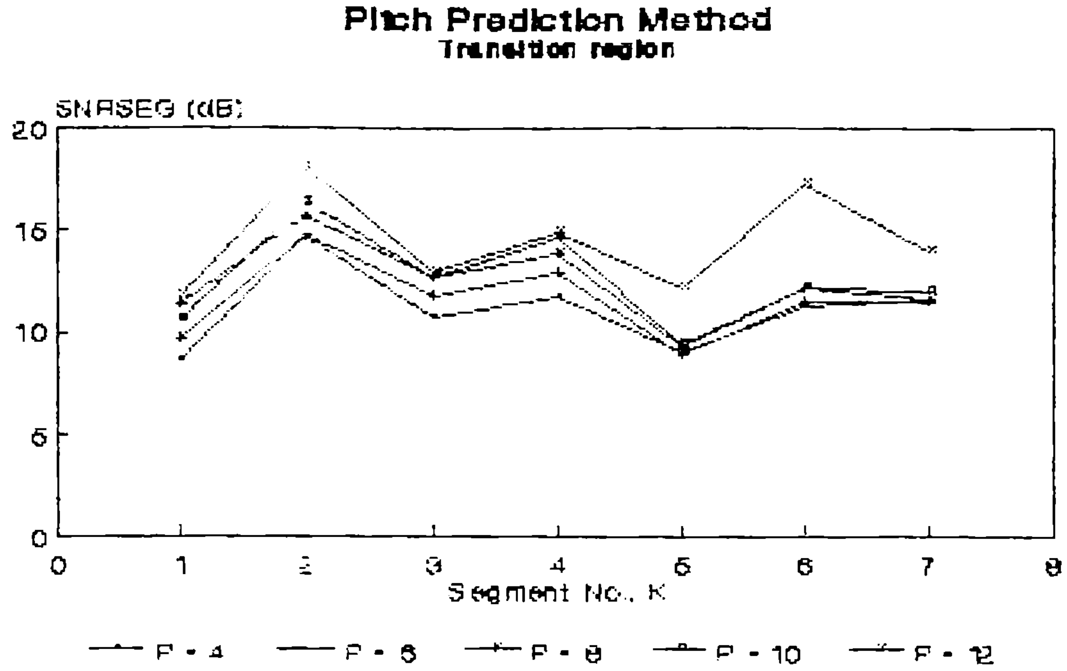


Fig.3.7 (e)

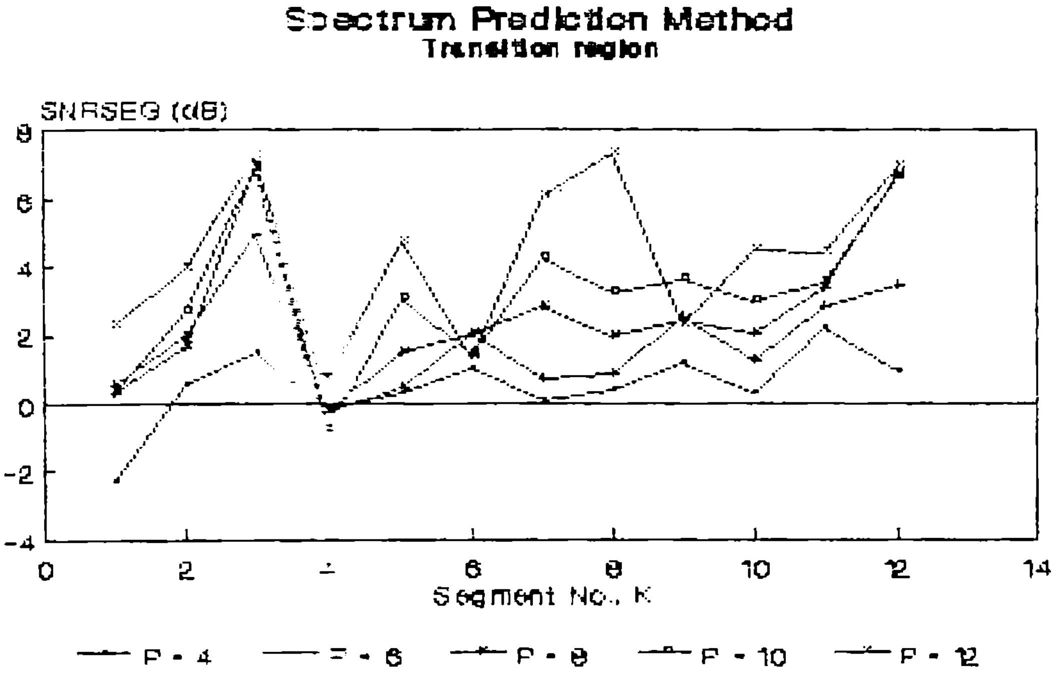


Fig.3.7 (f)

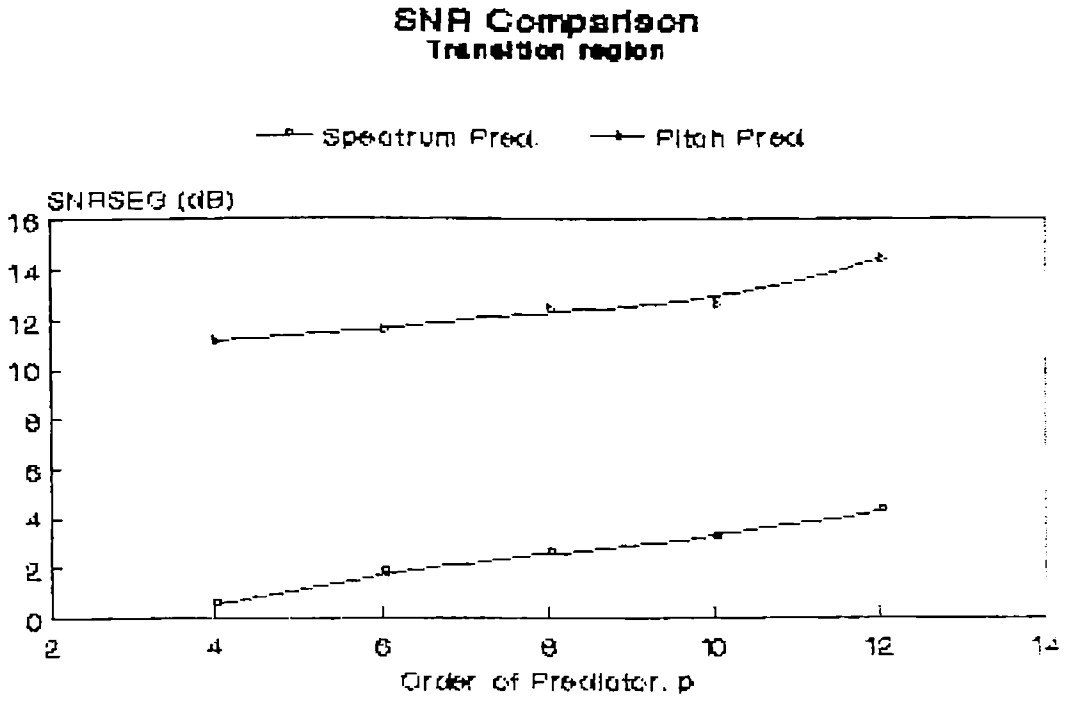


Fig.3.7 (g)

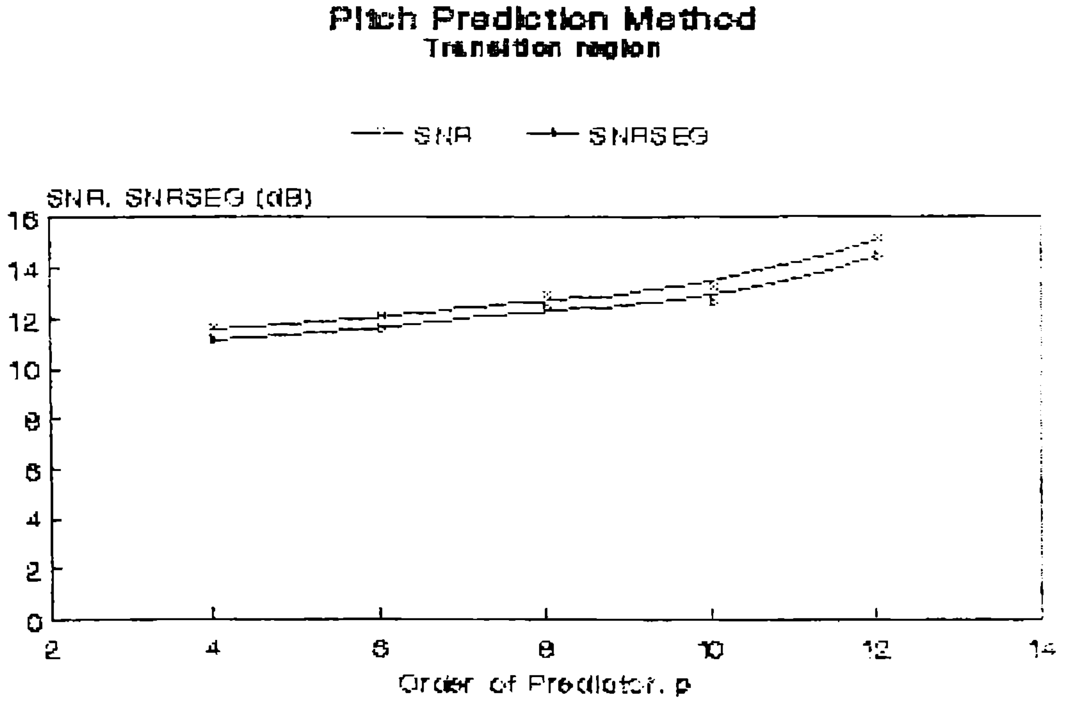


Fig.3.7 (h)

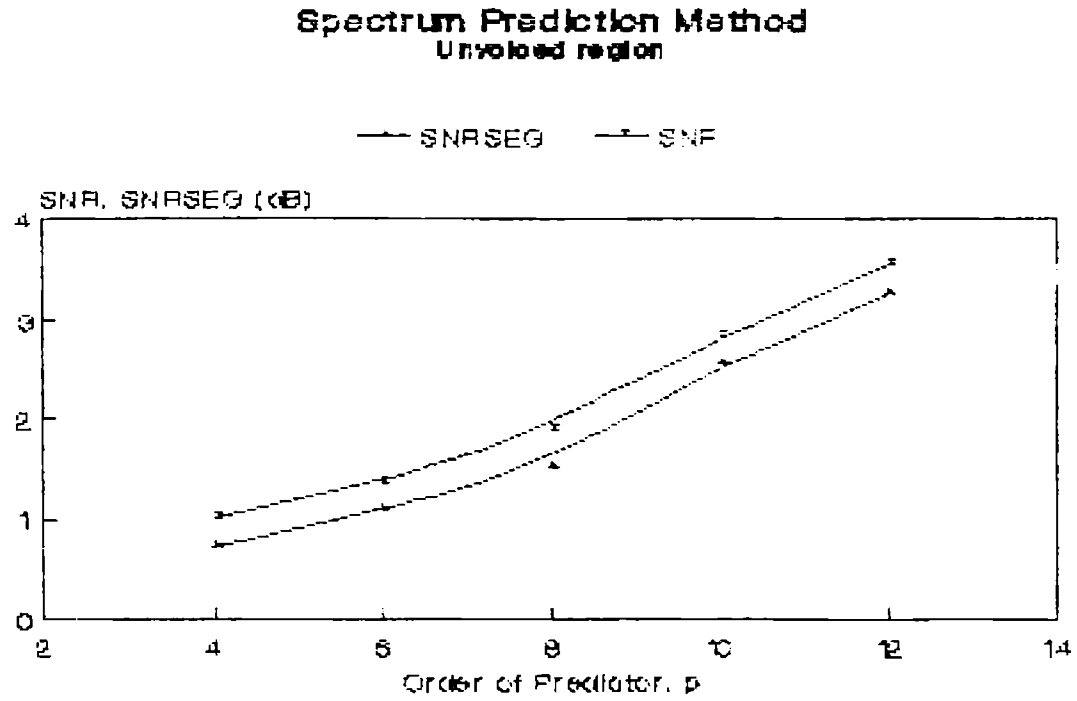


Fig.3.7 (i)

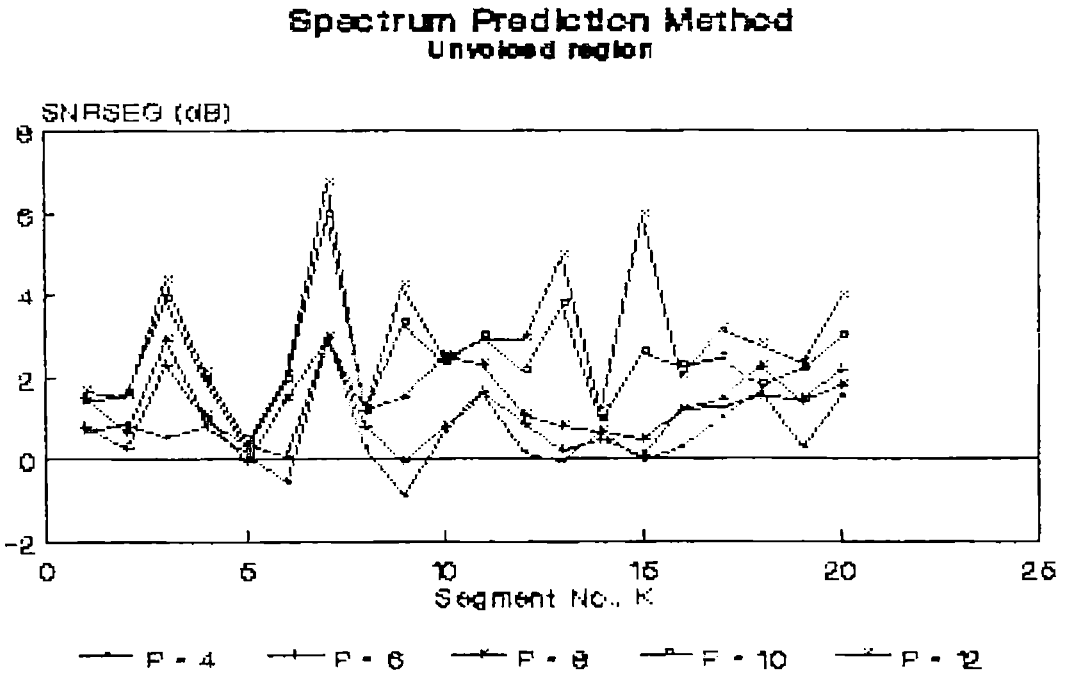


Fig.3.7 (j)

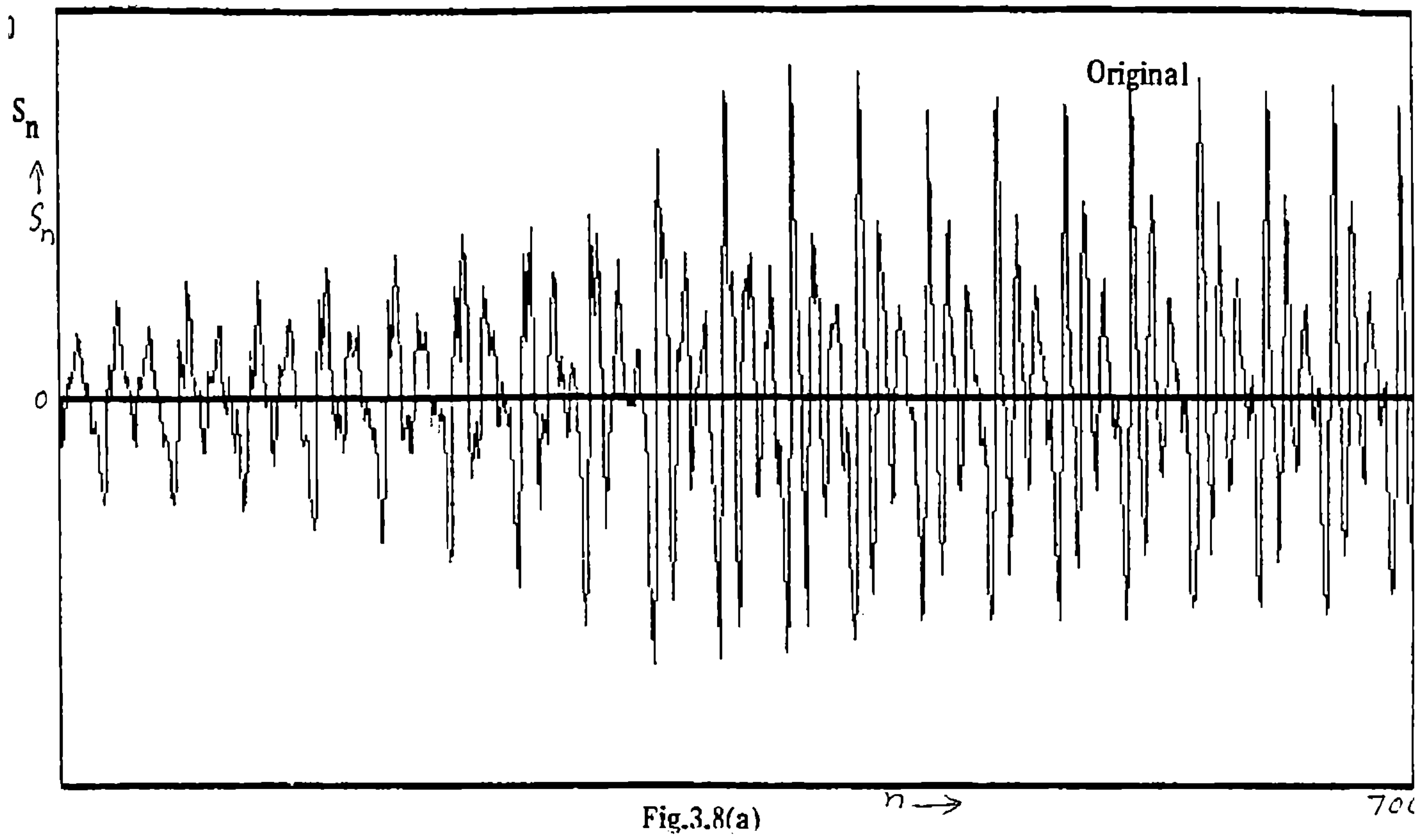


Fig.3.8(a)

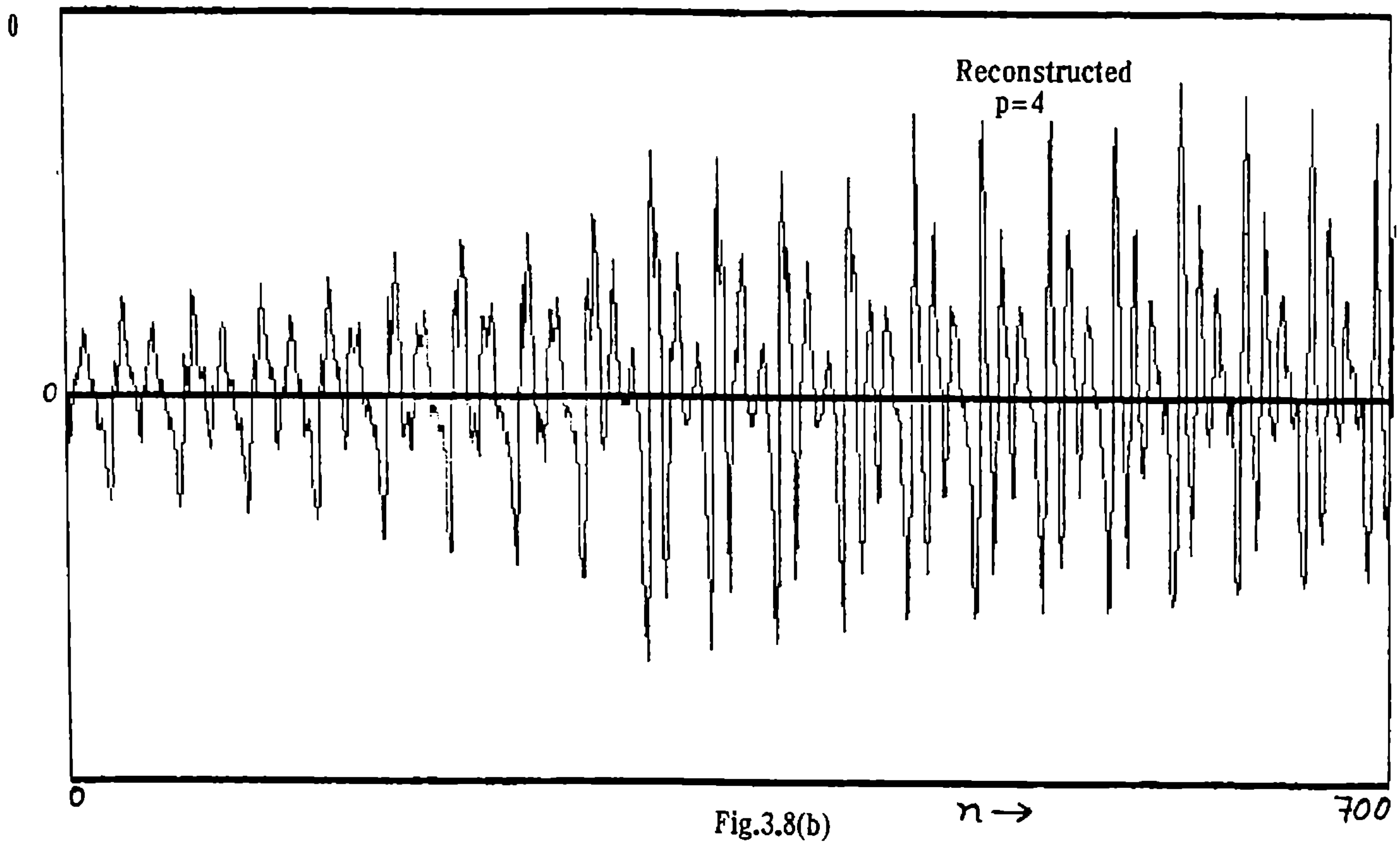


Fig.3.8(b)

Fig.3.8(a–d) Plots of original and reconstructed waveforms – Using pitch prediction – Voiced region.

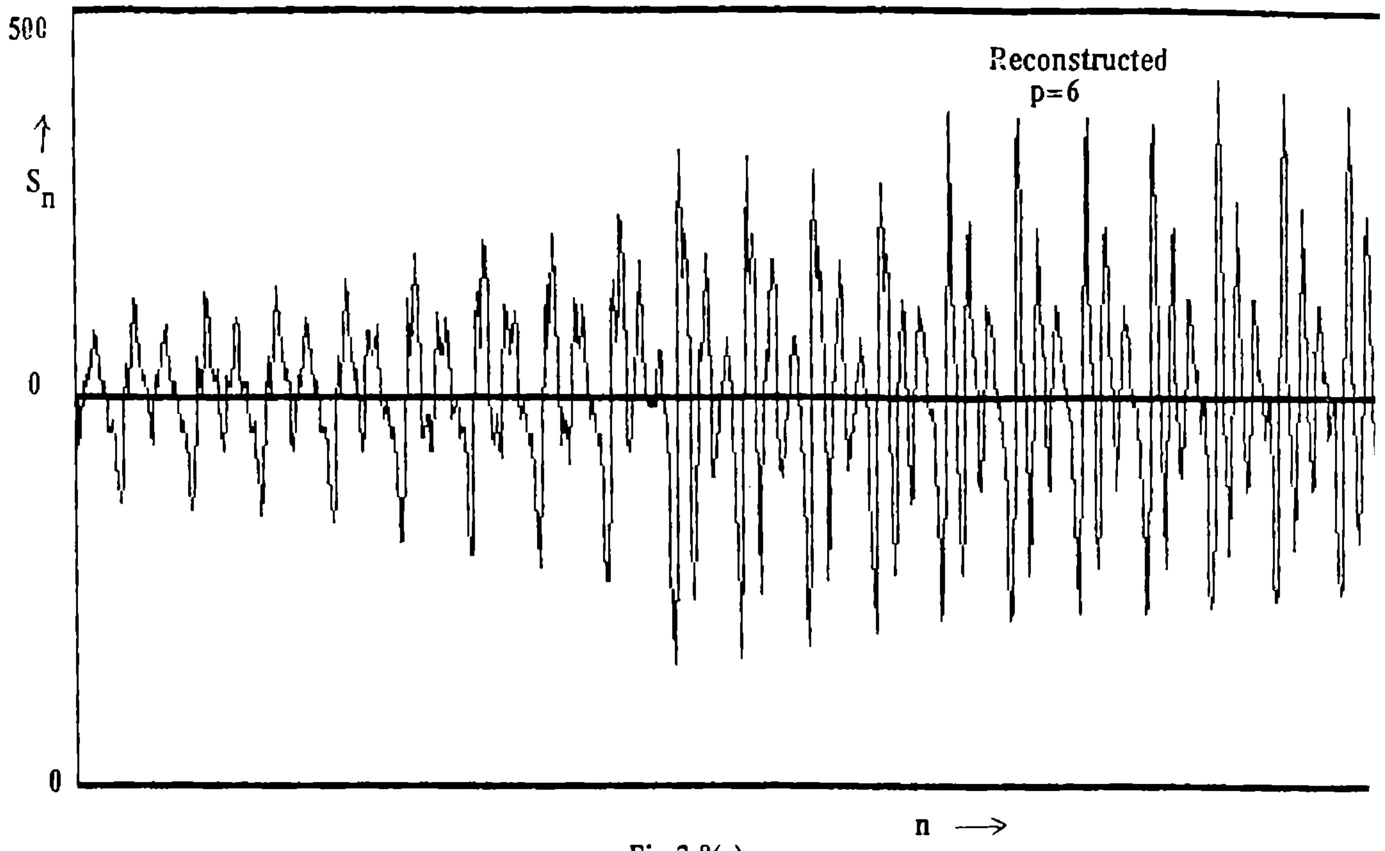


Fig.3.8(c)

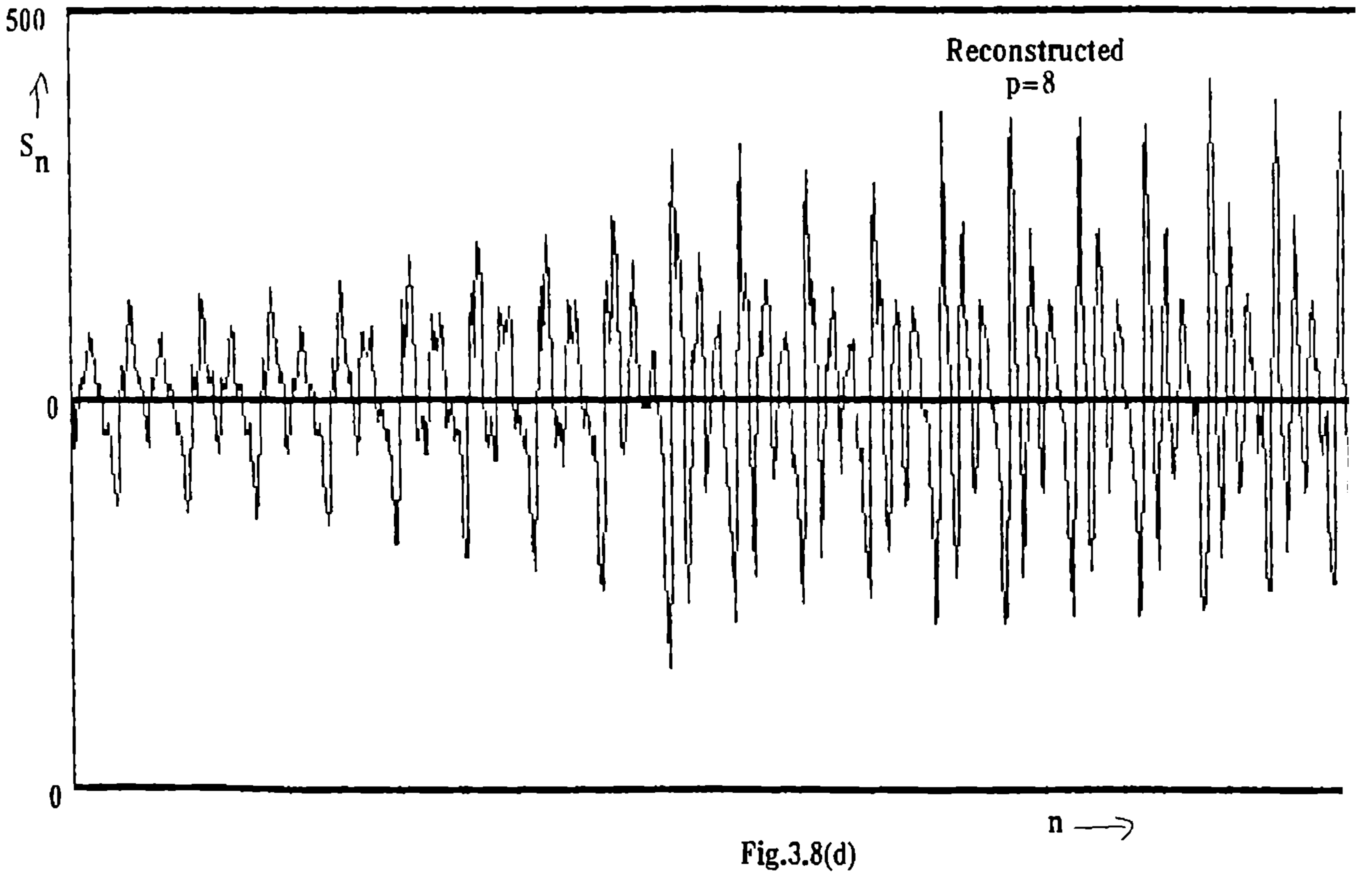


Fig.3.8(d)

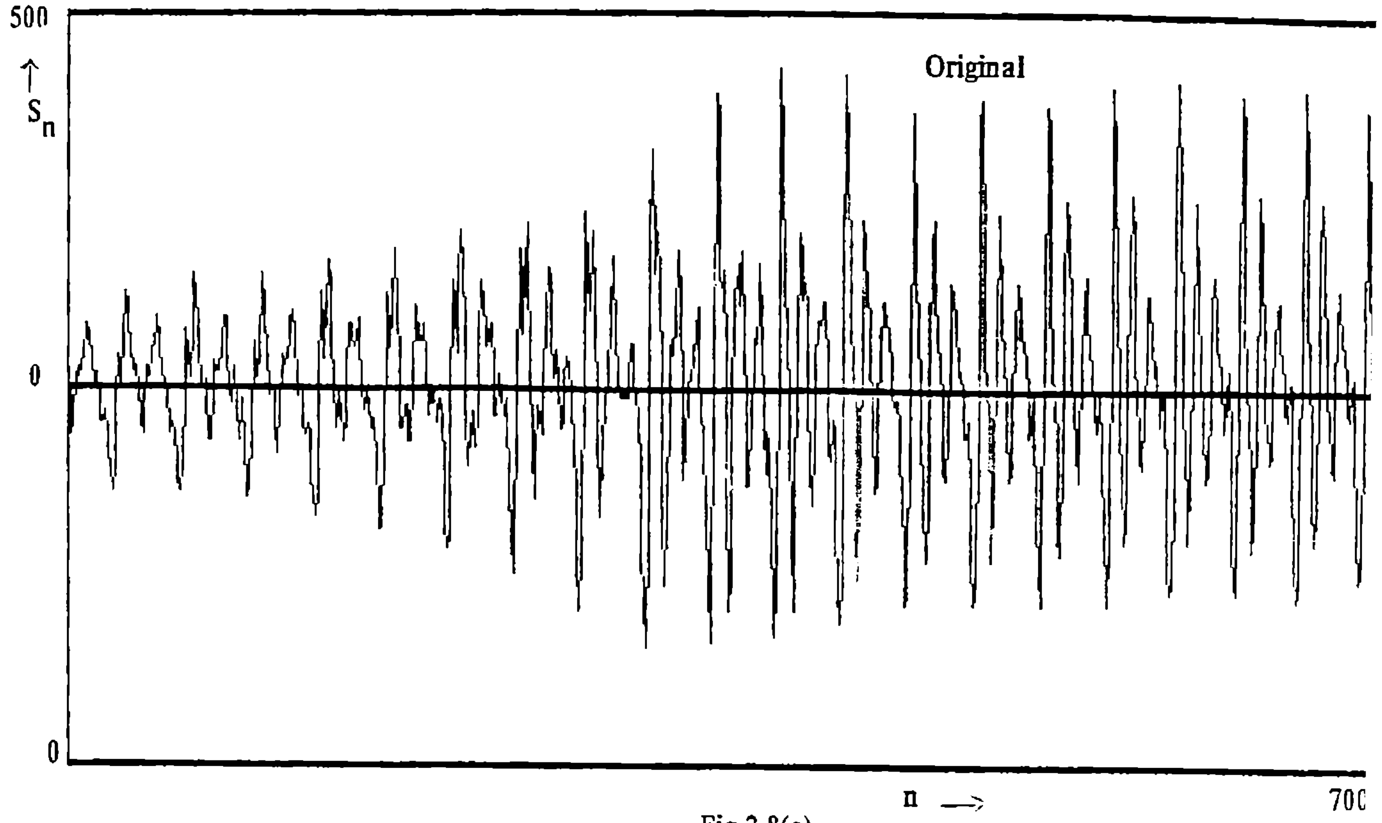


Fig.3.8(e)

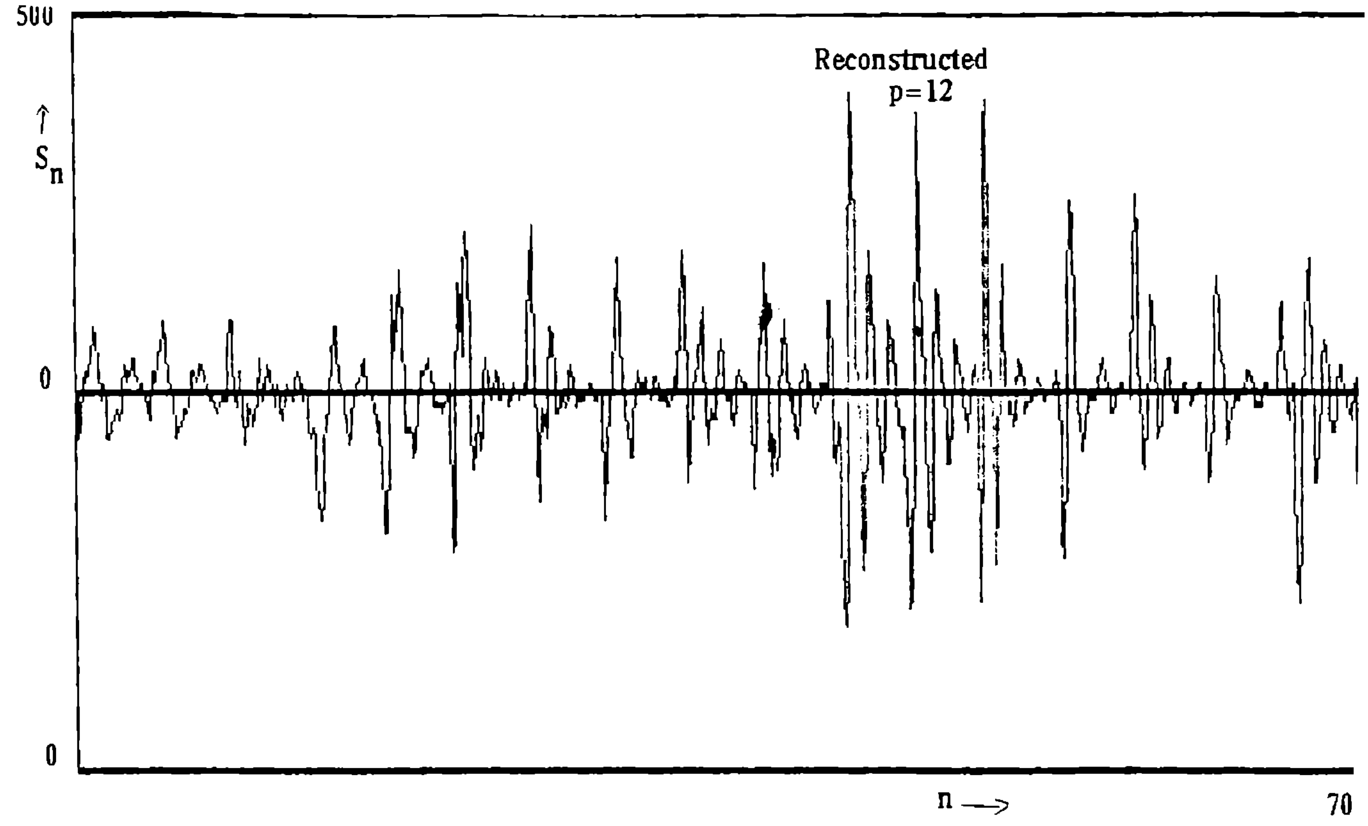


Fig.3.8(f)

Fig.3.8 (e–h) Plots of original and reconstructed waveforms – Using spectrum prediction – Voiced region.

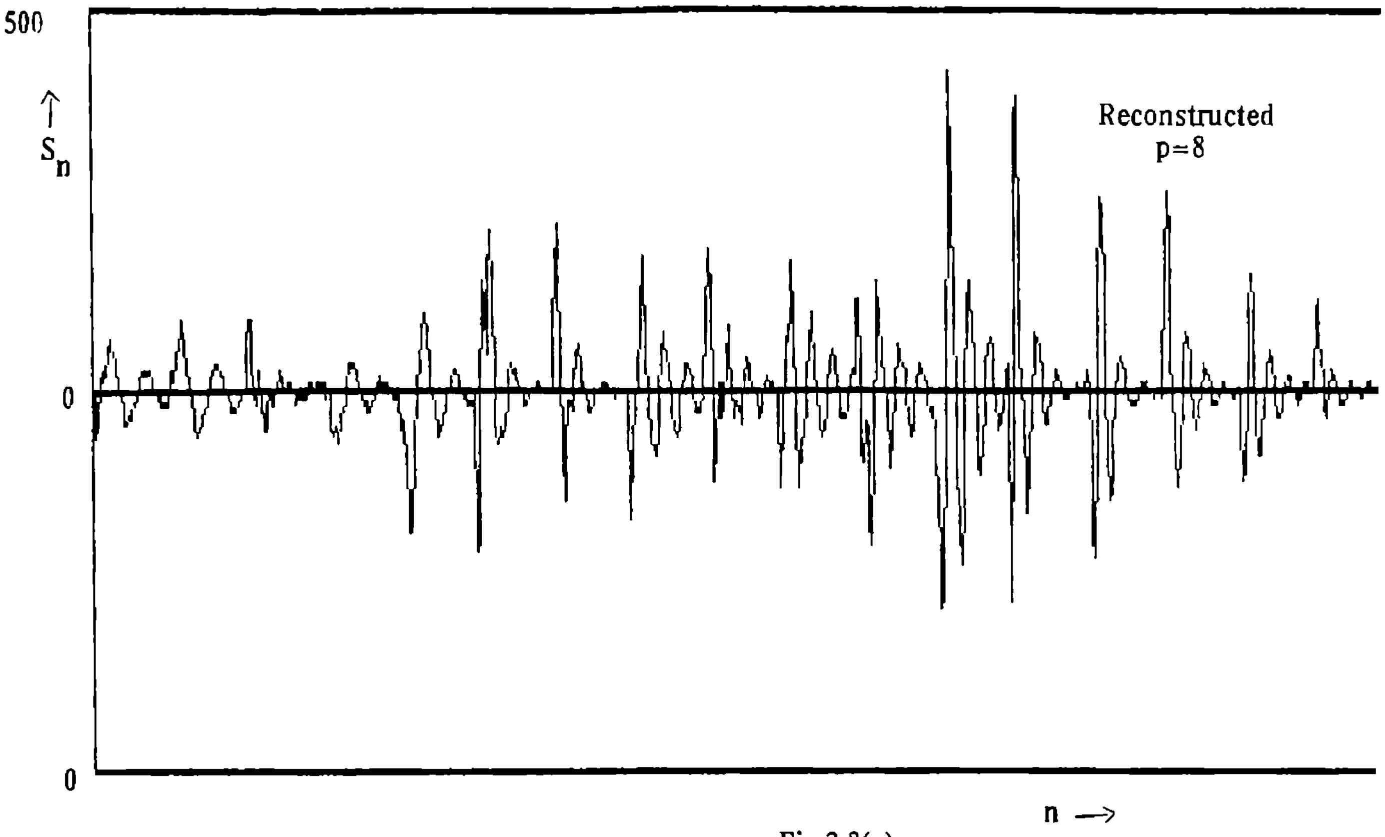


Fig.3.8(g)

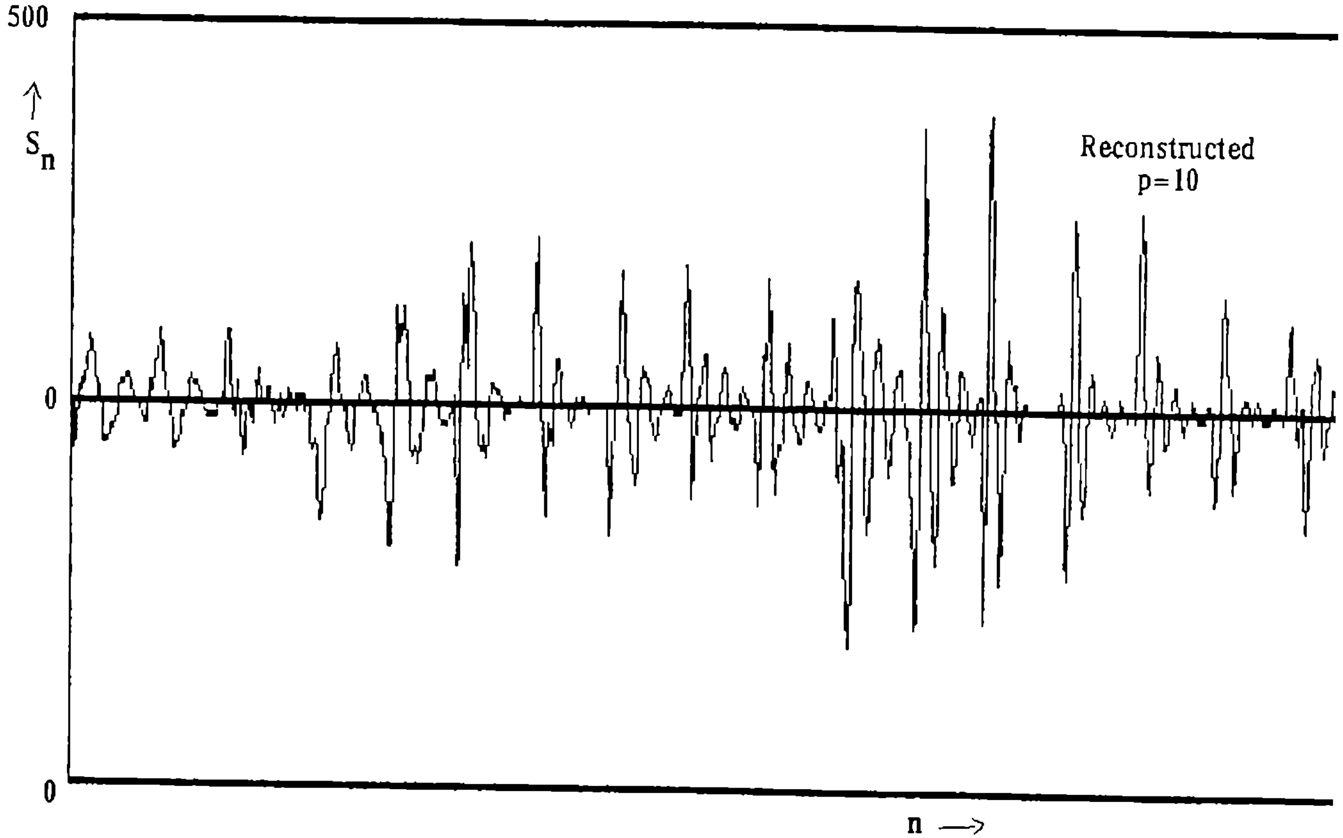


Fig.3.8(h)

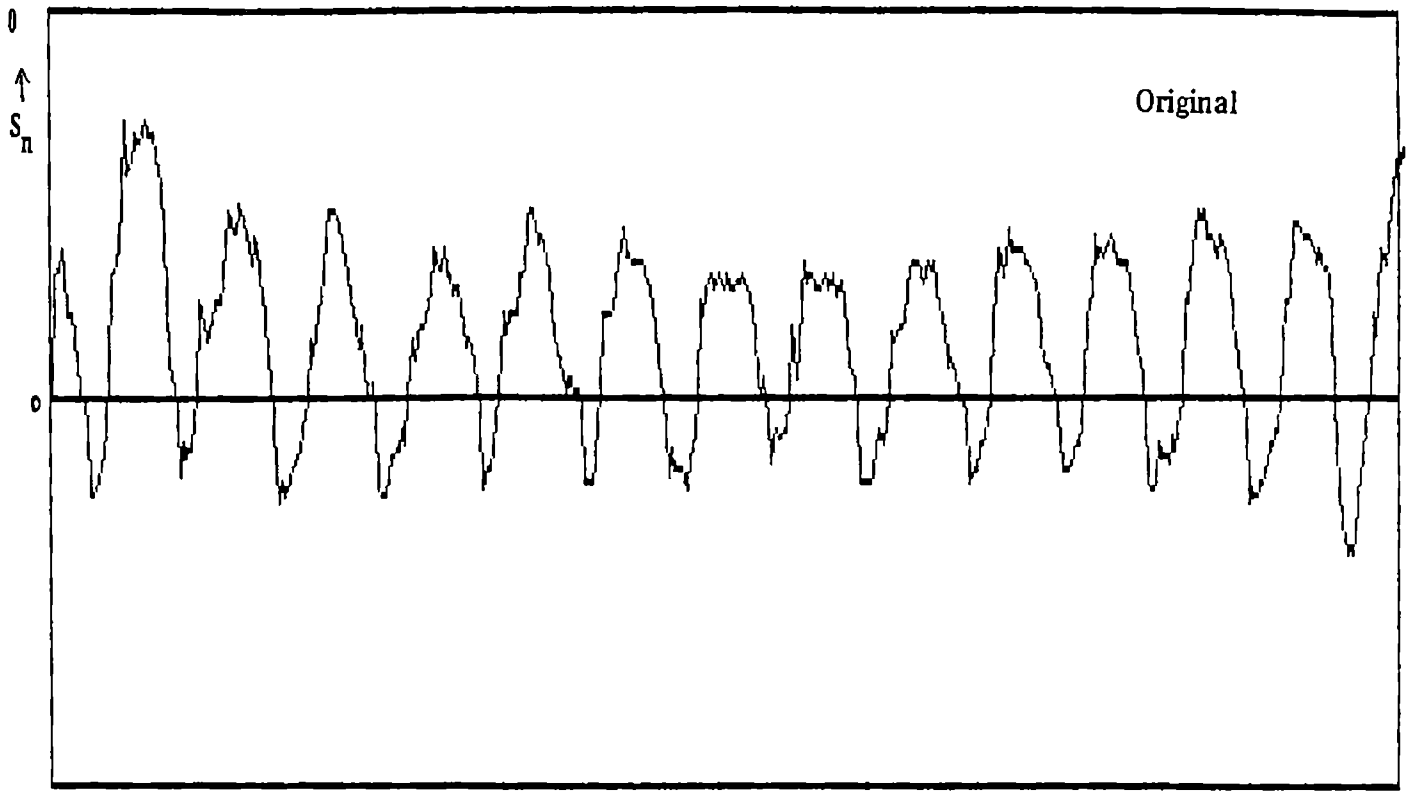


Fig.3.9(a)

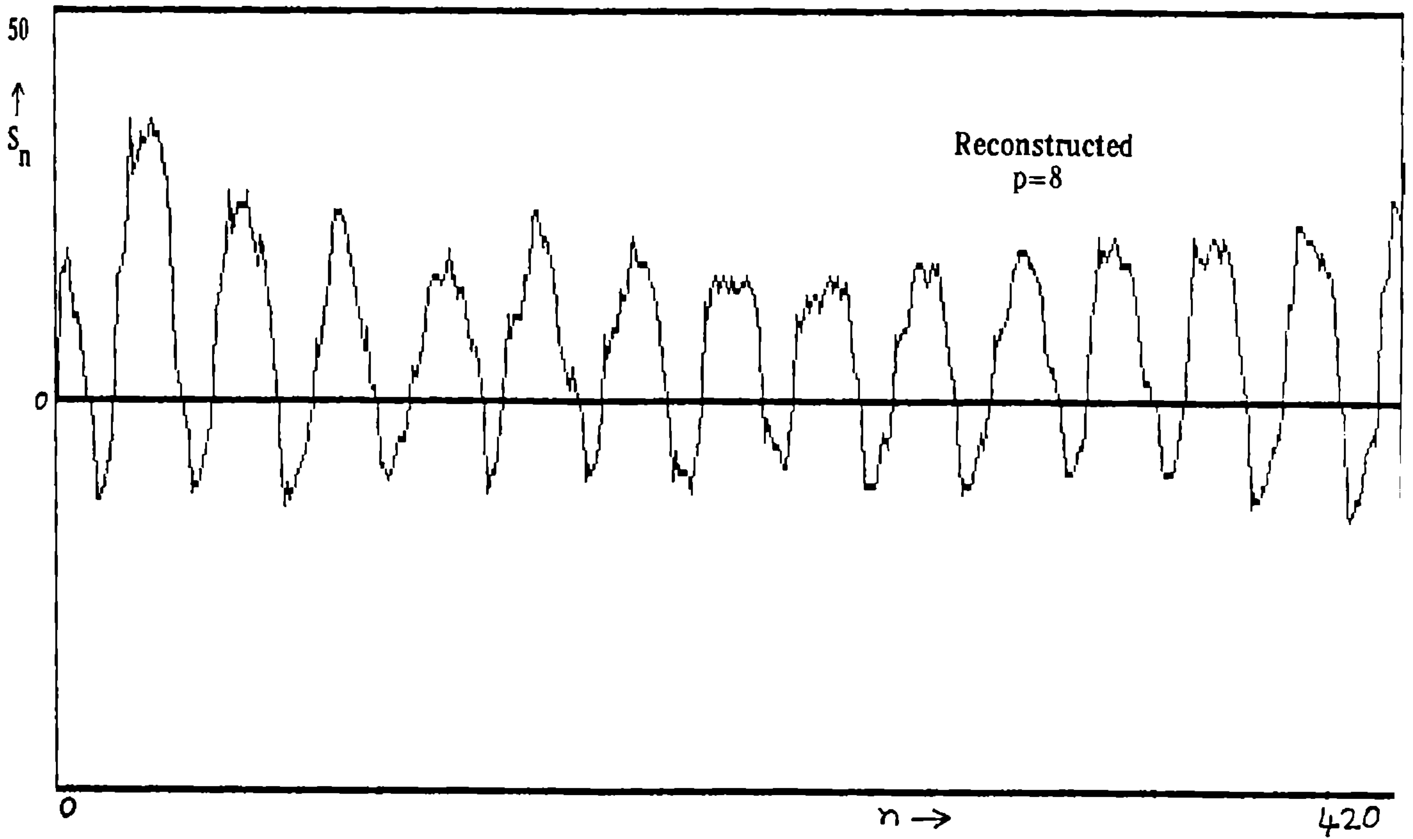


Fig.3.9(b)

Fig.3.9 (a–d) Plots of original and reconstructed waveforms – Using pit prediction – Transition region.

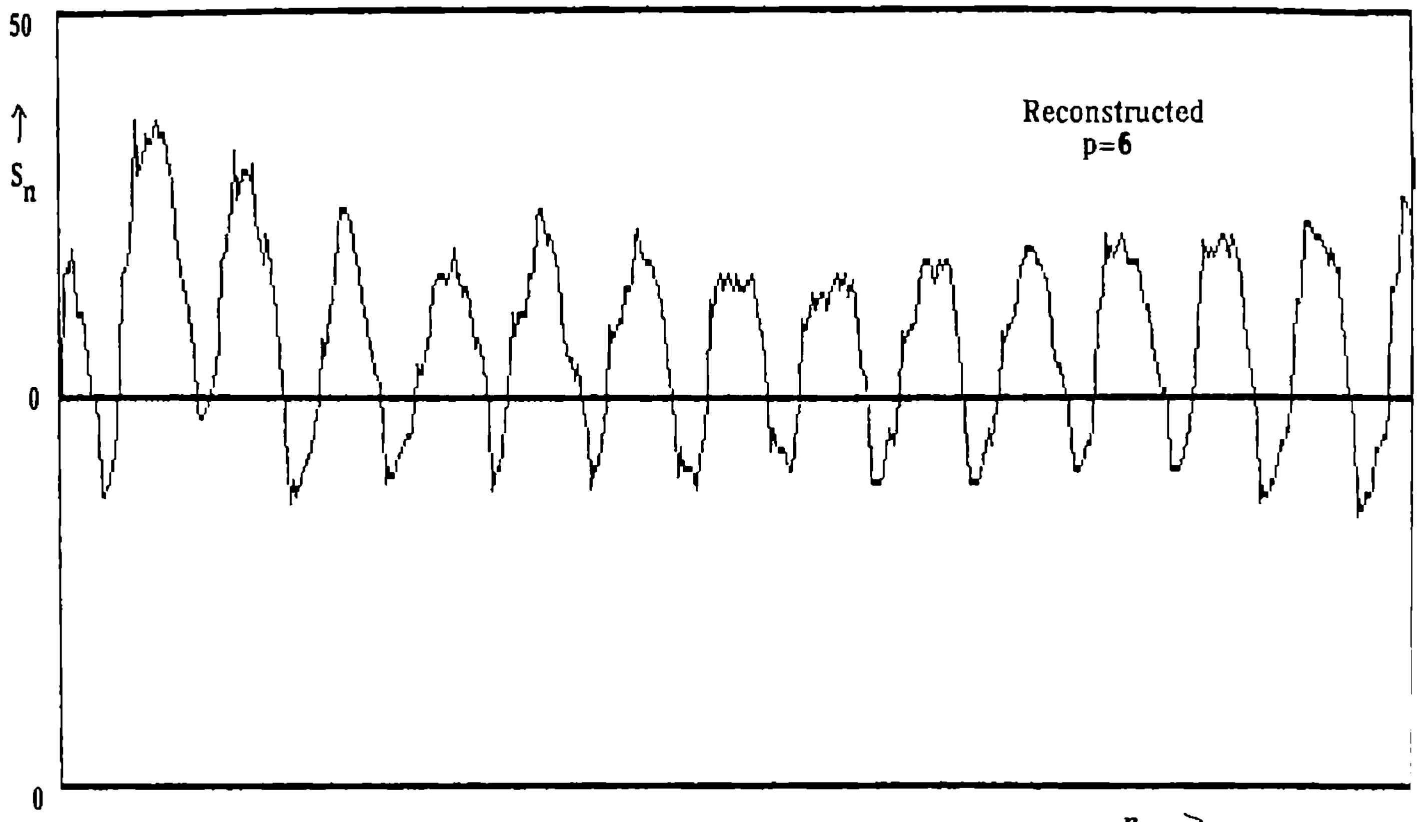


Fig.3.9(c)

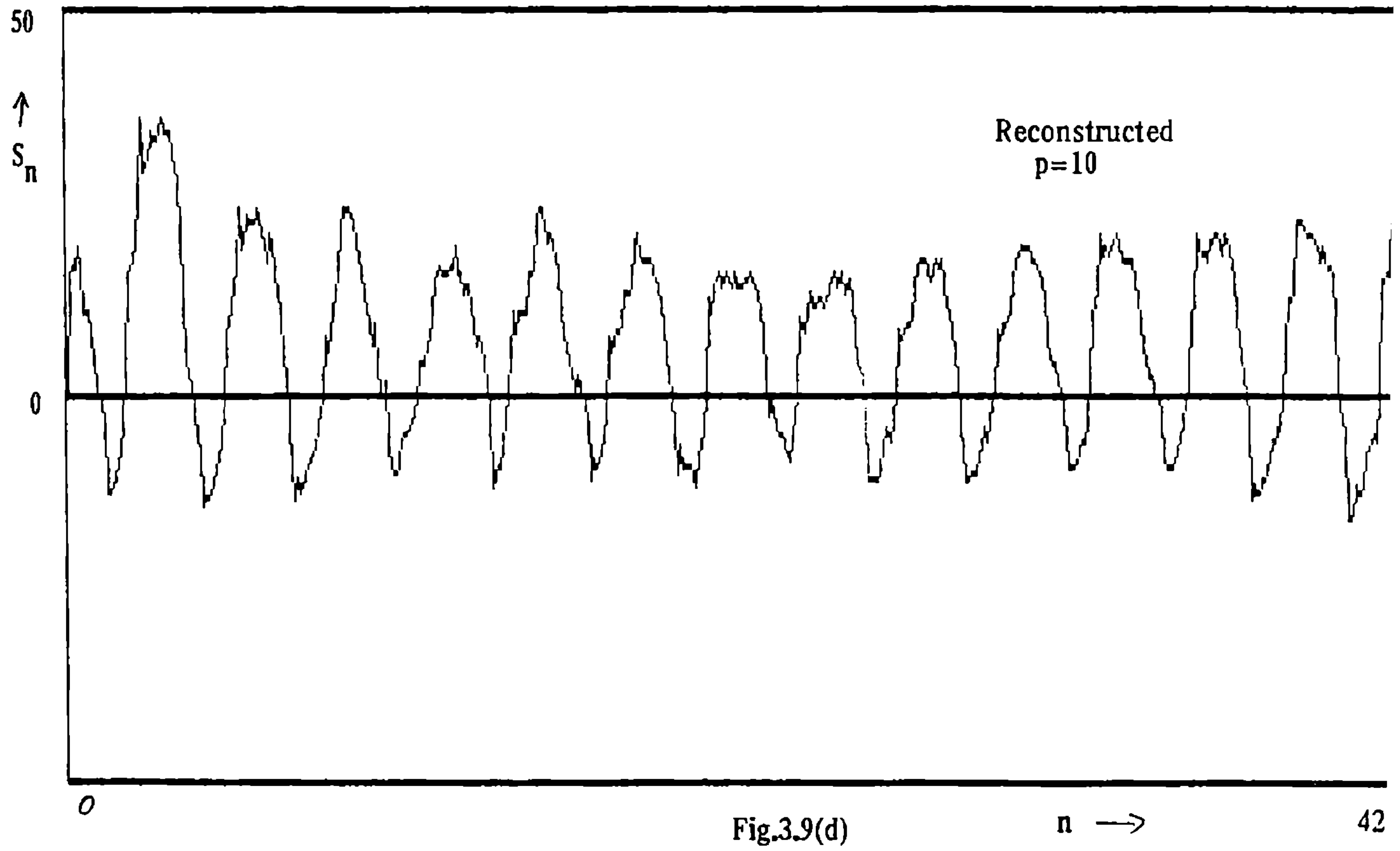


Fig.3.9(d)

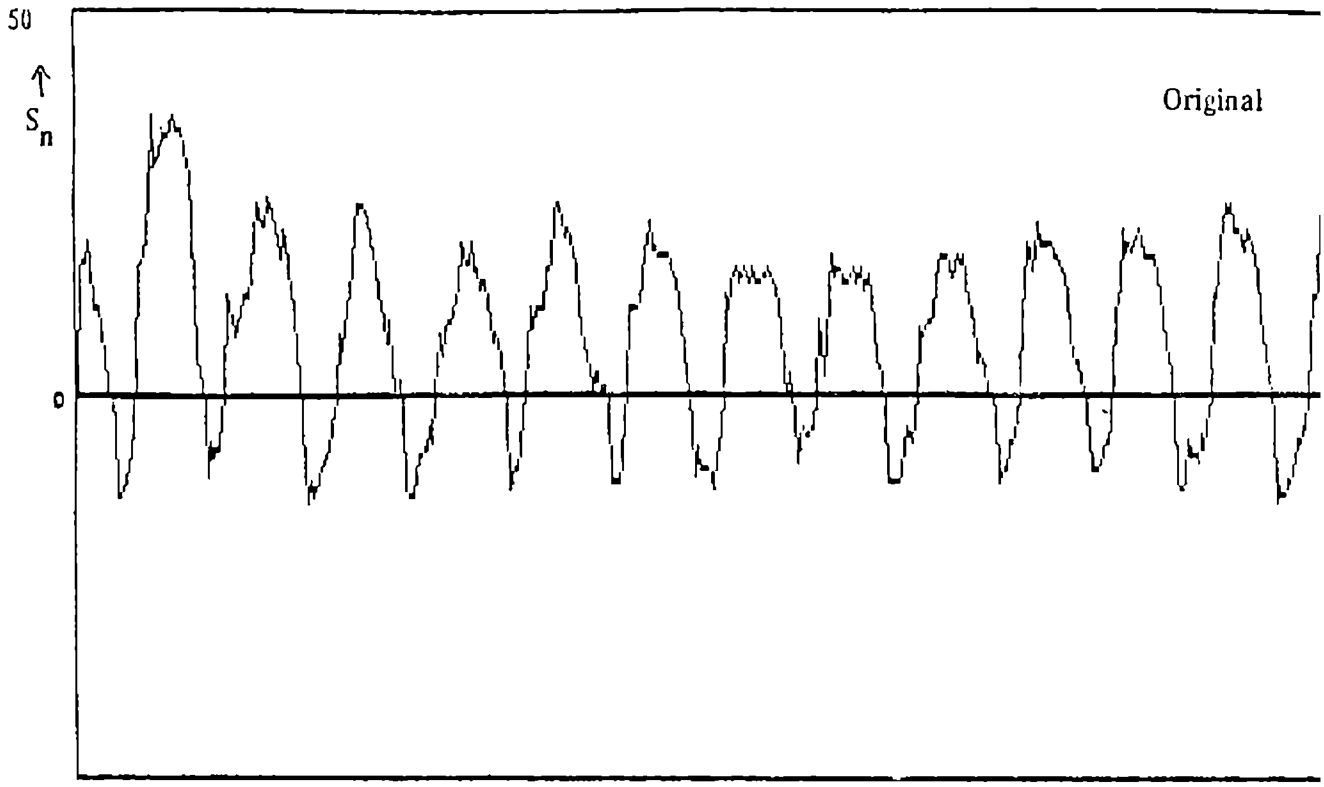


Fig 3·9 (e)

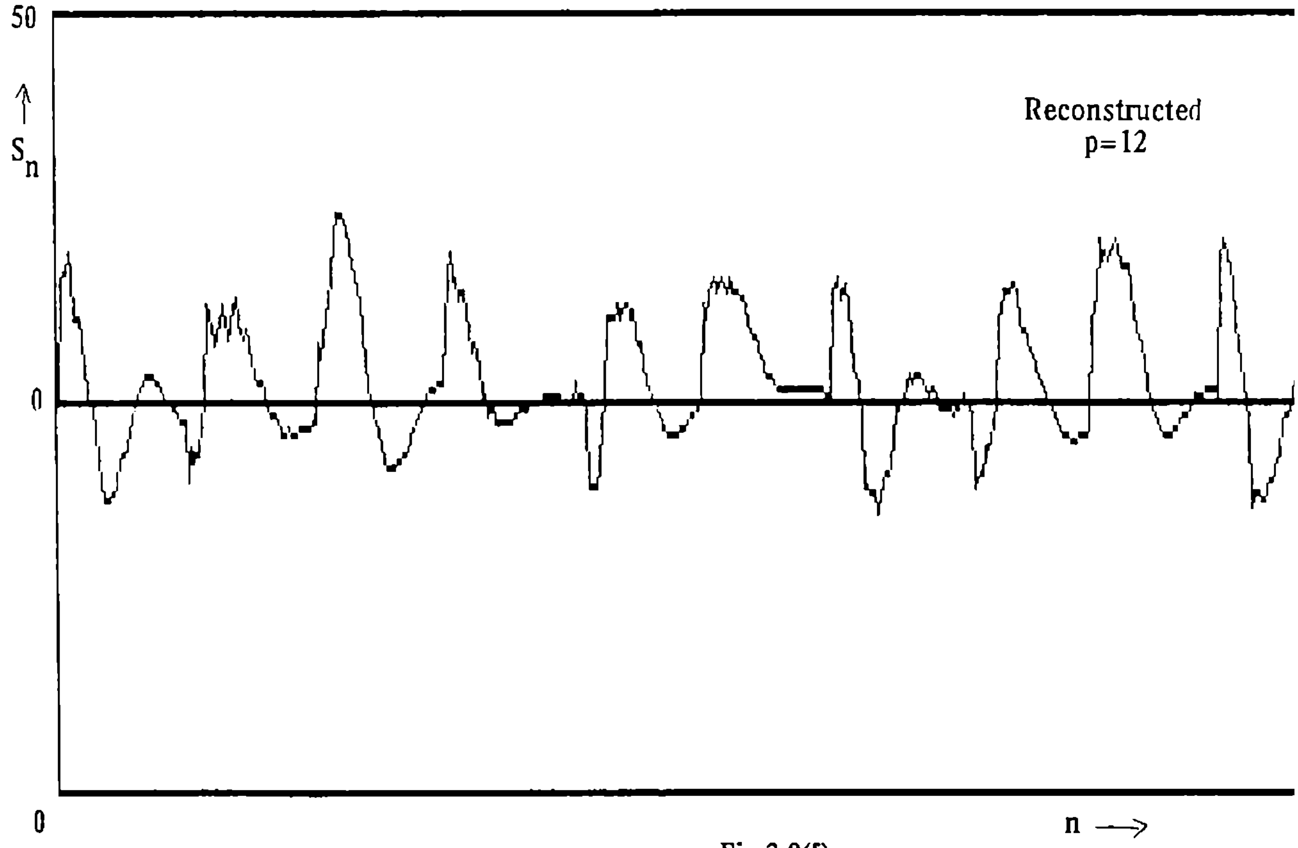


Fig.3.9(f)

Fig.3.9 (e–h) Plots of original and reconstructed waveforms – Using spectrum prediction – Transition region.

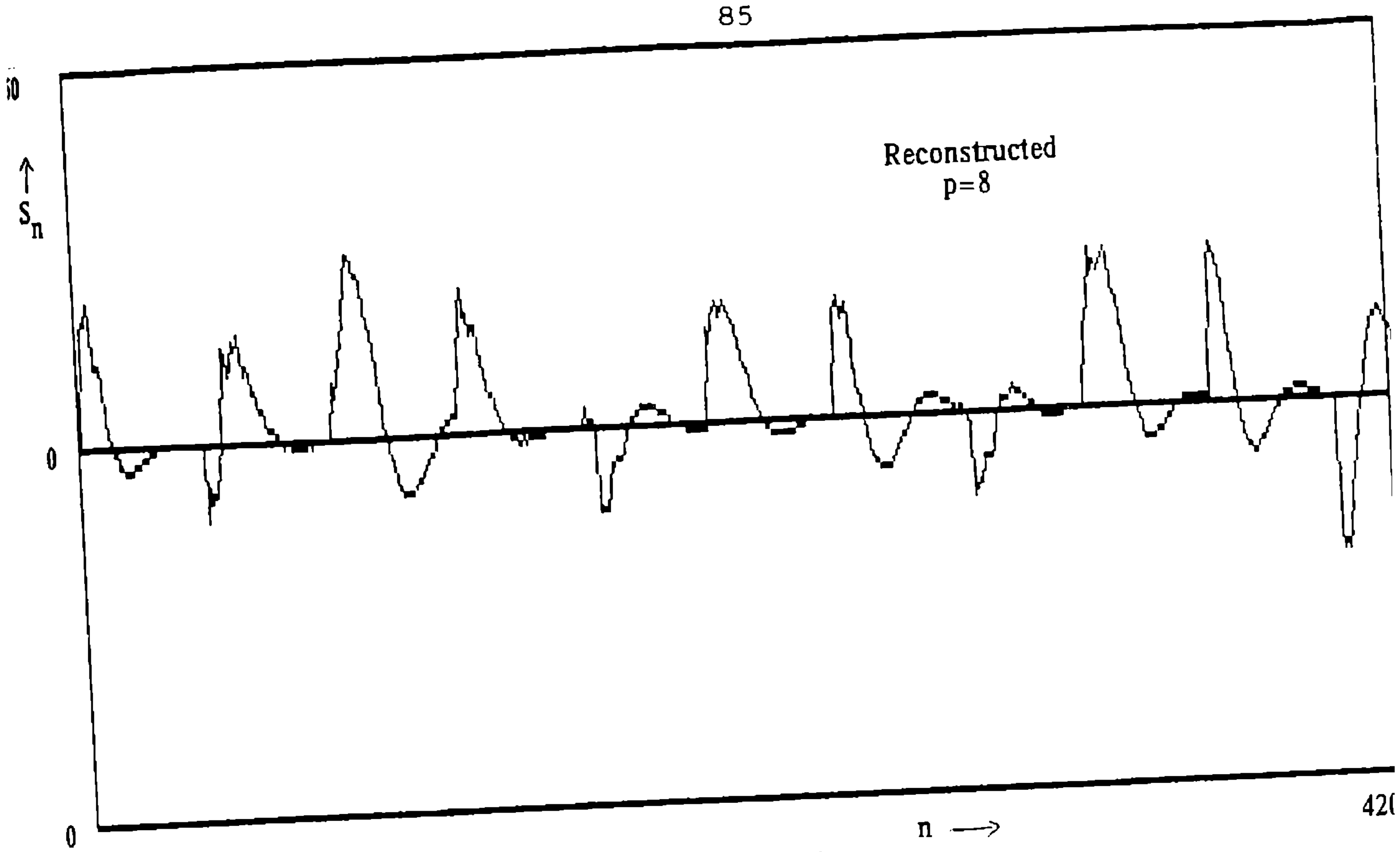


Fig.3.9(g)

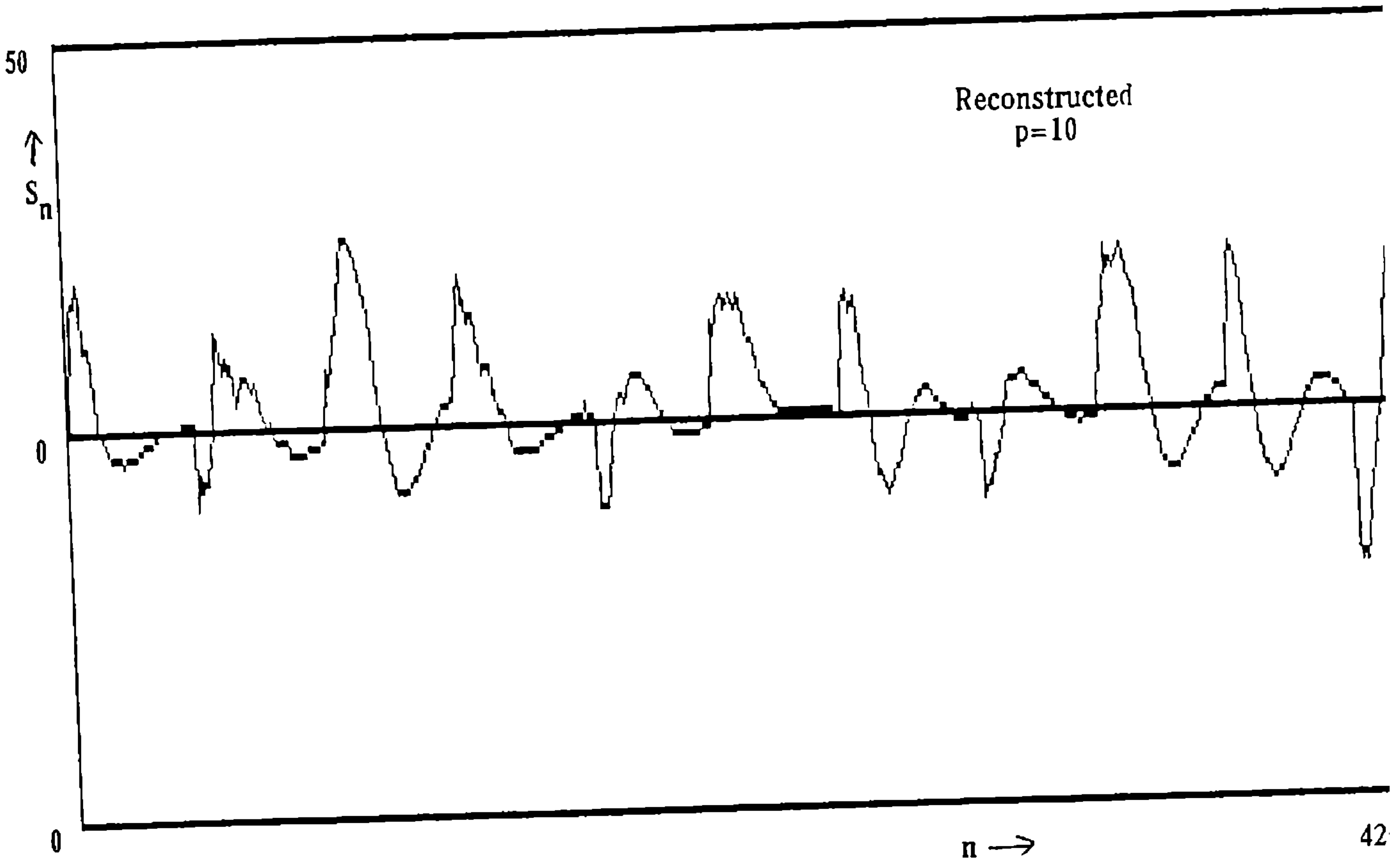


Fig.3.9(h)

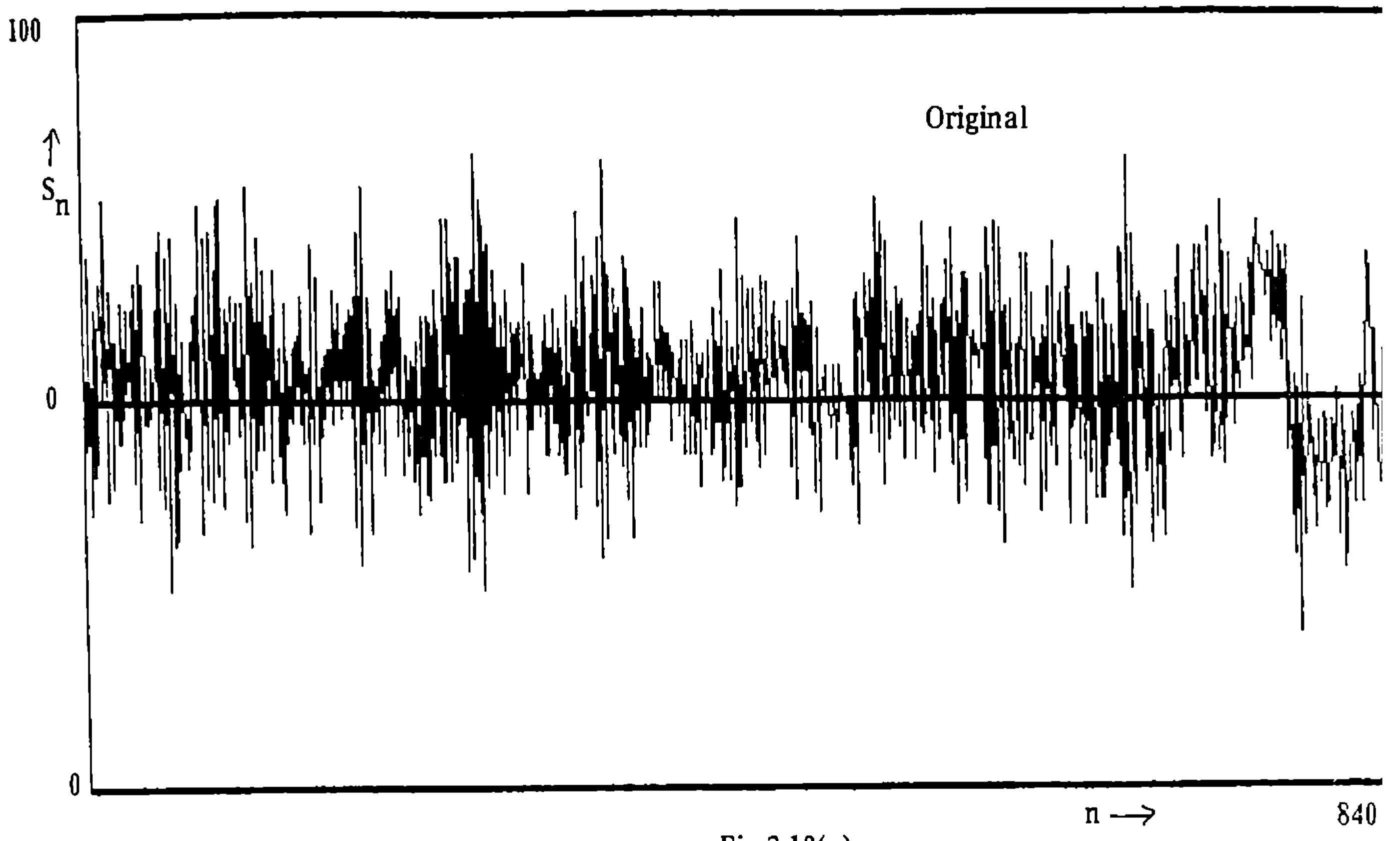


Fig.3.10(a)

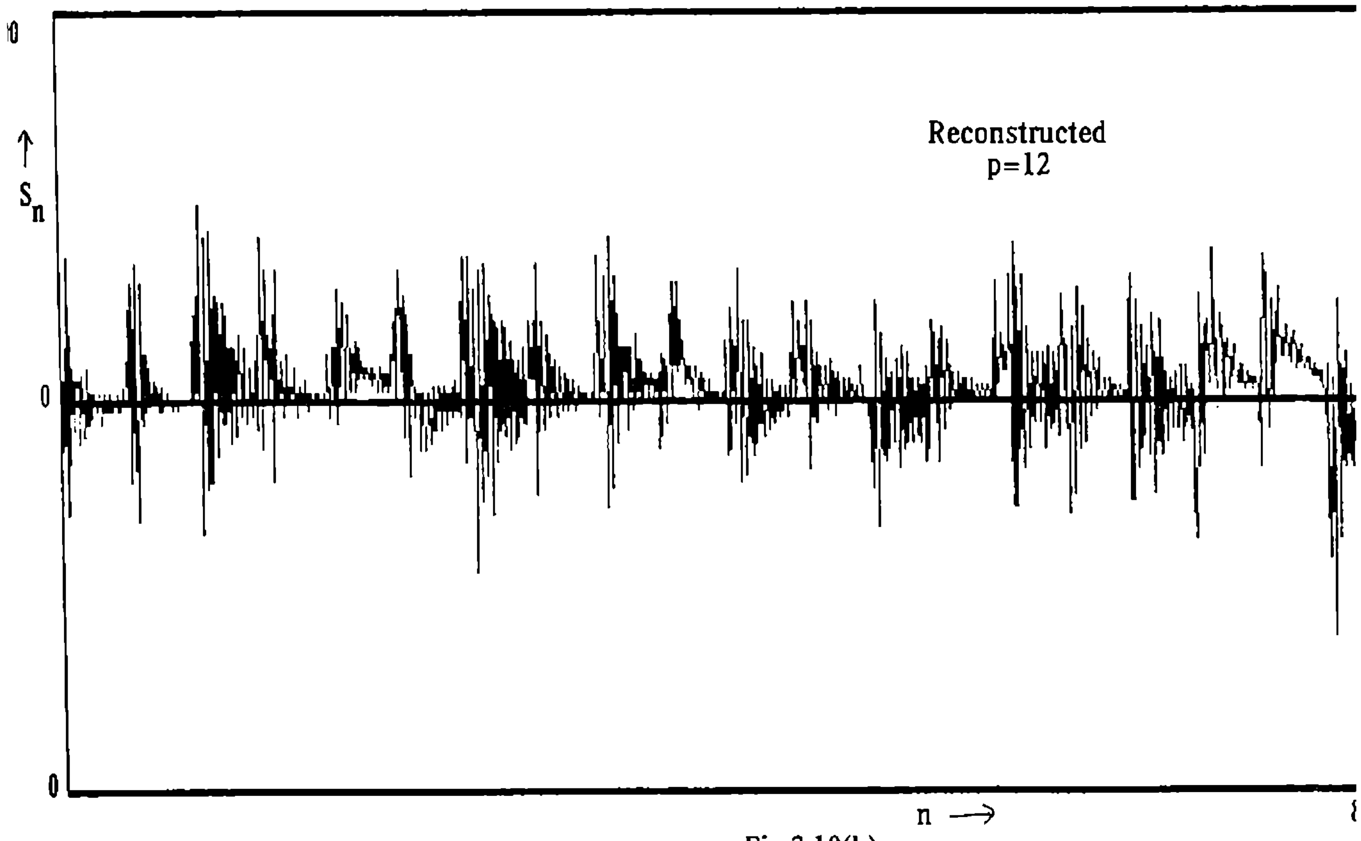


Fig.3.10(b)

Fig.3.10 (a-d) Plots of original and reconstructed waveforms - Using spectrum prediction - Unvoiced region.

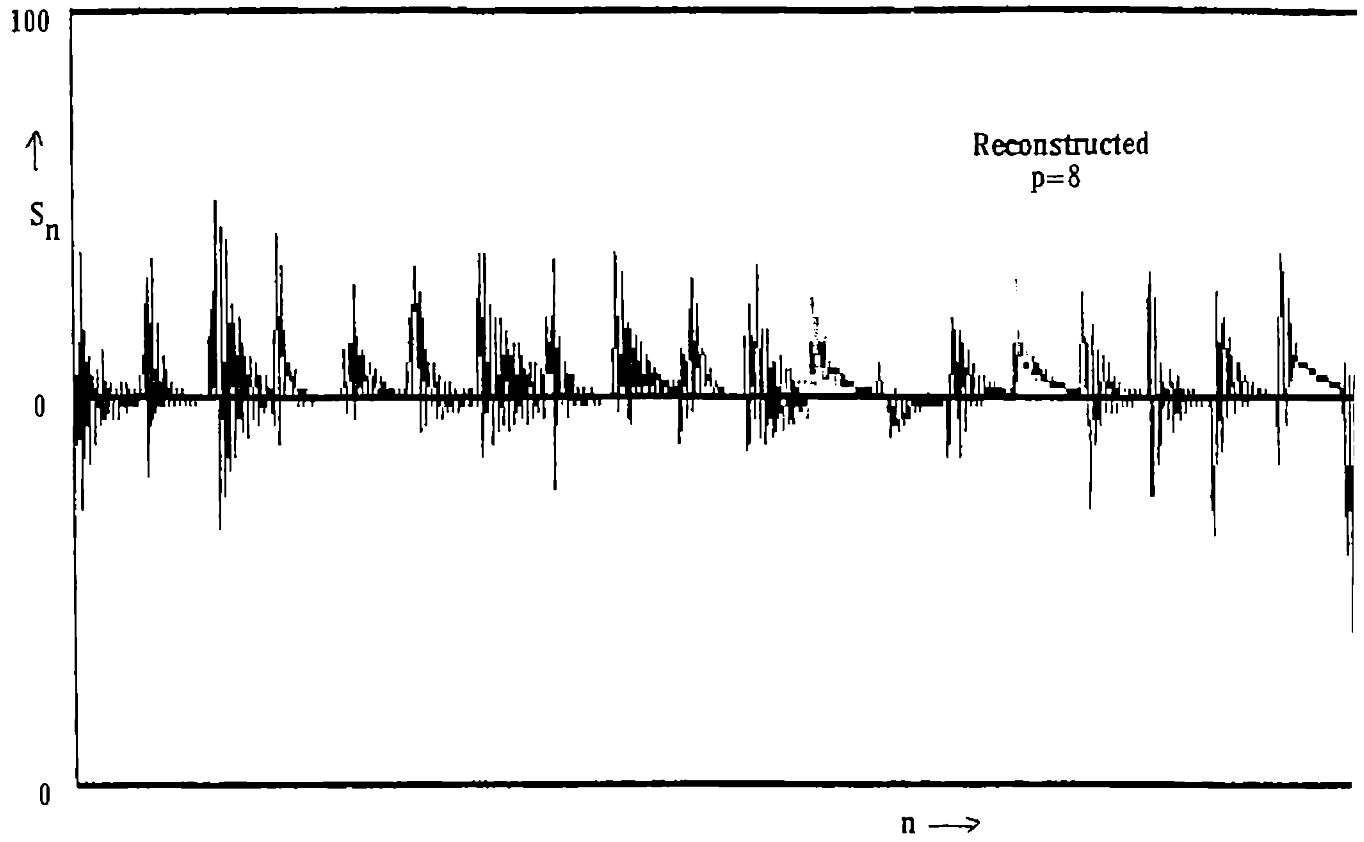


Fig.3.10(c)

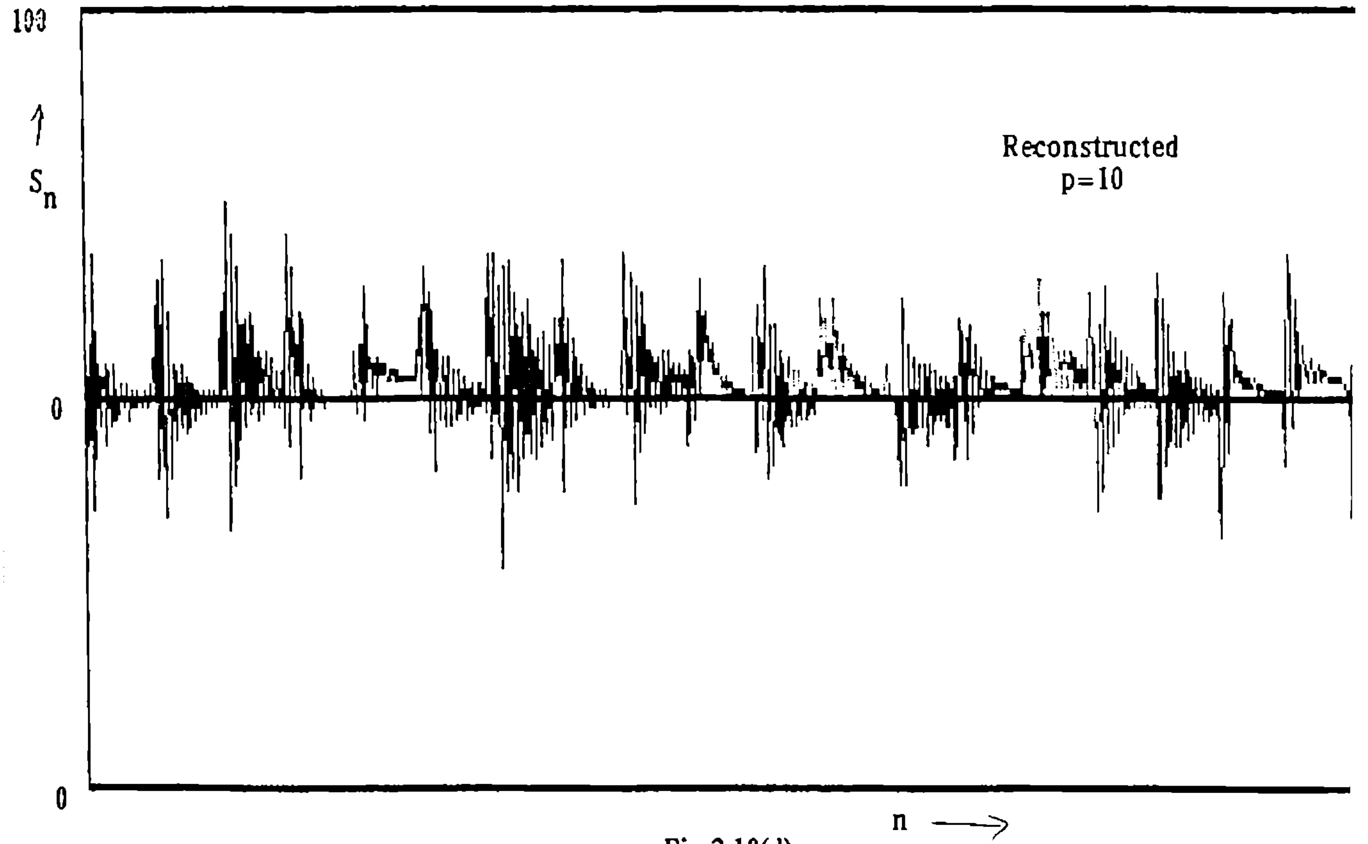


Fig.3.10(d)

Figures 3.7(a to j) illustrate the variations in the value of SNRSEG for various parameters like, the order of predictor p, and k, the segment number of the speech data used.

To obtain a comparative view of the original and the reconstructed waveforms, the reconstructed waveforms are plotted for 3 nearly optimum values of p, for the different regions (please refer figs.3.8 to 3.10). The reconstructed waveforms are quite acceptable.

Computational Savings

When the covariance matrix is reduced to the autocorrelation matrix, it can be noted that the number of matrix elements that are to be evaluated is reduced from p(p+1)/2 to p, where p is the order of the predictor. That is, for p = 8, the number of matrix elements to be evaluated reduces to just 8, from 36. This means that a reduction in computation, from $(p^2+p)N/2$ multiplications and additions to pN multiplications and additions, is possible (Above values are derived in Appendix II).

3.9.1.3 Evaluation of the Predictor Coefficients

Gauss-Jordan elimination method was used to solve for the vector a, from the equation ϕ a = ψ .

It has been noted that (as explained in Appendix II), using this method, the evaluation of the predictor coefficients, requires approximately $(p^2/2 + 3p/2)$ divisions, multiplications and subtractions. The square-root or Cholesky decomposition method requires $p^3/6 + O(p^2)$ operations, while Durbin's method requires $p^2+O(P)$ operations [11]. Hence, the elimination method used was comparable to that of Durbin's method.

3.9.2 Reduction In the Complexity of the Coder

In the modified coder, as mentioned earlier, the difference between the actual and the predicted value of the signal samples is not transmitted to the receiver. At the transmitter, using a block of sample values, the predictor coefficients are evaluated, and they are transmitted to the decoder (receiver) along with the decoded version of the first few sample values. At the receiver, using these parameters, the present value of a sample is estimated based on the earlier predicted values. The predictor at the receiver, is updated by transmitting the predictor parameters afresh every block.

3.9.2.1 Coding of Side Informations

The coding of the side informations (corresponding to the amplitudes of the first few samples every block), is achieved by a transformation or mapping.

A one-to-one mapping between the present sample value S_n and the corresponding difference value D_n, between the present and the preceding sample value, is used. The first sample value S_1, the step size

Δ, and the codes corresponding to the difference samples $\{D_n\}$ are to be transmitted.

It is stated that [2] short time pdf's of speech segments are described by a bell-shaped Gaussian pdf, irrespective of whether the speech segments are voiced or unvoiced. For a signal with Gaussian probability, only 32% of the instantaneous sample values will go above the standard deviation of the samples considered. Using these facts, the mapping was developed.

Let $D_n = S_n - S_{n-1}$ be the difference signal and let σ_d be the standard deviation of these difference samples. Then for voiced segments,

$$\sigma_d = \left[\sum_{n=2}^{p+M} (D_n - \overline{D}_n)^2 \right]^{\frac{1}{2}} \qquad (3.27)$$

where $\overline{D}_n$ is the mean value. The upper limit of the summation changes to p, for unvoiced segments. The difference samples are to be coded and transmitted to the receiver. The quantisation levels were made relative to the standard deviation of the difference samples.

Different levels of quantisation — 3-level, 7-level and 8-level —were tried and an 8-level quantizer was found to be optimum. The effects of the quantisation on the SNR values are shown in Table 3.5.

Table 3.5

Effect of quantisation on the SNR value

Files	Using original S_n		Using quantised D_n values					
			3 - level		7 - level		8 - level	
	Av.SNR	Av.SNRSEG	Av.SNR	Av.SNRSEG	Av.SNR	Av.SNRSEG	Av.SNR	Av.SNRSEG
1	15.827	12.712	3.559	3.367	13.150	11.525	13.328	11.555
2	10.270	9.410	2.404	2.196	7.238	6.765	7.887	7.434
3	12.786	12.378	9.419	9.187	12.339	11.914	12.443	12.127

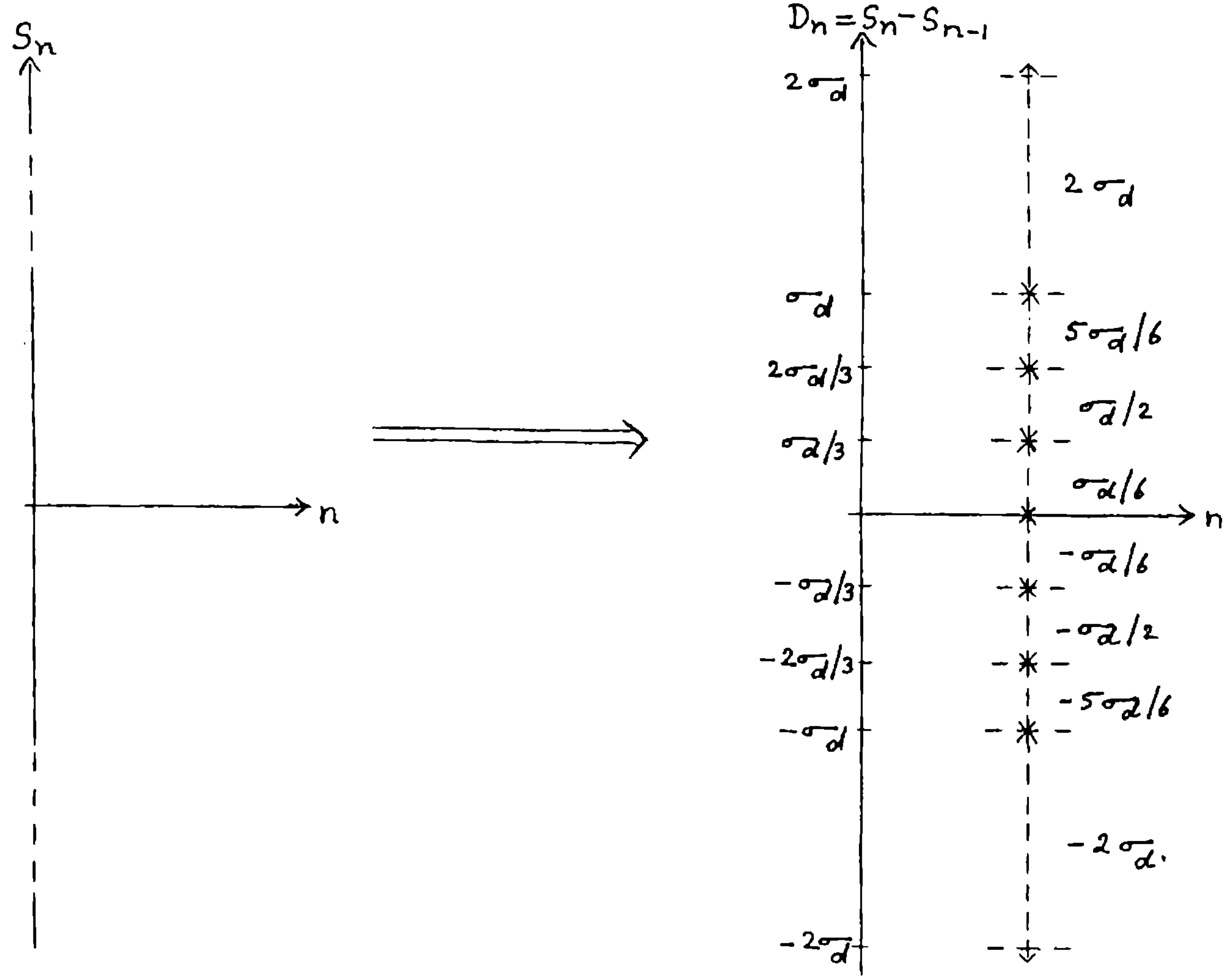


Fig.3.11 S_n-to-D_n Mapping

The S_n-to-D_n mapping using an 8-level quantiser, in fig.3.11. Using a mid-rise quantizer, based on the pdf of the difference signal as explained earlier, the values of D_n below σ_d are quantised into 3 equal steps, at $\sigma_d/6$, $3\sigma_d/6$ and $5\sigma_d/6$. The values of D_n above σ_d is taken as equal to $2\sigma_d$.

The above S_n-to-D_n mapping and further encoding is summarised as follows:

1. SI should be transmitted.
2. Find standard deviation σ_d and fix up step size $\Delta = \sigma_d$.
3. Then,

a) If $D_n > \sigma_d$, then $D_{nq} = 2\Delta$

b) If $2\sigma_d/3 \leq D_n \leq \sigma_d$, then $D_{nq} = 5\Delta/6$

c) If $\sigma_d/3 \leq D_n < 2\sigma_d/3$, then $D_{nq} = \Delta/2$

d) If $0 \leq D_n < \sigma_d/3$, then $D_{nq} = \Delta/6$

e) If $-\sigma_d/3 \leq D_n < 0$, then $D_{nq} = -\Delta/6$

f) If $-2\sigma_d/3 \leq D_n < -\sigma_d/3$, then $D_{nq} = -\Delta/2$

g) If $-\sigma_d \leq D_n < -2\sigma_d/3$, then $D_{nq} = -5\Delta/6$

h) If $D_n < -\sigma_d$, then $D_{nq} = -2\Delta$

D_{nq} is the quantised version of D_n.

4. Transmit codes corresponding to D_{nq}, to the decoder.

5. At the decoder, decode to obtain D_{nq} and hence the final version of the coded S_n's; from which the predicted estimate is computed.

Thus, the high-order predictor at the transmitter of an APC system is replaced by a simple delay unit, thereby reducing the hardware complexity of the system. The computational load is also reduced much, as a prediction is not needed at the transmitter, and the system becomes more real-time. Also, if needed, the time interval in between the transmission of the predictor parameters, can be effectively used for time-division multiplexing between other signals or systems.

3.9.3 Reduction in Bandwidth Requirement

At the first glance, one might expect a minimum reduction in bandwidth of the order of f_s KHz (f_s is the sampling frequency of the signal), over the APC system of Atal and Schroeder, as the difference signal is not transmitted. But a coded version of the first few sample values have to be transmitted every block, to maintain a better SNR value. But the predictor parameters and side informations tolerate coarse quantisation and slow updating and hence much excessive channel capacity is not required. The work done in this direction is explained, in detail, in the following chapters.

3.10 The Modified Block Adaptive Coder

The block diagram of the modified block adaptive coder

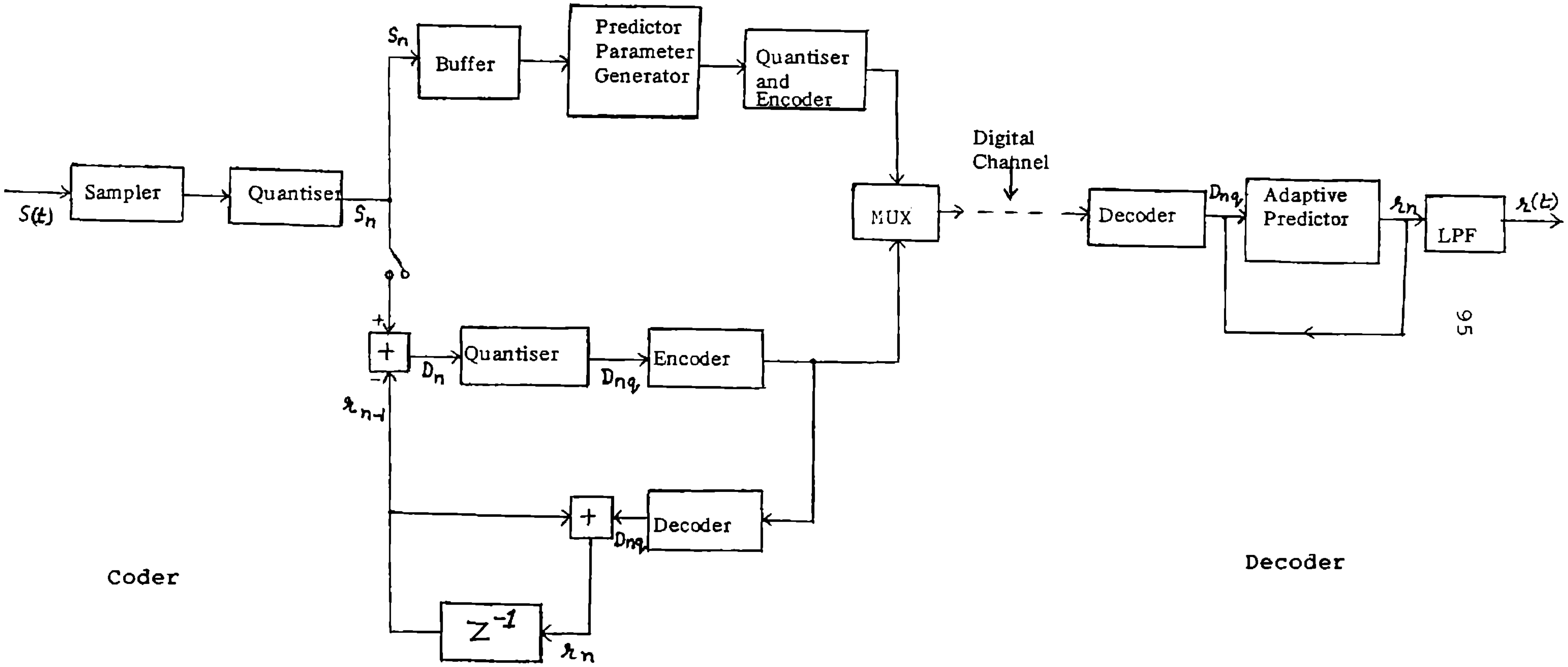


Fig.3.12 The Modified Block Adaptive Predictive Coding System

(MBAC) is shown in Fig.3.12. The input signal s(t) is low pass filtered to 4 KHz, sampled at 8 KHz and encoded to 12 bits, to form the speech samples $\{S_n\}$. Using blocks of data values, of appropriate length, the predictor parameters are computed and are suitably encoded and then transmitted to the receiver.

At transmitter, the difference between the adjacent sample values is obtained, for the first few sample values, using a delay unit as shown in figure 3.12. These difference sample values are quantised, encoded and transmitted. They are also decoded, and used at the transmitter to form the reconstructed samples r_n, from which the next delayed version of the sample is obtained. This reduces the quantisation error effect.

At the receiver, the difference signal is added to the previous reconstructed values to obtain the predicted values of the first few samples. (The first sample S_1, is to be transmitted). For the rest of the samples, the predictor forms an estimate, based on the predictor parameters received. They are low-pass filtered to get the reconstructed speech signal. The predictor is updated every block of 4M samples. The value of M may change from block to block, depending on the correlation of the input samples.

The actual simulation of the modified coder is explained in the next chapter.

3.11 Signal-to-Noise Ratio of the Modified Coder

The signal-to-noise ratio representation of the modified coder is not straightforward as in a DPCM system with a linear predictor, or the APC system, as explained in section 3.5. The total error in the reconstructed signal is due to (i) error in prediction (ii) error due to the new prediction estimate being based on the earlier predicted values and (iii) the error in quantising the predictor parameters and side informations.

The performance of the system can be measured in terms of the standard segmental signal-to-reconstruction error ratio, given by

$$\text{SNRSEG (dB)} = \frac{1}{M} \sum_{m=1}^{M} \text{SNR}(m) \text{ dB} \tag{3.29}$$

where,

$$\text{SNR}(m) = \frac{E_s^2}{E_r^2} \tag{3.30}$$

E_s^2 is the variance of the signal in the current block and E_r^2 is the variance of the corresponding prediction errors.

To summarise, a predictive coding system is presented in the first half of this chapter. The latter part presents the actual modifications involved in the development of the modified block adaptive predictive coder.

Chapter 4

SIMULATION STUDY ON THE MODIFIED BLOCK ADAPTIVE CODER

The actual simulation process of the Modified Block Adaptive Coder is presented in detail in this chapter.

4.1 Speech Data Used

To study the feasibility of the coder on all types of sounds and phonemes, in the English language, the speech data base is so chosen as to contain almost all the phonemes in English. Two sets of speech texts are used in the work. In the first set, consisting of four phrases, shown in Table 4.1, care is taken to include more of vowels and nasals, which are to a great extent speaker dependent. Here it can be noted that one word ends at one region and the next word starts at a different region, distinctly apart in the vocal tract, so that merging of words is avoided. The second set consists of eight sentences, shown in Table 4.2, with a total duration of 21 secs. These sentences are chosen since they are phonetically well-balanced. For both* sets, speech data from a male and a female speaker were collected.

Speech data were collected using a PC based speech digitizer. A

* Data set 1 was obtained from Dr.Babu P.Anto and data set 2 was obtained from Dr.N.K.Narayanan, both former research scholars of the Department of Electronics, Cochin University of Science and Technology.

Table 4.1

	Phrases used	Male	Female
1.	Drop coin after tone	B1	T1
2.	Push blue after speech	B2	T2
3.	Close door after party	B3	T3
4.	Right move close lock	B4	T4

Table 4.2

Sentences used	Male	Female
1. The pipe began to rust while new	Ml	Fl
2. Cats and dogs hate each the other	M2	F2
3. Oak is strong and also gives shade	M3	F3
4. Thieves who rob friends deserve jail	M4	F4
5. Open the crate but do not break the glass	M5	F5
6. Add the sum to the product of these three	M6	F6
7. Joe brought a young girl	M7	F7
8. A lathe is a big tool	M8	F8

dynamic microphone was used as the input to the system. An ordinary A/c room was chosen for the work. The system noise was adjusted to a minimum at the beginning of the data acquisition session. In real life environment, there will be a lot of background noise and hence it will be always better to study the performance of a system in a noisy environment. Hence an ordinary A/c room was chosen as the venue for the data acquisition. It is reported by Babu P.Anto [79] that when an anechoic chamber was chosen as the speech chamber, the absence of normal ambient noise tends to make the speaker speak with slight difficulty.

The speech waveform was band-limited to 4 KHz, sampled at 8 KHz and encoded to 12 bits and stored in the system RAM and then transferred and stored as files on disks. These sampled data were later used for the processing.

4.2 Voiced/Unvoiced/Silent/Transition Classification

As explained in section 3.2, depending on the statistical properties of the speech waveform, there are mainly two different categories of speech sounds, namely voiced and unvoiced. The vowels and the nasals (like /a/, /i/, /I/, /e/, /æ/, /ʊ/, /u/, /o/, /ɔ/, /m/, /n/, /η/ etc.) are voiced sounds with high energy and high correlation, whereas, the fricatives and plosives (like /s/, /ʃ/, /f/, /θ/, /p/, etc.) are unvoiced sounds with low energy and low correlation.

Speech is a non-stationary quasi-periodic waveform. Though it is not fully periodic, the voiced regions remain periodic for about 4 to 5 pitch periods [14]. Unvoiced regions are almost uncorrelated, while transition regions show sharp and sudden changes. Hence, while processing speech segments based on parametric methods, special care must be taken to classify the different regions. The silent regions do not have much signal information, and hence no processing need be done for those regions. Hence it was thought feasible to develop a method which will detect all the four regions correctly. This is actually a difficult problem in speech analysis.

As such, many algorithms are available in the literature for voiced/unvoiced detection. The aim of all these algorithms is to find out certain features of the speech signal, that can help in this decision. Atal and Rabiner [80] have considered the voiced/unvoiced classification as a pattern recognition problem. Five features like zerocrossing rate, correlation coefficient, energy, LPC predictor coefficients and prediction error energy have been considered. Rabiner and Sambur [81] have presented an LPC distance measure for voiced-unvoiced-silence detection. Knoor [82] has proposed another technique, by filtering the speech and comparing the rectified filter outputs.

Three methods are presented for the detection of voiced/unvoiced/silent/transition regions. These algorithms are based on certain basic statistics of the speech waveform and were developed after extensive arithmetic work.

In the first method, fixing up two threshold values for the short-time zero crossing rate (STZCR), an initial silent/unvoiced detection is done. Next, for blocks having values of STZCR in between the two thresholds, periodicity of autocorrelation functions (ACF), product of STZCR and short-time energy (STE), and S_{rms} to S_{mean} ratio are used for the further classification. In the next two methods also, the initial detection step is the same as explained in method I. For the final detection, normalised correlation coefficient ρ_1 for lag 1 was used in method II, while the periodicity of the ACF's and the value of a correlation factor β were used as measures for detection in the third method.

4.2.1 Algorithm I for Voiced/Unvoiced/Silent/Transition Detection

The silent region in a speech waveform will have very low energy. But certain unvoiced sounds like, /f/, /θ/, /p/, /k/, /s/, etc. at the beginning or end of a speech sequence, and also nasals at the end, have very low energy [2, 83], and can be mistaken to be a silent region. But the STZCR for a silent region is very low, compared to the other regions. Energy of voiced speech is concentrated below about 3 kHz, while that of unvoiced speech is found at higher frequencies. Hence the STZCR of unvoiced segments should be higher than that of voiced segments. But there are regions where overlapping can take place, in both the STE and the STZCR domains [3, 80] of the voiced and the unvoiced segments. But voiced regions show a periodic nature while unvoiced regions do not. Transition regions can have energy and zerocrossing rate relatively at all levels,

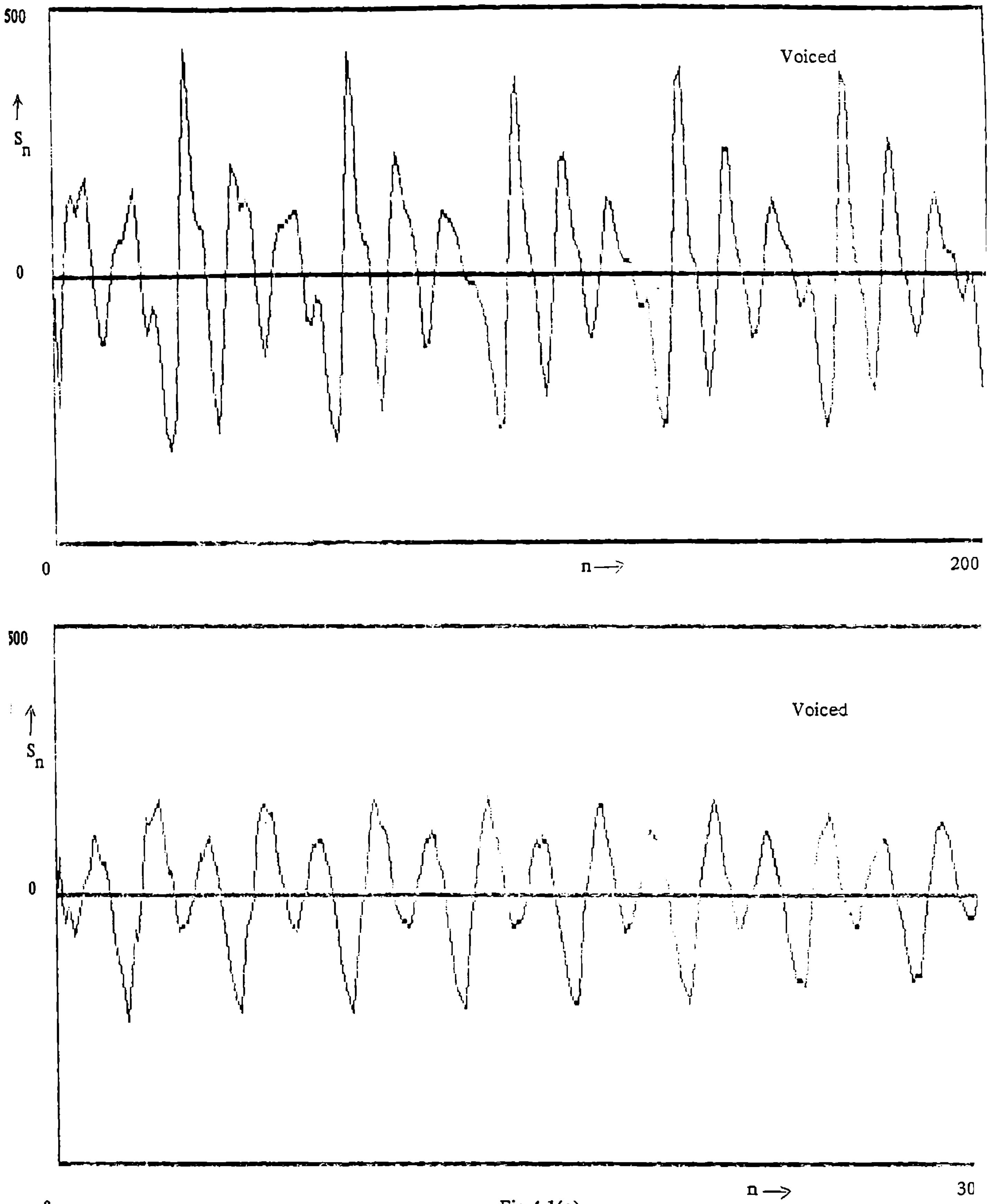


Fig.4.1(a)

Fig.4.1(a-d) Typical waveforms of varic:s regions in speech signals.

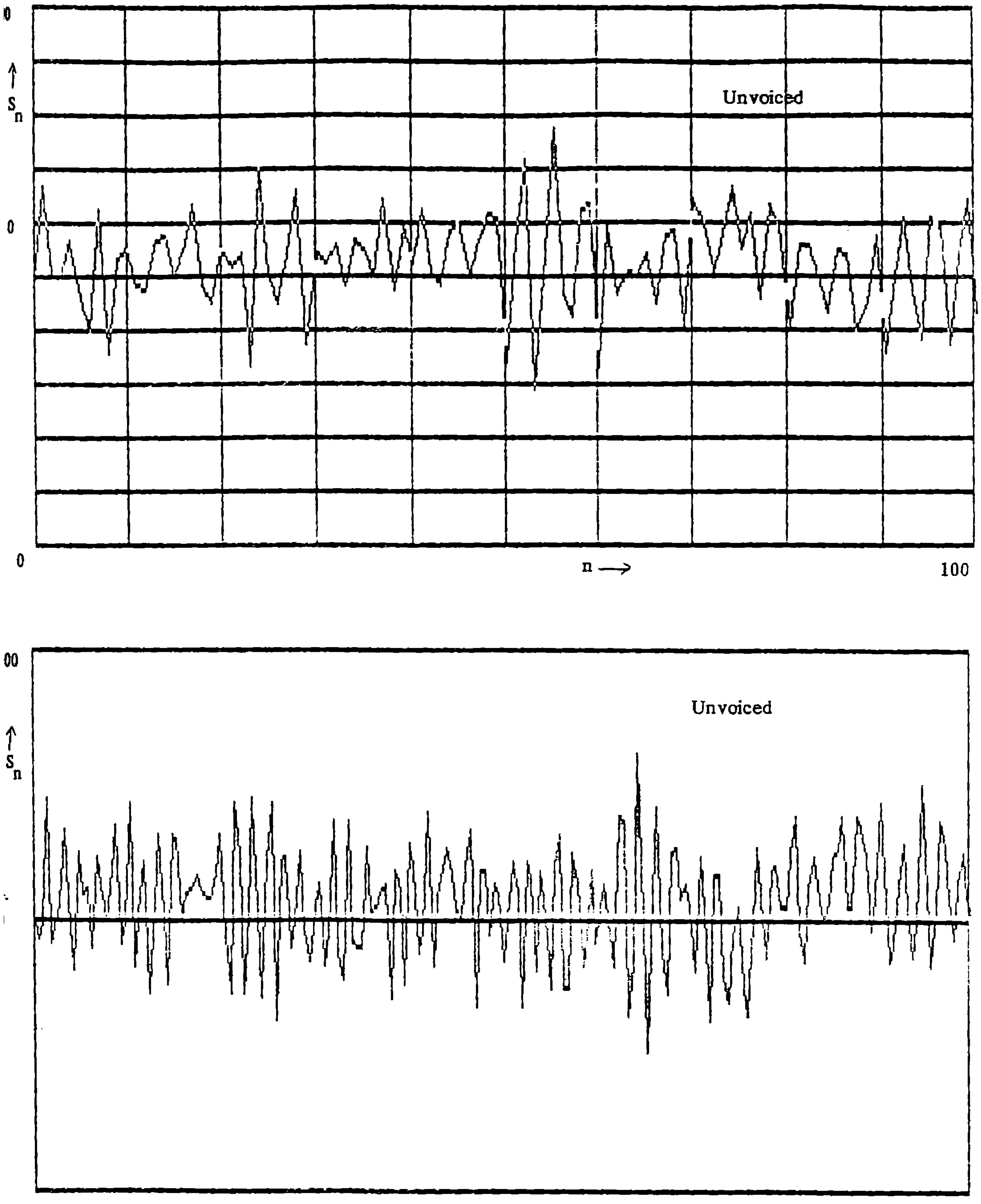


Fig.4.1(b)

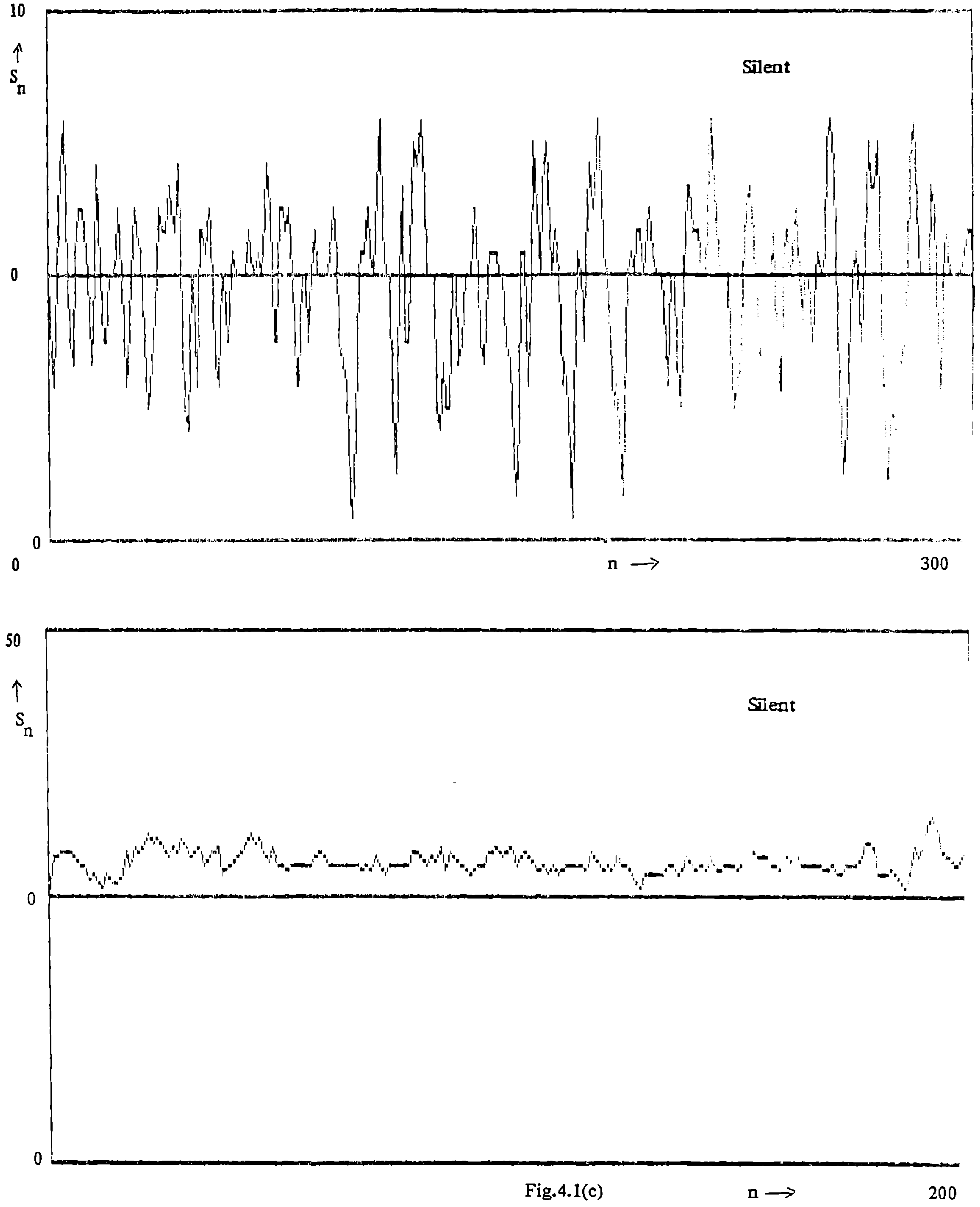


Fig.4.1(c)

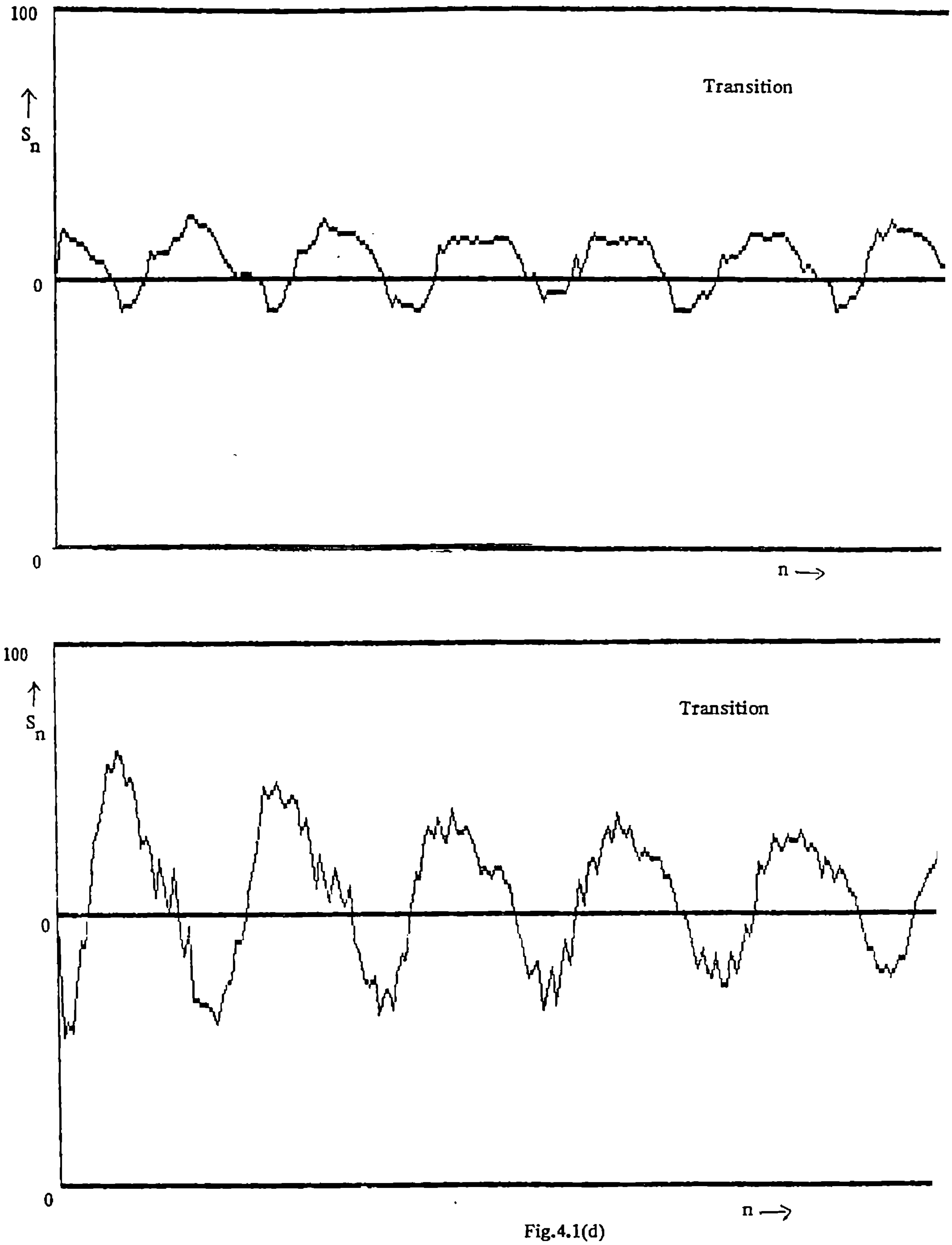


Fig.4.1(d)

between the two extreme ends, for the silent and unvoiced regions. Transition can occur between any combination, from among voiced, unvoiced and silent regions, and hence the energy, ZCR and also the duration of the transition regions will be relative to both the preceding and succeeding segments.

Three sentences, each of duration 2.5 secs, spoken by three speakers, were considered for the training session. They were, the sentence "The pipe began to rust while new"(data files M1 and F1 in table 4.2) spoken by a male and female speaker and the sentence, "Close door after party" (data file B3 in table 4.1) spoken by another male speaker. Typical plots of the various regions, present in the above data files are shown in figure 4.1a to 4.1d.

It is reported [14, 43] that the vocal tract 'rings' for a duration of about 10 msecs. Also, under natural speaking conditions, the pitch period can vary from 2 msec. for very high-pitched females and children, to 20 msec for very low-pitched males [3, 43, 44]. For a sampling frequency of 8 kHz, it corresponds to a maximum of 160 samples. In the present study, the maximum pitch period was less than 8 msec, and the block length of 160 samples, helps in validating the periodicity in the peak positions of the correlation coefficients, within a block. Hence the best block length for the processing was fixed as 160.

The speech data were first grouped into blocks of 160 samples, and each block was manually classified into the four different regions, by plotting on a VDU. Parameters like energy, ZCR, S_{rms} to S_{mean} ratio (RMR), ZCR to energy ratio (ZER), product of ZCR and energy (ZEP), normalised correlation coefficient (NCC), periodicity of ACF's, correlation factor β denoting change in amplitude from one pitch period to the next, etc. were evaluated every block, and studied, and the following algorithm was developed. The values of the different parameters obtained for the different regions of the different data files, are summarised in Table 4.3.

The different steps involved in algorithm I is explained below.

Step 1: Check for STZCR

(a) If STZCR $<$ 200, then Silent Region.

(b) If STZCR $\gtrsim$ 3500, then Unvoiced Region.

If STZCR is between 200 and 3500, then it can belong to any of the four regions. Hence check for the periodicity of the speech segment, by noting the consecutive peak positions of the ACF's R(J), of the original samples, in the segment under consideration. It is verified by the author (please refer Table 3.1), that the ACF's of the unclipped sample values are enough, for pitch period determination, as against the centre-clipped method developed by Sondhe [76], or the normalised correlation coefficient method of Atal and Schroeder [44]. Let M1 denote the position of the first maxima of R(J), (leaving the peak at J = 0) and M2 be the distance, in number of samples, between the first and second maxima. Or in other words, M1 and M2 represent the number of samples within consecutive pitch periods [as

Table 4.3

Results of the Analysis of the data files, for voiced/unvoiced/silent/transition detection

Region	Data Files	STZCR	M1	M2	M3	RMR	STE	ZEP	ZER	ρ_1	β
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
SILENT	F1	0–100	16	15	15	1.0–1.08	(.79–3.0)10^{-5}	(.5–2.0)10^{-3},	(1.4–5.1)10^{6}	.94–.965	.82—.95
			45	25	21				0.0		
	M1	500–1650	49	21	24	2.7–3.7	(1.2—7.3)10^{-6}	(.7–8.0)10^{-3}	(.68–6.3)10^{8}	.20–.48	.33–.59
		2500–2650	64	24	33	3.5–3.8	(1.2–2.2)10^{-5}	(3.0–5.0)10^{-2}	(1.1-2.1)10^{8}	.13–.25	.32–.46
		850–1850	47	39	54	3.0–3.9	(.8–5.9)10^{-5}	(1.2–9.0)10^{-2}	(.26–1.0)10^{8}	.46–.50	.44–.88
	B3	1950–2650	30	60	37	1.25–3.5	(.4–5.4)10^{-5}	(.9–9.0)10^{-2}	(.9–5.7)10^{8}	.01–.56	.38–.47
UNVOICED	F1	3450–3700	42	16	17	4.2–6.3	(1.2–1.6)10^{-3}	(3.04–6.07)	(2.2–3.9)10^{6}	–.17–+.013	.26–.37
			15	23	38						
	M1	3500–3700	30	49	36	8.0–11.0	(3.9–13)10^{-3}	40–80	(.5–2.4)10^{5}	.15–.295	.64–.69
		3000	36	39	35	2.4	3.7x10^{-5}	.11	8.1x10^{7}	.27	.6
	B3	3300-4500	15	23	101	7–11	(.18–.79)10^{-3}	1.9–2.8	(4.4–7.9)10^{6}	-.24–+.32	.16–.30
			53	15	72						

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
VOICED	F1	200–400	34	34	34	23.5–28.0	(.1–.15)10^{-2}	17–47	(2.5–9.0)10^{4}	.96–.98	.94–1.03
		700–1950	38	38	38	15–20.0	(1–12.3)10^{-2}	10.5–166	(1.2–11.8)10^{4}	.77–1.01	.77–1.09
		2050–2800	29	29	30	14.1–14.7	(2.6–6.4)10^{-2}	55–163	(3.1–8.0)10^{4}	.42–.72	.77–.91
	M1	1150–2000	40	39	40	14.4–31.7	(.7–8.0)10^{-2}	20–155	(2.1–23)10^{4}	.46–.83	.78–1.08
		2000–2900	51	50	52	14.8–23.4	(0.7–4.6)10^{-2}	20–86	(4.2–31)10^{4}	.43–.57	.73–.97
		3100–3450	49	49	49	11.0–13.5	(3.0–3.9)10^{-2}	100–124	(4.5–5.0)10^{4}	.30–.40	.73–.80
	B3	750–1700	39 49	39 49	39 49	14.8–53.8	(1.0–18.6)10^{-2}	19–223	(0.6–13.9)10^{4}	.78–.94	.81–1.15
TRANSITION	F1	200–1350	31	31	30	2.0–4.4	(.9–6.0)10^{-3}	.4–7.0	(.4–10.1)10^{5}	.93–.97	.75–1.01
	M1	1100–2350	71	75	—	40–45.0	(2.7–3.6)10^{-3}	3.0–8.2	(.6–4.0)10^{5}	.60–.88	.56–.66
		2350–2750	57	56	—	40–45.0	(.4–2.9)10^{-3}	.96–6.2	(7.0–60)10^{5}	.55–.66	.61–.96
	B3	300–750	48	47	46	25–66	(0.8–12.0)10^{-3}	0.6–8.1	(8.0–98)10^{5}	.93–.97	.84–1.1
		800–1500	55	20	35	14.9–27.1	(2.0-8.0)10^{-3}	2.0–8.7	(1.3–5.5)10^{5}	.86–.91	.78–.85

M1, M2, M3 — Distances between consecutive peak positions of the ACF's, R(J), in terms of lag J.
Z E R — STZCR/STE
RMR — S_{rms}/abs (S_{mean})

shown in Fig.3.6].

A transition region, being influenced by its preceding and succeeding regions, can show a periodic nature. So is the case with the silent region also (Refer Table 4.4). The author has also noted that under rare conditions, especially for male speakers, the inter word and intra word silence regions show a pseudo-random nature, with an ZCR value higher than usually expected (for example, STZCR goes upto 2900 for male speakers while it is below 200 for female speakers).

Based on the above points, step 2 is evolved:

Step 2:

(a) If M2 $\neq$ M1 $\pm$ 2, then Unvoiced Region or Transition region or Silent Region.

(b) If M2 = M1 $\pm$ 2, then Voiced Region or Transition Region or Silent Region.

In general, the amplitude of the data samples in the silent region is less than that in the unvoiced region, and that of the samples in the unvoiced region is less than those in the voiced region. Hence, a specific relationship can be developed between the RMS and mean values of the samples in the different regions.

Table 4.4

Analysis results to show that silent and transition regions can also show periodicity

Date files	Regions	M1	M2	M3	STE	ZEP
F1	Silent	16	15	15	$(.79-3.0)10^{-5}$	$(.5-2.0)10^{-3}$
	Transition	31	31	30	$(.9-6.0)10^{-3}$	0.4-7.0
M1	Transition	57	56	--	$(.4-2.9)10^{-3}$	.96-6.2
B3	Transition	48	47	46	$(.8-12.0)10^{-3}$	.6-8.1

M1, M2, M3 - Number of samples in consecutive pitch periods.
STE - Short time energy
ZEP - Product of short time energy and zerocrossing rate.

Table 4.5

Analysis results to show that product of STE and STZCR is a better parameter than their ratio, for silent/unvoiced detection

Data files	Regions	STE	ZEP	ZER
F1	Silent	$(.79\text{-}3.0)10^{-5}$	$(.5\text{-}2.0)10^{-3}$	$(1.4\text{-}5.1)10^{6*}$
B3	Silent	$(.8\text{-}5.9)10^{-5}$	$(.12\text{-}.7)10^{-1}$	$(2.6\text{-}10.0)10^{7}$
M1	Unvoiced	3.7×10^{-5}	0.11	8.1×10^{7}

ZER = STZCR/STE

* It can be noted from Table 4.3 that this value of ZER actually falls in the range of unvoiced region.

The fact that the STZCR is larger for unvoiced speech than for voiced speech, and also that, the STE is lesser for unvoiced speech compared to voiced speech are made use in defining the ratio of STZCR to STE as a parameter for voiced/unvoiced classification [84, 85]. The ratio STZCR to STE (ZER) should be very distinct for the two regions, being very high for unvoiced regions and very low for voiced regions. But the author has found that, the values of STZCR and STE, in general, being both low for silent region and both comparatively high for unvoiced region, will result in the value of the ratio of STZCR to STE being almost of the same order, for both silent and unvoiced regions. The author has also noted that the product of STZCR and STE (ZEP) is a better parameter for the classification, in this context [Please refer Table 4.5].

To take into consideration those unvoiced and silent regions which go undetected by step 1, due to the STZCR value of the unvoiced region being slightly less than the threshold value fixed for those regions, and the STZCR of the silent region being slightly higher than that fixed for those regions, step 3 has been evolved. The ratio of S_{rms} to absolute S_{mean} value (RMR) has been taken as the parameter for detection. Similarly, to take care of the rare instances, where a low energy unvoiced region overlaps a silent region, step 4 has been formulated.

Step 3:

(a) If STZCR $\geq$ 2800 and

$0.0001 < \text{STE} < 0.1$ and

$4 < \text{RMR} < 14$,

then Unvoiced Region.

(b) If STZCR $<$ 2800 and

STE $<$ 0.0001 and RMR $<$ 4,

then Silent Region.

Step 4:

(a) If STE $<$ 0.0001 and ZEP $<$ 0.1

then Silent Region.

(b) If STE $<$ 0.0001 and ZEP $>$ 0.1

then Unvoiced Region.

(c) Else, Transition Region.

Next, considering the second group (at step 2), consisting of voiced/silent/transition regions, criteria at step 5 is evolved.

Step 5:

(a) If STE $>$ 0.01 and ZEP $>$ 10,

then Voiced Region.

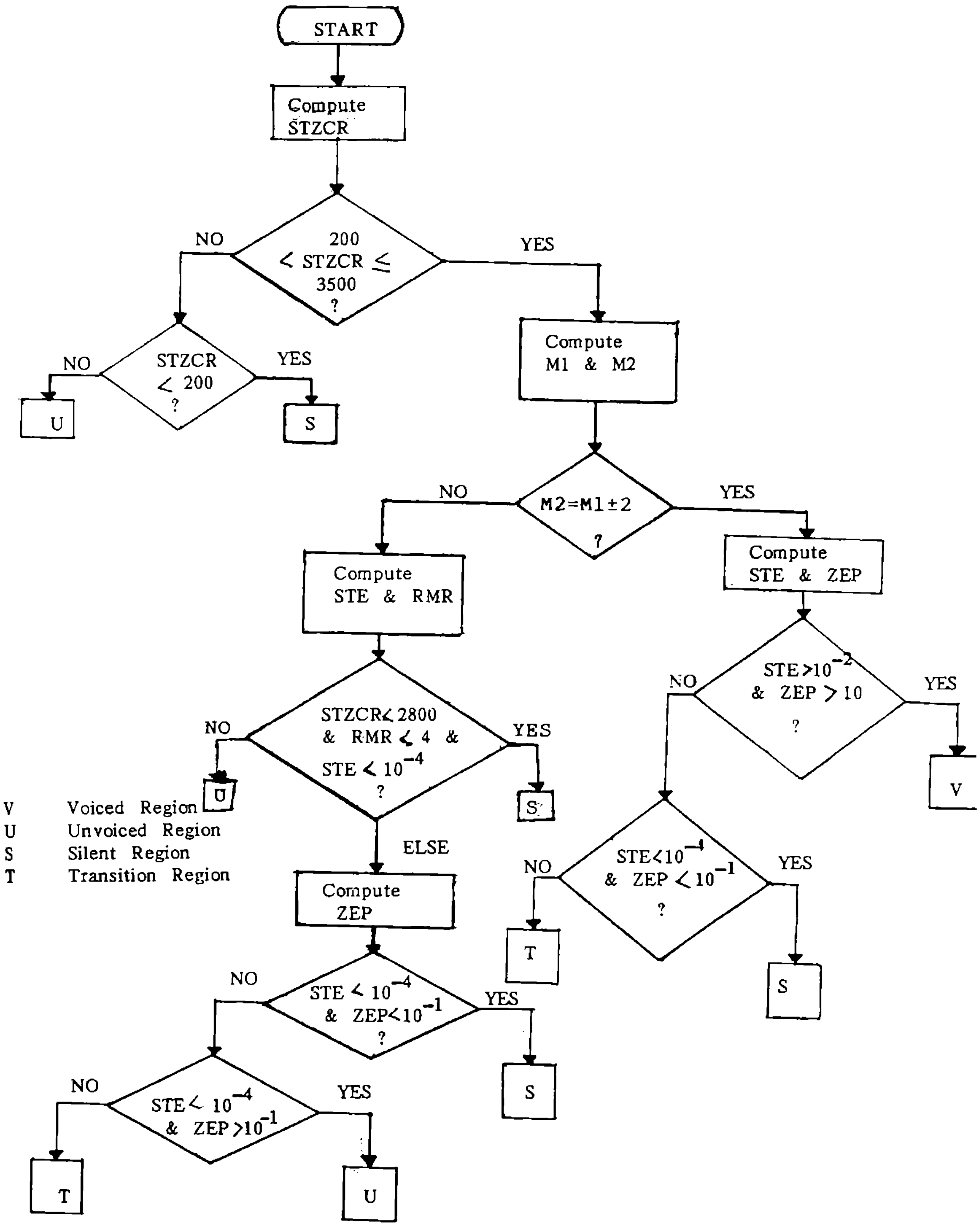


FIG.4.2 FLOW CHART FOR THE V/U/S/T DETECTION ALGORITHM I.

(b) If STE $<$ 0.0001 and ZEP $<$ 0.1,
then Silent Region.

(c) Else, Transition Region.

All the above steps need not be performed always. A flow chart for the above algorithm is shown in Fig.4.2

4.2.2 Algorithm II - Voiced/Unvoiced/Silent/Transition detection, based on correlation function

An elaborate classification as done in method I, is not always essential. Hence a simpler algorithm was developed, which also makes a voiced/unvoiced/silent/transition distinction. Algorithm II is explained by the following steps.

Step 1:

(a) If STZCR $<$ 200, then Silent Region.

(b) If STZCR $\geq$ 3500, then Unvoiced Region.

Step 2:

(a) If STE $<$ 0.0001 and ZEP $<$ 0.1, then Silent Region.

(b) If STE $<$ 0.0001 and ZEP $>$ 0.1, then Unvoiced Region.

Step 3:

(a) IF NCC1 $<$ 0.3, then Unvoiced Region.

(b) If NCC1 $>$ 0.3 and ZEP $>$ 10, then Voiced Region.

(c) If NCC1 $>$ 0.3 and ZEP $<$ 10, then Transition Region.

NCC1 is the normalised correlation coefficient for lag 1.

Step 1 is the same as in Algorithm I. Step 2 is intended to take care of those silent regions which may go undetected by step 1, due to their STZCR being greater than the threshold of 200. Also, those silent and unvoiced regions whose short-time energy may coincide, can be distinguished by using the energy ZCR product (ZEP) threshold set up in step 2.

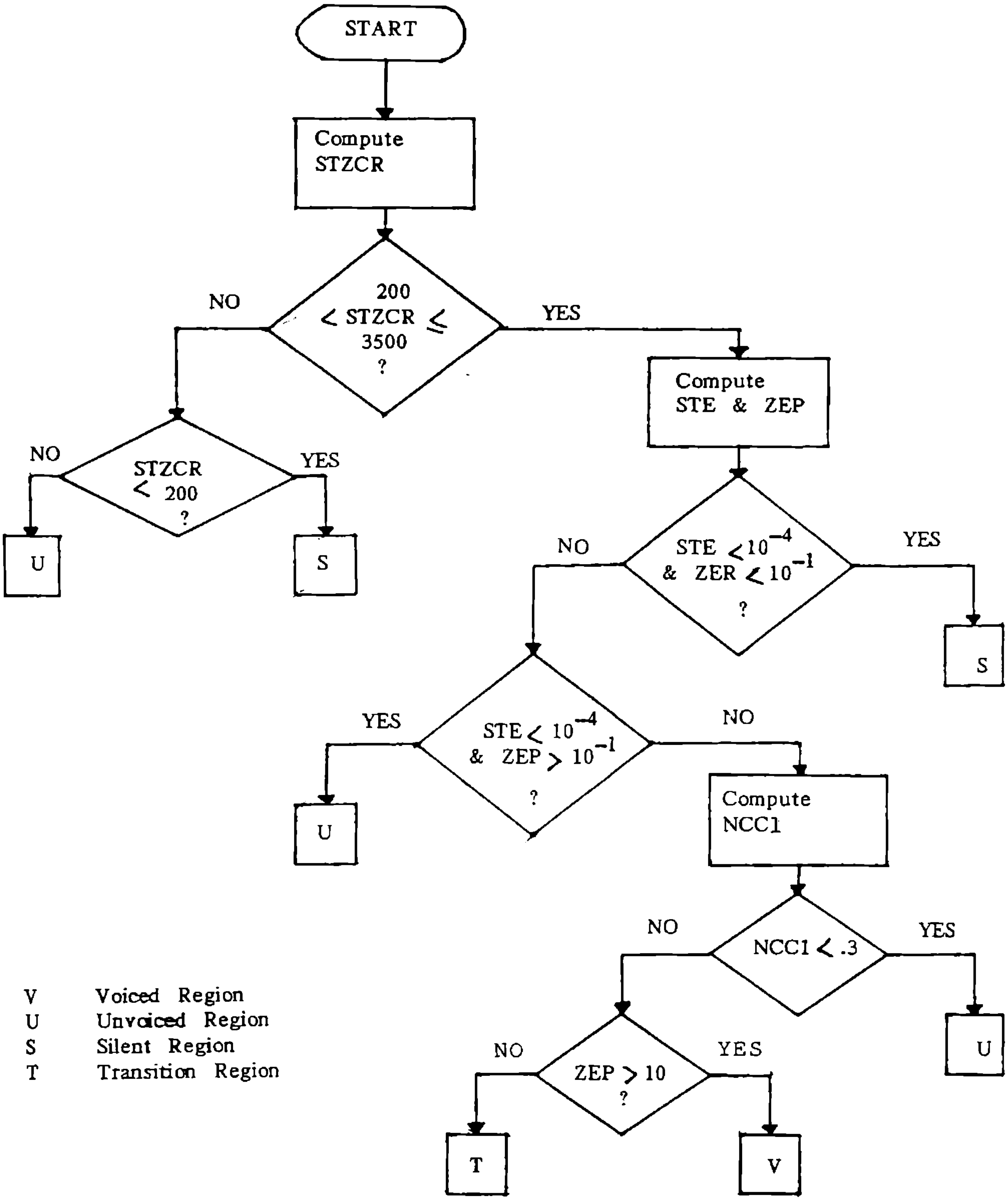


FIG.4.3 FLOW CHART FOR THE V/U/S/T DETECTION ALGORITHM II.

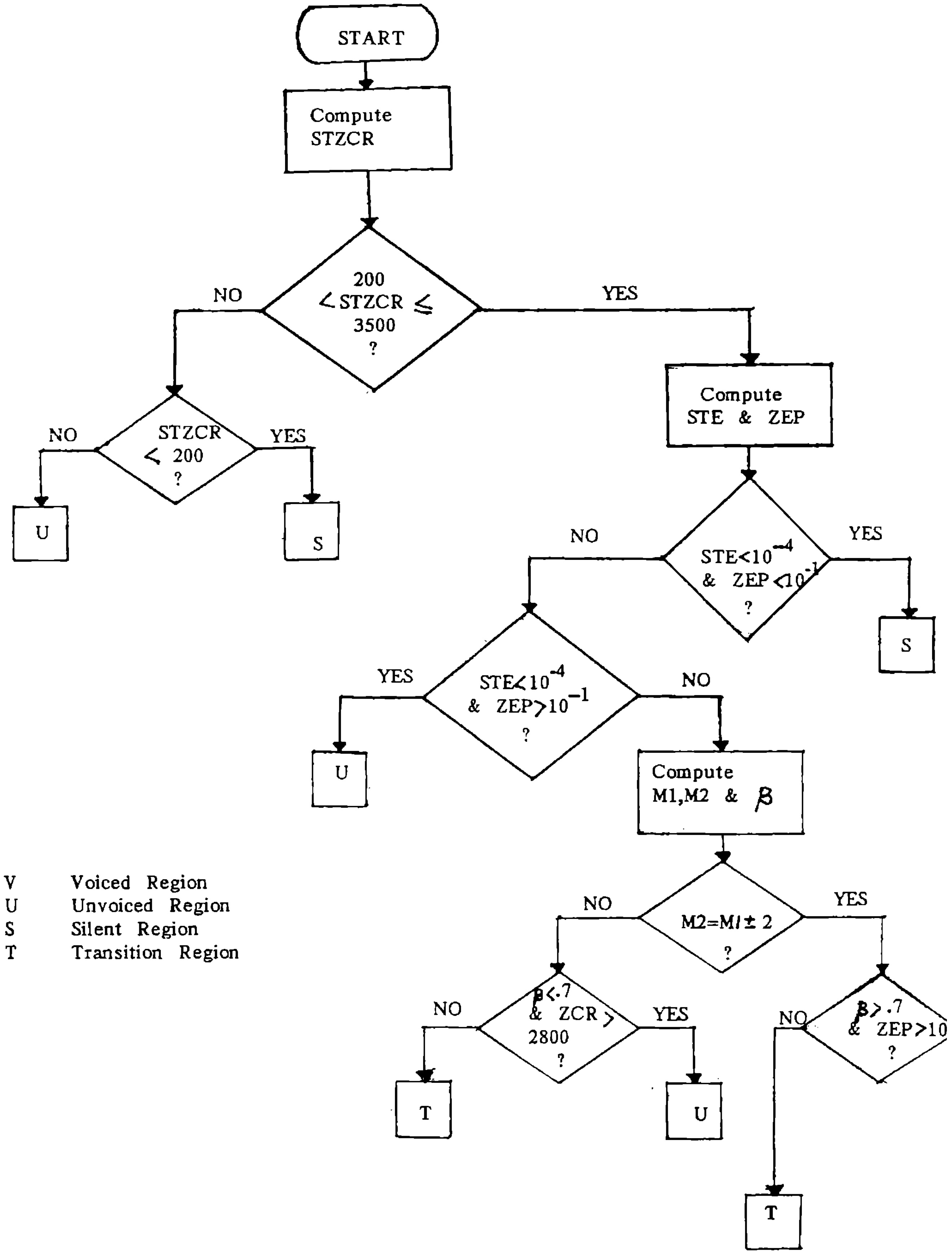


FIG.4.4 FLOW CHART FOR THE V/U/S/T DETECTION ALGORITHM III

Unvoiced speech signals are highly uncorrelated and the normalised correlation coefficient for lag 1, for such signals is found to have a negative value or a low positive value below 0.3 (Please refer Table 4.3). Data samples in the voiced region are highly correlated and hence the value of the correlation coefficient will be high. By fixing up a threshold value for ZEP, the transition regions can also be detected. The flow-chart for algorithm II is shown in Fig.4.3.

4.2.3 Algorithm III - Voiced/Unvoiced/Silent/Transition Classification based on a pitch correlation factor

In this approach, a new parameter, a pitch correlation factor β , which relates to the change in amplitude of the signal from pitch-to-pitch, is introduced.

$$\beta = \langle S_n S_{n-M} \rangle_{av} \Big/ \langle S^2_{n-M} \rangle_{av}$$

where, the $\langle \; \rangle_{av}$ denotes averaging over all the samples in the block under consideration. This factor is nearly equal to unity for the highly correlated voiced speech segments. The steps involved are shown below, and the flow-chart is given in Fig.4.4.

Step 1:

(a) If STZCR $<$ 200, then Silent Region.

(b) If STZCR $\geq$ 3500, then Unvoiced Region.

Step 2:

(a) If $STE < 0.0001$ and $ZEP < 0.1$, then Silent Region.

(b) If $STE < 0.0001$ and $ZEP > 0.1$, then Unvoiced Region.

Step 3:

(a) If $M2 = M1 \pm 2$, then Voiced Region or Transition Region.

(b) If $\beta > 0.7$ and $ZEP > 10$, then Voiced Region.

(c) Else, Transition Region.

Step 4:

(a) If $M2 \neq M1 \pm 2$, then Unvoiced Region or Transition Region.

(b) If $\beta < 0.7$ and $STZCR > 2800$, then Unvoiced Region.

(c) Else, Transition Region.

4.2.4 Simulation of the Algorithms

Two sets of speech data were used in this simulation experiment (i) for the training session of the algorithm and (ii) for validating the algorithm.

The speech data used for the training session were those of the speech files M1, F1 and B3, shown in tables 4.1 and 4.2. These sentences were specially chosen, since they contain most of the essential sounds like vowels, nasals, fricatives, plosives etc. Hence a wider range on the different parameters involved can be had, so that the algorithm developed can be perfect. The data values of each of these sentences were normalised, such that, the maximum value of the sample in the whole set is unity. The normalised speech sequences were then divided into blocks of size 160 samples. (A total of 311 blocks were considered).

The different blocks were manually classified into voiced, unvoiced, silent and transition blocks, by plotting on a graphics VDU of the computer. The algorithms were implemented on a PC using Turbo Pascal routines.

To compute the STZCR, the number of zerocrossings in each block of length 160 samples is found as ZN. Then, the STZCR can be computed as the number of zerocrossings per second, using the formula,

$$\mathrm{STZCR} = \mathrm{ZN.f_s/160} \qquad (4.1)$$

where f_s is the sampling frequency.

Now, the STE of each block is computed as the variance of the 160 normalised samples in the block. The root mean square value (S_{rms}) and the mean value (S_{mean}) are also computed and the ratio of the RMS to mean value is expressed as

$$\mathrm{RMR} = S_{rms}/\mathrm{abs}(S_{mean}) \qquad (4.2)$$

To check for the periodicity of the signal waveform, values of M1 and M2 are computed as the number of samples within two consecutive peak positions of the correlation coefficients R(J). Here, unclipped normalised samples within a block are considered. The autocorrelation coefficient R(J) is computed as:

$$R(J) = \sum_{n=J+1}^{160} S_n S_{n-J} \qquad (4.3)$$

M1 is determined by noting the value of J for which R(J) is maximum, for values of J greater than 15 (The minimum value of J = 16 is chosen (i) to avoid the high values of R(J) around J = 0, which may have values higher than R(M1), and (ii) taking into consideration that the lowest pitch period is 2 msec [3, 43, 44]). Similarly M2 is found by considering J > M1+15. Direct calculation method was used to compute R(J). The values of the above

Table 4.6

The parameter values useful for Algorithm I

Region	File Name	STZCR	STE	ZER	M1	M2	RMR
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Silent	F1	0-100	$(.79-3.0)10^{-5}$	$(.5-2.0)10^{-3}$	16	15	1.0-1.08
	M1	500-2650	$(.12-2.2)10^{-5}$	$(.7-5.0)10^{-2}$	49	21	2.7-3.8
	B3	850-2650	$(.4-5.9)10^{-5}$	$(.9-9.0)10^{-2}$	47	39	1.25-3.9
Unvoiced	F1	3450-3700	$(1.2-1.6)10^{-3}$	3.04-6.07	42	16	4.2-6.3
	M1	3100-3700	$(13-39.6)10^{-3}$	40-124	30	31	8-13.5
		3000	$3.7x10^{-5}$	.11	36	39	2.4
	B3	3300-4500	$(.18-.79)10^{-3}$	1.9-2.8	53	15	7-11

(contd...)

(1)		(2)	(3)	(4)	(5)	(6)	(7)
Voiced	F1	200-2800	$(.1-12.3)10^{-2}$	10-166	34	34	14.1-28.0
	M1	1150-2900	$(.7-8.0)10^{-2}$	20-155	40	39	14.4-31.7
	B3	750-1700	$(1-18.6)10^{-2}$	19-223	49	49	14.8-53.8
Transition	F1	200-1350	$(.9-6.0)10^{-3}$	.4-7.0	31	31	2.0-4.4
	M1	1100-2750	$(.4-3.6)10^{-3}$	.96-8.2	71	75	40-45
	B3	300-1500	$(.8-12.0)10^{-3}$	.6-8.7	55	20	14.9-66.0

parameters in the different regions are tabulated in Table 4.6.

Using algorithm I, the training sequence was first checked. All the segments, corresponding to all the four regions were detected correctly. To check the validity of the algorithm, it was tried on a speech data base 'outside' the training sequence.

The sentences in Tables 4.1 and 4.2, other than F1, M1 and B3 used for the training session, were considered. Each of these utterances were also divided into blocks of 160 samples, and were assigned as voiced/unvoiced/silent and transition regions, by manually inspecting the waveform. Detection of the various regions was then carried out using the algorithm, and the results were compared with those obtained by manual classification. It was observed that the total measured error probability of the algorithm is 4.298% (A total of 884 segments were considered).

The same above processes were done using algorithm II also. The parameter, normalised correlation coefficient for lag 1, was computed as

$$\mathrm{NCC1} = \sum_{n=2}^{160} S_n S_{n-1} \Big/ \left\{ \sum_{n=1}^{160} S_n^2 \cdot \sum_{n=2}^{160} S_{n-1}^2 \right\}^{\frac{1}{2}} \qquad (4.4)$$

Here, the measured error probability is 3.506%.

Algorithm III was next tested. The parameter β was computed using the expression,

$$\beta = \sum_{n=M+1}^{160} s_n s_{n-M} \Big/ \sum_{n=M+1}^{160} s_{n-M}^2 \qquad (4.5)$$

Here also, the training sequence was first checked algorithmically, and then the speech data outside the training sequence were used to validate the algorithm. The total measured error probability was obtained as 0.0% and 2.602%, for 'inside' and 'outside' the training sequence, respectively.

To summarise, three efficient algorithms, in the time domain, for the classification of the four different regions in a speech waveform, are presented. Compared to the simple algorithms for voiced/unvoiced detection, these algorithms require more number of computations, due to the introduction of the periodicity checking. But the simple ACF periodicity check has been found to be enough. The path length required for the detection of the different blocks vary, depending on the ranges of the values of the different parameters computed for that block. The third algorithm is found to give a better performance from the point of view of error probability. Hence the third algorithm is employed in the modified coder, for the voiced/unvoiced/silent/transition classification.

4.3 Simulation of the Modified Block Adaptive Coder (MBAC)

4.3.1 Simulation Description

Speech signals sampled at 8 kHz and encoded to 12 bits are used. As mentioned earlier, based on the existence of the correlation between the speech samples and also on the maximum pitch period of humans, the data samples are grouped into blocks of 160 samples.

It is reported by Harris [86] that it is a standard technique to use a tapered window function for weighting the data prior to spectral analysis, to reduce the leakage from other parts of the spectrum. Later, Paliwal [87] has proved that spectral flattening techniques like centre-clipping and inverse filtering deteriorates the pitch estimation performance. In the present work, care is taken not to use overlapping windows. This reduces the chances of mixing up of different regions within a block. If non-overlapping rectangular windows of specific length are used, wherever the full block does not belong to a particular region, based on the majority of the samples in the block, it can be correctly classified. This allows the correct and easy processing of the blocks further.

The blocks are next classified as voiced/unvoiced/silent or transition region, using the third algorithm explained above. If the block is detected as voiced, the first step is to determine the number of samples M in one pitch period of the signal. This is done by locating the position of the maximum value of the ACF's R(J), of the samples, for lags J above 15. The block length is now fixed as $N = 4M$, and β the correlation coefficient is calculated using the expression given by

$$\beta = \langle S_n S_{n-M} \rangle_{av} \;/\; \langle S^2_{n-M} \rangle_{av} \qquad (4.6)$$

where $\langle\ \rangle_{av}$ denotes the averaging over all the samples in the block, (N = 4M).

The block of $\{V_n\}$ samples where $V_n = S_n - \beta S_{n-M}$, are next calculated.

The autocorrelation coefficients R(J) for J = 0 to p are evaluated using the equation

$$R(J) = \langle V_n \cdot V_{n-J} \rangle_{av} \qquad (4.7)$$

p is taken as 4. The matrices ϕ and ψ are next formed from the coefficients R(J). ϕ is the p by p autocorrelation matrix and ψ is a p by one vector.

The predictor coefficients a_k's are obtained by solving for the matrix **a** in the matrix equation,

$$\phi\, \mathbf{a} = \psi \qquad (4.8)$$

The predictor parameters–a_k's, β and M –, and the first (p+M) samples are transmitted to the decoder, after suitably encoding them.

At the decoder, the parameters are decoded and used for the prediction of the sample values. The predicted value of the first (p+M) samples is the decoded version of itself. From n = (p+M+1) onwards, the sample values are predicted based on the earlier predicted values. The error in prediction and hence the SNRSEG are evaluated.

The predictor at the receiver is updated by transmitting the predictor parameters afresh every block.

If the block of samples under consideration is unvoiced, the block length is fixed as 40 samples and β is reduced to zero (ie., use spectrum predictor alone).

If the block is silent, no processing is needed and only the code indicating the region is to be transmitted.

If the segment considered belongs to the transition region, then the processing is the same as with the voiced segment, but with block length N = 2M and order of predictor p = 8.

4.3.2 Rate of Transmission of the Predictor Parameters and Side Informations

It has already been established that the reflection coefficients [2, 44] or frequency and bandwidth of the poles [14] can be effectively

Table 4.7

Analysis results of percentage content of the different regions in a phonetically well-balanced sentence

Speech file	Total No. of segments	No.of segments under				Percentage content of			
		VR	UR	SR	TR	VR	UR	SR	TR
M1	112	53	8	20	31	47.32	7.14	17.86	27.68
M4	120	52	12	20	36	43.33	10.00	16.70	30.00
M6	140	68	6	33	33	48.60	4.29	23.50	23.50
F1	112	62	8	19	23	55.36	7.14	16.96	20.54
F4	130	52	5	27	45	40.00	3.85	20.77	34.62
F6	135	54	5	38	38	40.00	3.70	28.15	28.15
B3	100	32	7	31	30	32.00	7.00	31.00	30.00
		Average				43.80	6.16	22.13	27.78

quantized and transmitted, without producing any perceptible effect on the synthesised speech. Atal et al [88] has reported that if predictor polynomial roots are used for transmission, 5 bits per root are adequate to preserve the quality of the synthesised speech so as to make it essentially indistinguishable from the speech synthesised from the unquantized parameters.

Based on the values, as suggested by Atal et al [88], the number of bits required per frame is 33 for voiced blocks, 53 for transition blocks, 62 for unvoiced frames and 2 for silent blocks, and the number of bits per second is 1.452 kb/s, 4.717 kb/s, 12.44 kb/s and 100 b/s respectively (Refer Appendix III).

From Table 4.7, it can be noted that on an average, in the phonetically well-balanced sentences chosen for the simulation study, the percentage of occurrences of the different regions are around, 40% for voiced, 10% for unvoiced, 20% for silent and 30% for transition regions. Hence, the total number of bits required to transmit the predictor parameters is 3.2379 kb/s (Computations shown in Appendix III).

Considering the transmission of the few difference samples ($D_n = S_n - S_{n-1}$) for every block, it has been verified in section 3.9.2 that, 3 bits per sample is adequate to retain almost same value of SNRSEG. (Refer Table 3.5). Hence, the number of bits required per frame, for the transmission of this side information, is 164 for voiced region, 176 for transition segments, 53 for unvoiced segments and nil for silent region. The corresponding bits per second are 7.216 kb/s, 15.664 kb/s and 10.60 kb/s respectively. The total number of bits per second required for the transmission of the side information is calculated to be 8.6456 kb/s; or the overall bit rate becomes 11.8835 kb/s. (Please refer Appendix III for details).

To brief the matters presented in this chapter, three algorithms for voiced/unvoiced/silent/transition classification and their simulation results are presented. The simulation procedure of the modified coder is also described.

The results of the computer simulation of the modified coder are presented in detail in the next chapter.

Chapter 5

SIMULATION RESULTS OF THE MODIFIED BLOCK ADAPTIVE PREDICTIVE CODER

5.1 Introduction

The overall efficiency of any communication system depends on various aspects like the design simplicity of the coder/decoder, the computational load involved in the analysis of the input signal, the transmission capacity of the system and finally on the quality of the reconstructed signal. Taking all the above factors into consideration a modified block adaptive predictive (MBAC) coder was developed. This coder is explained in section 3.10. The reduction in computation and hardware complexity attained over the standard APC system has also been explained in chapter 3. The actual simulation of the coder is presented in chapter 4. The various results obtained from the simulation of the MBAC is described in this chapter.

5.2 The Design Procedure for the Block Adaptive Coding Algorithm

The complete design procedure for the block adaptive coding algorithm can be summarised as follows:

1. Classify the given segment into voiced/unvoiced/silent/transition region.

2. If the segment is voiced, select distant sample–based prediction (DSP)

method for analysis. Block length for processing is 4M, where M is the number of samples within one pitch period of the signal, and p, the order of the predictor is taken as 4.

3. If the segment is unvoiced, select near sample based prediction (NSP) for analysis. Block length is chosen as 40 samples and p as 12.

4. If the segment is a transition segment, then distant sample based prediction method, with N = 2M and p = 8, is chosen.

5. If the segment is silent, then no processing is needed.

6. Evaluate the predictor parameters.

7. Perform the prediction of the samples using the decoded predictor parameters, based on the earlier predicted values.

8. Evaluate the SNRSEG values to check for the quality of the reconstructed signal.

The flow chart of the MBAC design algorithm is shown in figure 5.1.

5.3 Experimental Results

The MBAC coder shown in Fig.3.12 was simulated on a

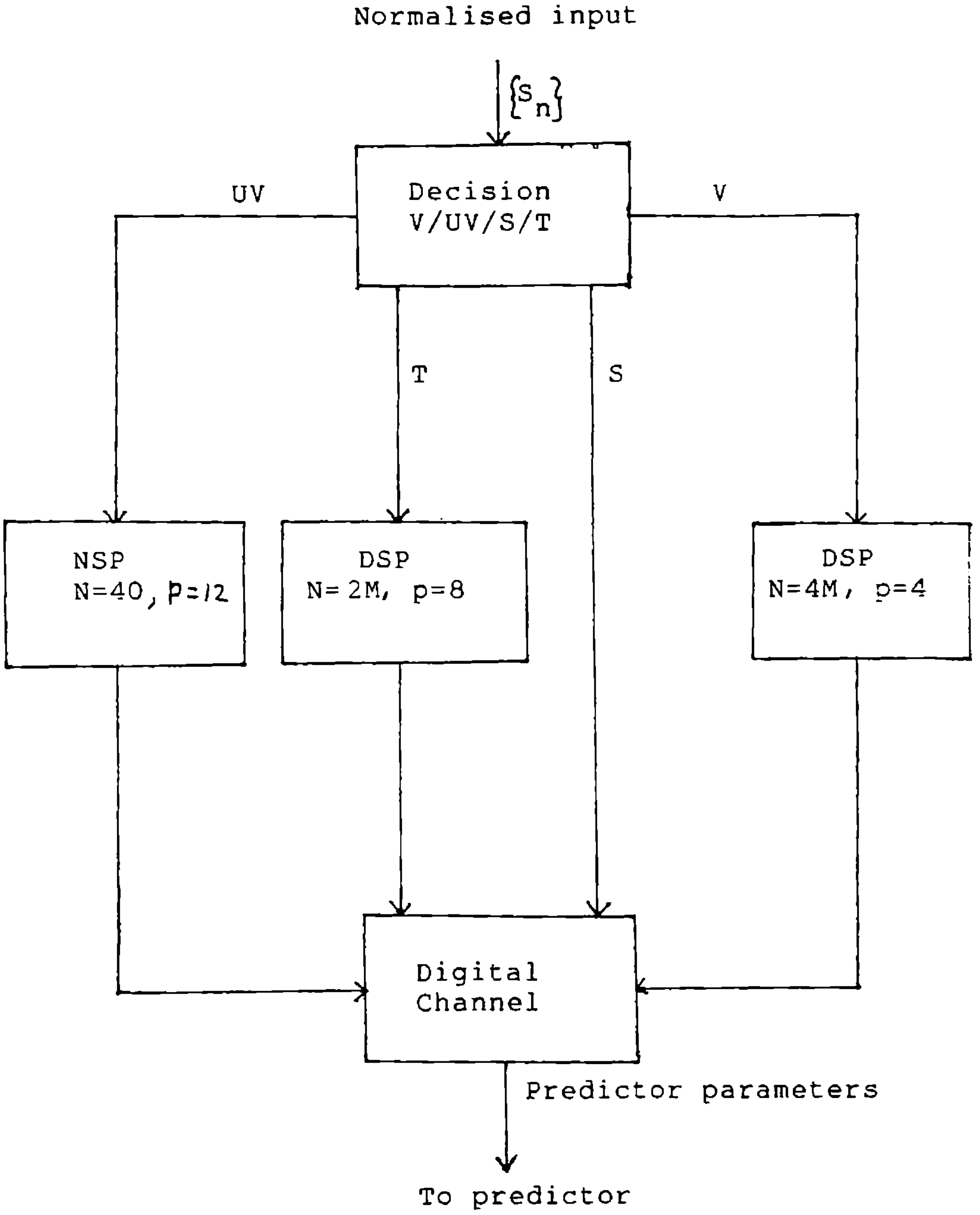


Fig.5.1(a) Flow-chart of the MBAC design algorithm
(a) Coder

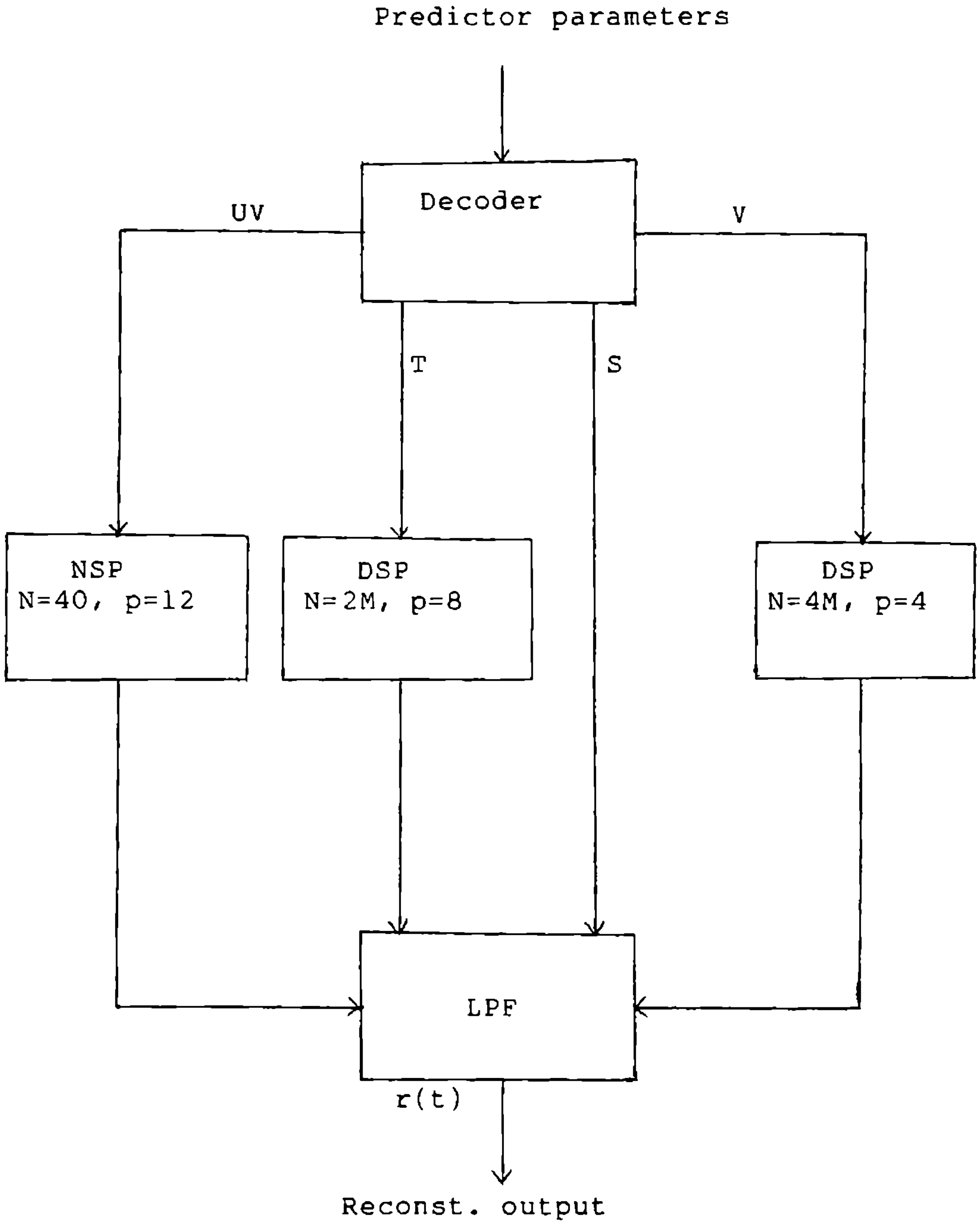


Fig.5.1 Flow-chart of the MBAC design algorithm
(b) Decoder

computer using Turbo Pascal routines. Initially the MBAC coder was tried on fully voiced speech segments. Regions corresponding to ten different voiced phonemes were chosen. In continuous speech applications, phonemes within a word suits the situation more than isolated phonemes. Hence such data obtained from the speech files shown in table 4.1 were considered.

The data values were normalised and grouped into segments of length 160 samples. The different predictor parameters– M, β and a_k's- were evaluated for each block, as mentioned in section 4.3. Using these parameters, the speech samples were reconstructed, based on the earlier predicted values and the SNR (dB) value was computed. This was repeated for all the segments, in the data set, and the average values were computed. The same was done using all the data sets. Speech data from three speakers - 2 male and 1 female were considered. A list of the values of the different parameters for the various speech sounds, of the different speakers are shown in table 5.1

It was noted that long vowels, like /ɔ/ in /dɔr/ 'door' and /u/ in /muv/ 'move' and /blu/ 'blue' show higher and steadier gain (around 12 to 16 dB) than short vowels like /ɑ/ in /drop/ 'drop' and /ʊ/ in /pʊʃ/ 'push' (around 12 to 14 dB). The phonemes /ɑ/ and /I/ showed still lesser gain (around 8 to 10 dB). The phoneme /I/ in 'right' gave the lowest gain (around 5 to 7 dB). The same pattern was followed for the different speakers though the gain varied slightly. These gain values clearly indicate the extent of correlation that exists within speech samples, in the

Table 5.1

Analysis-synthesis Results of Different Voiced Sounds, for Different Speakers

Sl. No.	Speaker	Speech Sound	Average values of		
			M	β	SNRSEG (dB)
(1)	(2)	(3)	(4)	(5)	(6)
1	M1	/ɑ/ in AFTER	51	0.93	8.316
	M2		57	1.00	8.418
	F1		35	0.98	8.402
2	M1	/ɑ/ in PARTY	47	0.94	8.903
	M2		57	0.92	9.589
	F1		37	0.93	8.064
3	M1	/o/ in 'DROP'	46	0.96	12.198
	M2		46	0.98	8.245
	F1		32	0.94	13.051
4	M1	/ɔ/ in 'DOOR'	45	0.97	16.739
	M2		51	0.98	10.861
	F1		33	0.98	12.877
5	M1	/o/ in 'COIN'	49	1.01	15.667
	M2		53	0.96	9.611
	F1		32	0.96	9.330

(contd...)

(1)	(2)	(3)	(4)	(5)	(6)
6	M1	/ʊ/ in 'PUSH'	44	0.94	11.690
	M2		46	0.96	12.690
	F1		28	0.98	14.795
7	M1	/u/ in 'MOVE'	45	1.01	15.503
	M2		46	0.98	12.025
	F1		32	0.98	16.314
8	M1	/u/ in 'BLUE'	43	0.99	15.331
	M2		46	0.98	15.331
	F1		32	0.96	14.665
9	M1	/i/ in 'SPEECH'	42	0.97	11.552
	M2		49	0.96	10.837
	F1		38	0.88	8.695
10	M1	/I/ in 'RIGHT'	45	0.99	5.404
	M2		50	0.98	7.042
	F1		36	0.90	5.036

M1: Male speaker 1, M2: Male speaker 2, F1: Female speaker.

Table 5.2

Results obtained on using the MBAC on the differenτ regions in a speech signal

Files	Regions	Av. SNR (dB)	Av. SNRSEG (dB)
1	V	13.3285	11.5548
	UV	2.1679	2.0843
	T	13.3640	12.0710
2	V	7.8869	7.4343
	UV	3.1992	2.7682
	T	12.5001	12.0186
3	V	10.4112	8.3678
	UV	2.4502	2.2596
	T	9.1378	6.0317
4	V	4.8139	4.7215
	UV	3.7593	3.0734
	T	8.8158	8.4415
5	V	11.2058	9.1901
	UV	2.1864	2.0067
	T	18.1590	17.6841
6	V	7.6269	7.1917
	UV	3.0022	2.7751
	T	5.9239	5.3228
7	V	6.8704	6.6623
	UV	5.9533	5.4161
	T	11.0812	10.3102

V: Voiced region, UV: Unvoiced region, T: Transition region

Table 5.3

Results of the Study of the MBAC on Continuous Speech Segments

Data set	1	2	3	4	5	6	7
SNR (db)	11.95	9.03	8.94	6.68	12.69	6.41	8.33
SNRSEG (dB)	10.58	8.57	6.73	6.46	11.48	5.94	7.87

case of the different phonemes and also in the different contexts.

The modified coder was next tried on the various regions present in speech signal. Voiced, unvoiced and transition regions of different speech files in Tables 4.1 and 4.2, were separated and used for the processing. It was noted that the average SNRSEG values for the voiced regions ranged from 6.7 dB to 11.6 dB (SNR ranged from 6.9 to 13.3 dB), for different voiced sounds and for different speakers. Similarly, transition regions obtained an SNRSEG value from 5.32 dB to 17.69 dB, while the unvoiced region showed a lower value, ranging from 2.08 to 5.42 dB. The results obtained from this experiment is tabulated in Table 5.2.

The next step was to study the effect of the modified coder on continuous speech. Speech signals of duration 1 to 2 secs. (from Tables 4.1 and 4.2) were used for this simulation study. As explained in section 5.2, the system was made block adaptive with N = 4M and p = 4 for voiced segments, N = 2M and p = 8 for transition regions and N = 40 (5 msec duration) and p = 12 for unvoiced segments. The predictor parameters were evaluated and a final reconstruction was done. The SNR values were computed for each block and a final average value (SNRSEG) was computed. The gain values obtained for different speech data are shown in Table 5.3. It can be noted that, on the whole, the SNR value varied between 6.41 dB and 12.69 dB. Figures 5.2(a to f) give a comparative study between the original waveform and the waveform reconstructed using the MBAC coder, for the voiced, unvoiced and transition segments. It can be noted that the

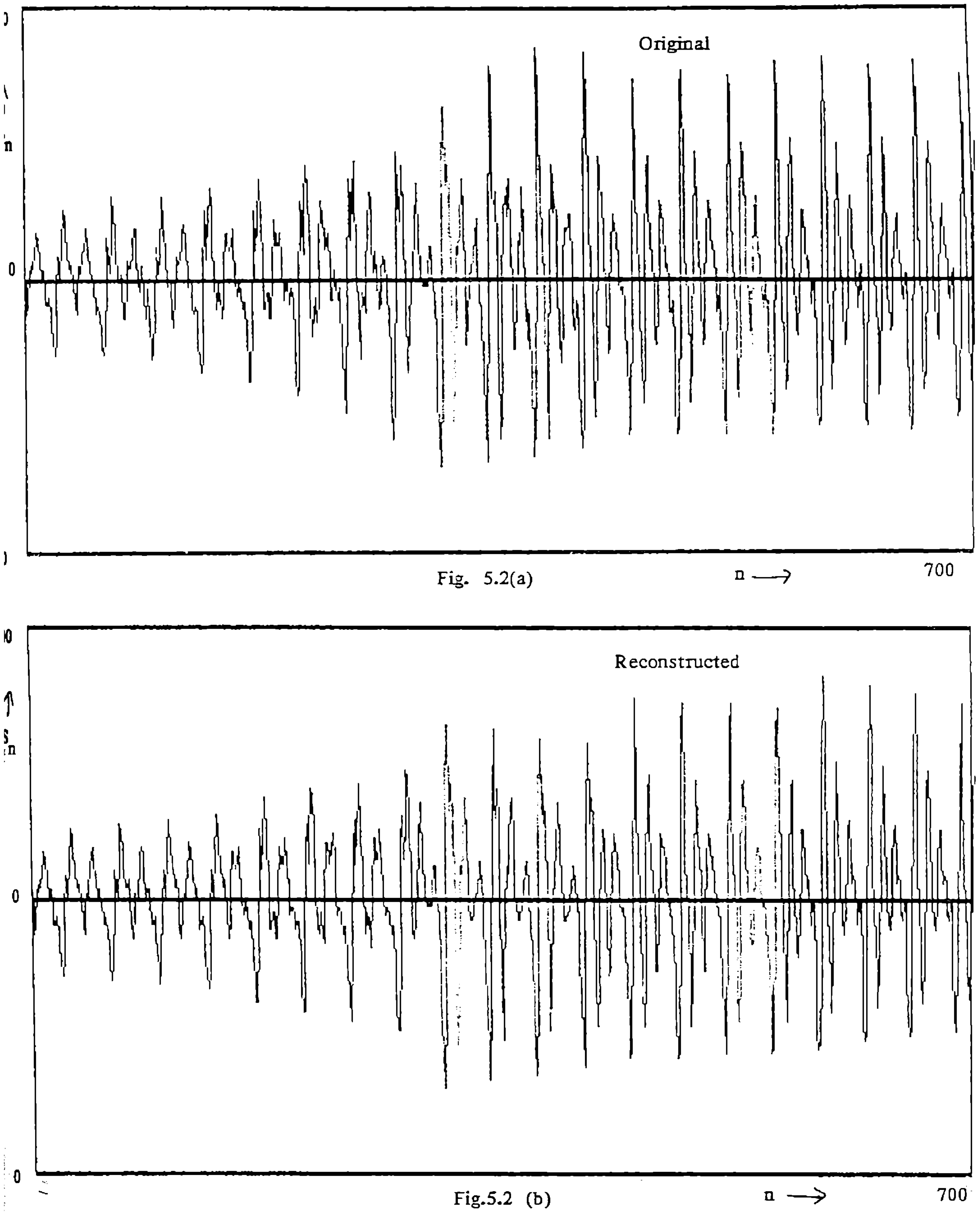


Fig. 5.2(a)

Fig.5.2 (b)

Fig.5.2(a&b) Plots of original and reconstructed waveforms - Using MBAC system - Voiced region.

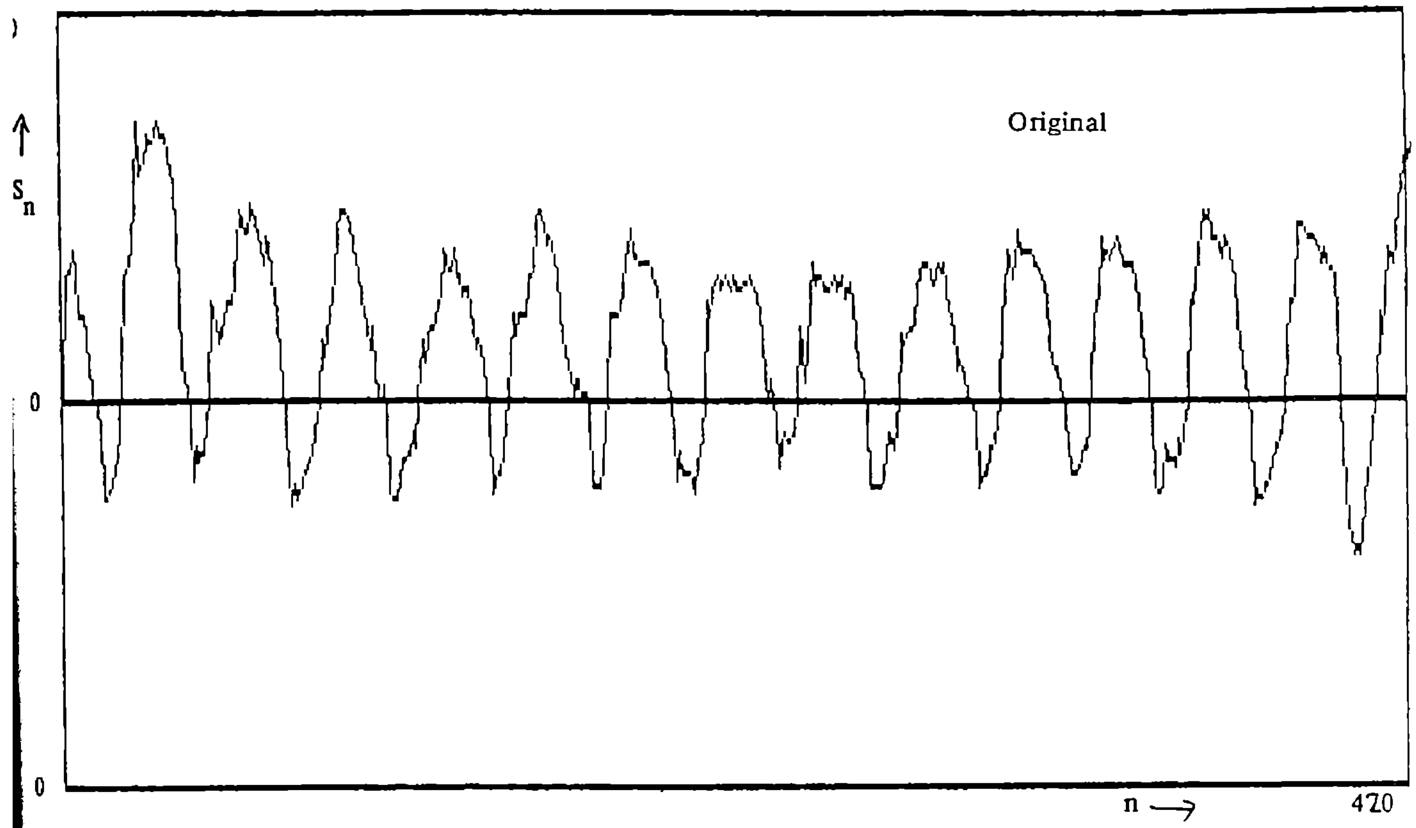


Fig.5.2(c)

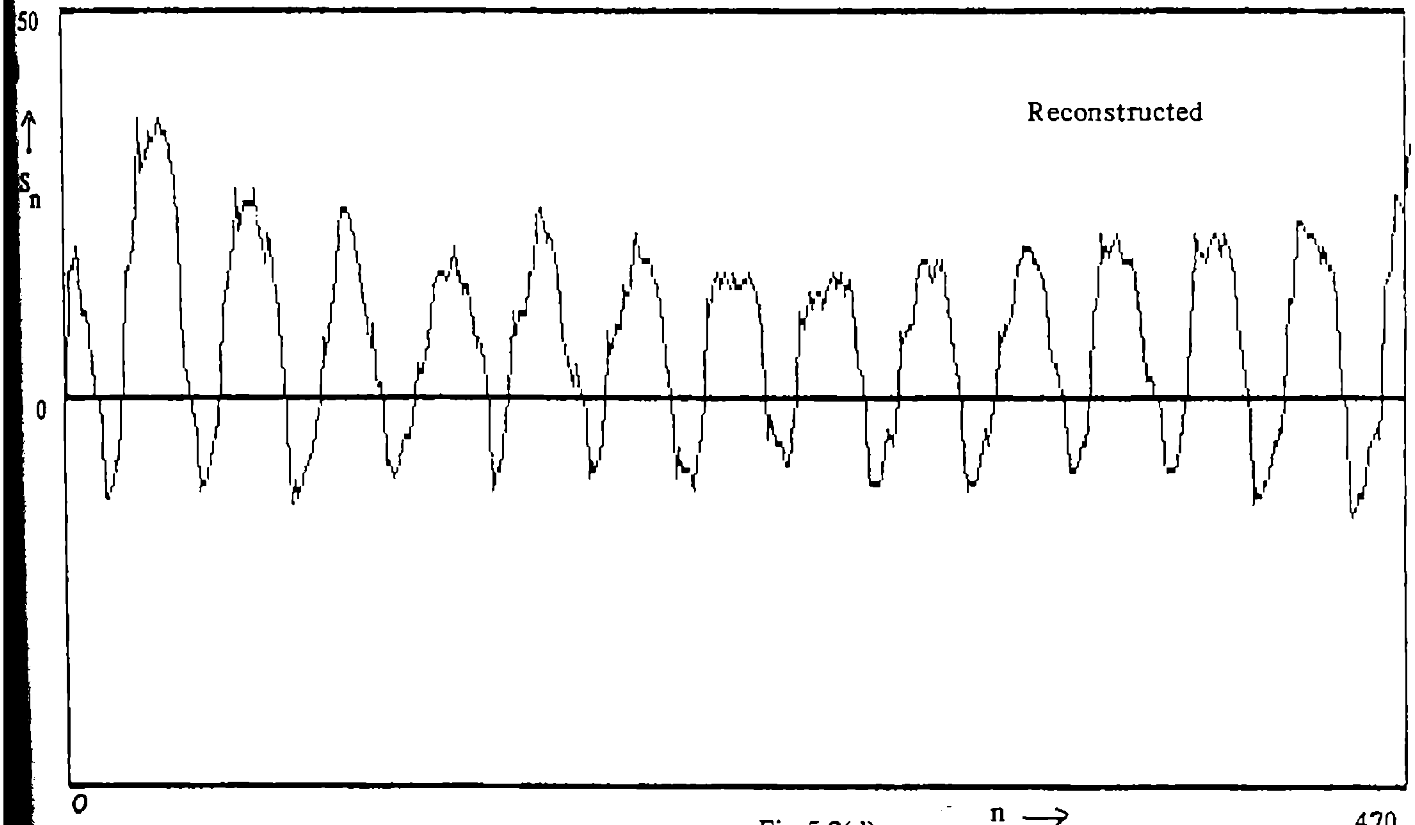


Fig.5.2(d)

Fig.5.2(c&d) Plots of original and reconstructed waveforms – Using MBAC system – Transition region.

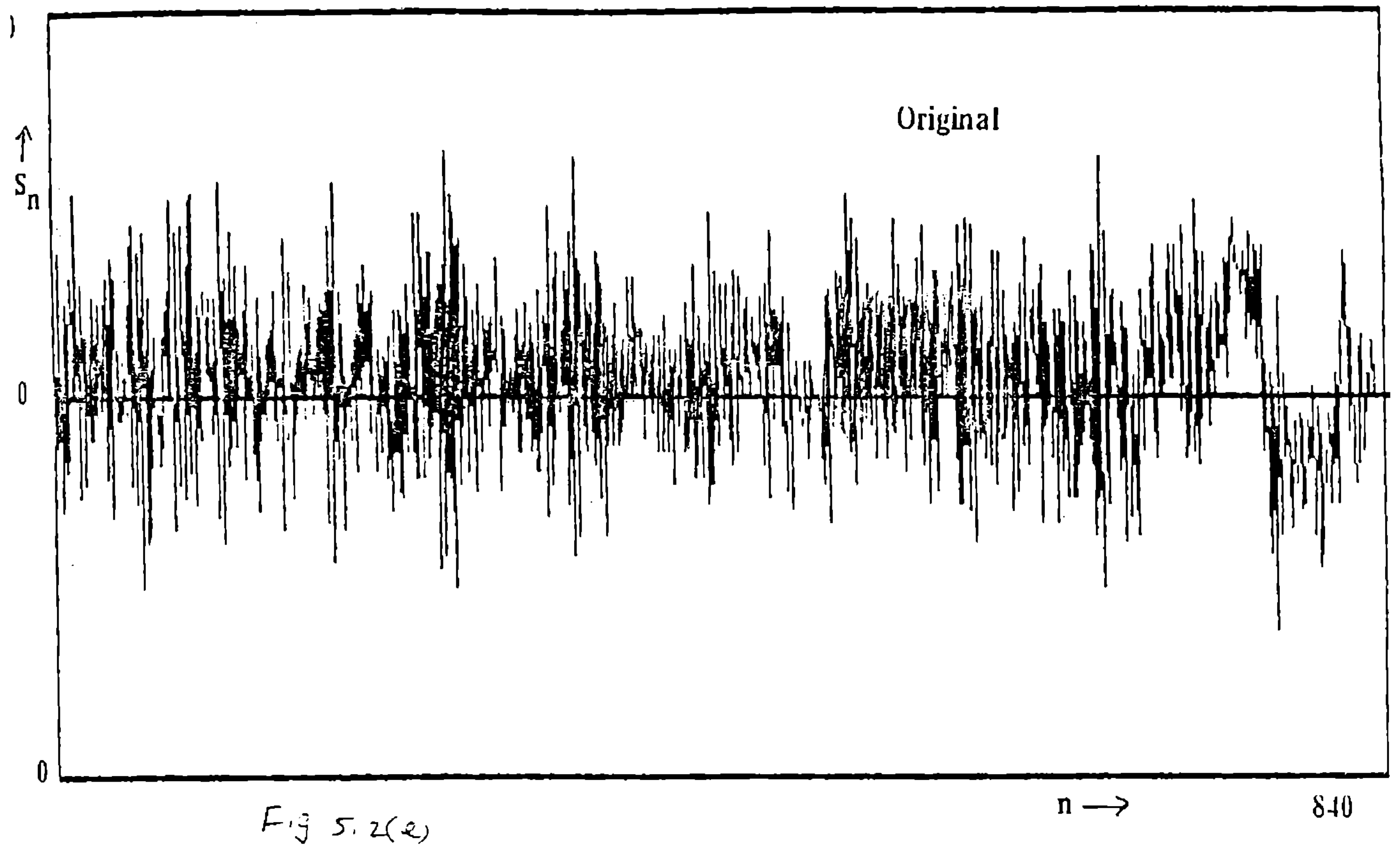


Fig 5.2(e)

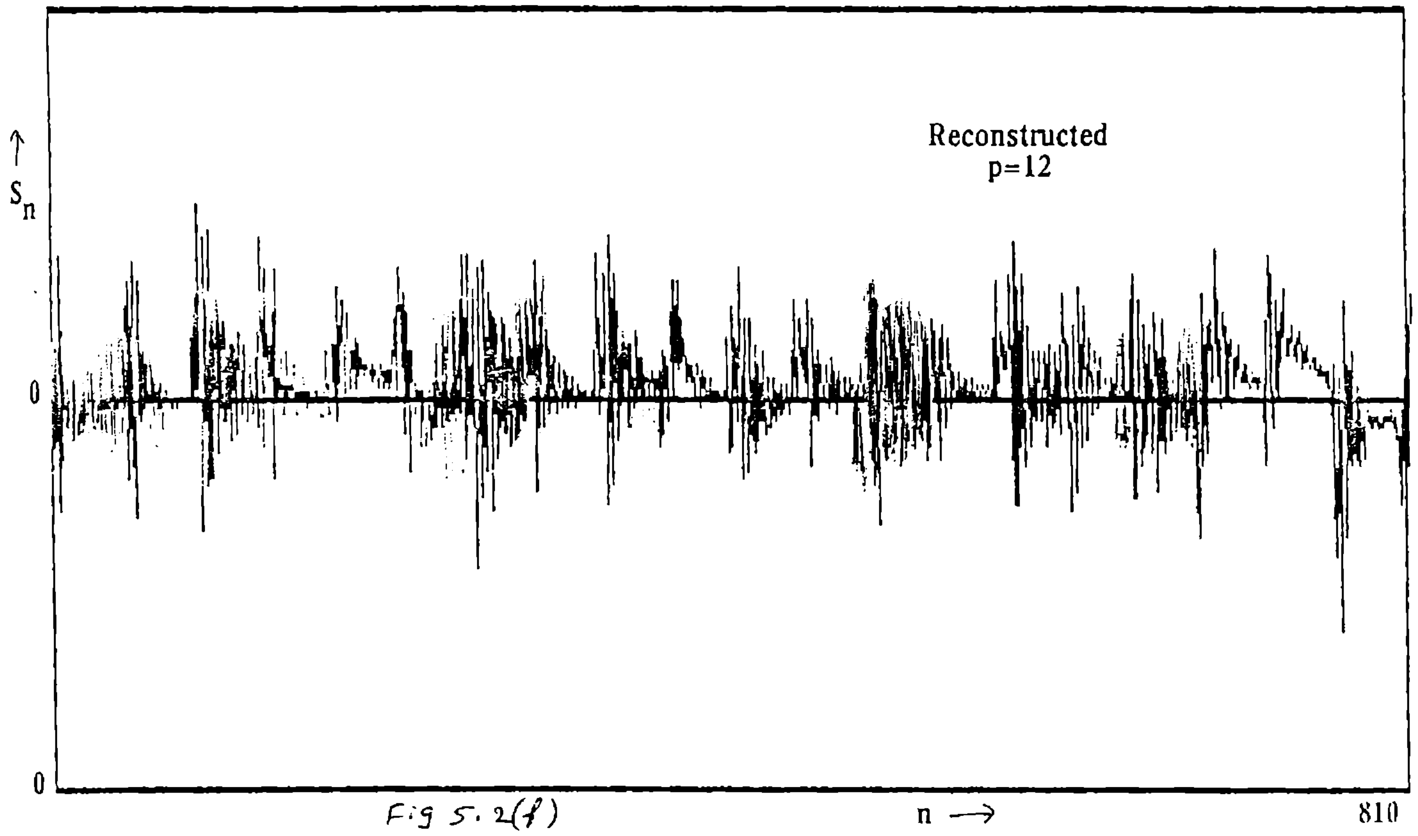


Fig 5.2(f)

Fig.5.2(e–f) Plots of original and reconstructed waveforms - Using MBAC system - Unvoiced region.

reconstructed waveforms in the voiced and transition regions are very similar to the original ones, proving the effifiency of the coder. The unvoiced segments, though very low in number, exhibited only lower quality of reconstruction.

A comparative study between an LPC system and the MBAC system is presented in Table 5.4. It can be noted that the MBAC system is far superior to the LPC system with a gain in SNR, of 5 to 8 dB when the same data samples were used. The MBAC system requires more computations and also the bit rate is higher, compared to an LPC system.

5.4 On the Applicability of MBAC System to Sounds in Malayalam

A trial on the feasibility of the coder on speech sounds, in our regional language, Malayalam, was thought of as a good idea. Phonemes in Malayalam, which are not normally present in English are considered. As a coder, the performance for this special set of phonemes was very good.

The special sounds like /r̲/ in /vaɾ̲a/ ’vara’, /ḷ/ in /vaḷa/ ’vala’, /ḻ/ in /maḻa/ ’mazha’, /ṇ/ in /tuṇ/ ’thoon’, /ḷ/ in /vaːḷ/ ’vaal’, /ṅ/ in /teːṅːa/ ’thaenga’, /n̄/ in /n̄aṇṭu/ ’njantu’, /ṇ/ in /kaṇaku/ ’kanaku’, were some of the phonemes chosen. Of these, some are common to Hindi. All the words chosen do not represent the full spelling of Malayalam words. They were so chosen as to study the interaction between the different sounds, in different contexts.

Table 5.4

A Comparative Study between an LPC and a MBAC

Parameters	System used	
	LPC	MBAC
1. Method of prediction	Spectrum prediction	Pitch prediction
2. Order of predictor, p	12	4 to 12 (adaptive)
3. Mode of excitation	Single pulse	Multi-pulse (First p or P+M samples known)
4. SNR (dB)	1.6 to 4.0	6.7 to 12.0
5. Bit rate	7.2 kb/s	11.88 kb/s
6. Complexity	Simpler	Medium complexity

Table 5.5

Representation of Malayalam sounds based on International Phonetic Alphabet (IPA)

With respect to place of articulation ⟶

With respect to manner of production ⟶

	Bilabial	Labio-dental	Dental	Denti-Alveolar	Alveolar	Retroflex	Palatal	Velar
Stop	പ് p ബ് b ഫ് ph ഭ് bh		ത് t ദ് d ഥ് th ധ് dh		ററ് t̲	ട് ṭ ഡ് ḍ ഠ് ṭh ഢ് ḍh	ച് c̄ ജ് ɟ ഛ് ch ഝ് ɟh	ക് k ഗ് g ഖ് kh ഘ് gh
Fricative					സ് s	ഷ് ṣ	ശ് s̄	ഹ് h
Nasal	മ് m		ന(ഉഈ്) n		(ഩ)ന n̲	ണ ṇ	ഞ n̄	ങ ṅ
Lateral					ല് l	ള് ḷ		
Trill				ര് r̲	റ് r			
Approximant (continuant)		വ് v				ഴ് l̲	യ് y	

Table 5.6

Results of Analysis -- Synthesis of Special Sounds in Malayalam that are not present in English

Sl. No.	Speech sound	Average values of			Average SNRSEG (dB)
		M	β	ρ_M , ρ_1	
(1)	(2)	(3)	(4)	(5)	(6)
1	/ḷ/ 'LA' in VALA	17	.95	.84, .81	11.920
	'ള' in വള				
2	/l̲/ 'ZHA' in MAZHA 'ഴ' in മഴ	16	.98	.84, .82	8.395
3	/r̲/ 'RA' in VARA 'റ' in വറ	20	.86	.76, .50	6.999
4	/ṇ/ 'ENN' in THOON 'ൺ' in തൂൺ	14	.99	.85, .85	21.168

(contd...)

(1)	(2)	(3)	(4)	(5)	(6)
5	/ṇ/ 'NA' in KANAKU 'ണ' in കണക്ക്	18	1.00	.86, .89	19.313
6	/ṇ/ 'NA' in KANA 'ണ' in കണ	17	.98	.82, .83	10.479
7	/n̄/ 'NJA' in NJANTU 'ഞ' in ഞണ്ട്	18	1.0	.88, .93	20.627
8	/ṅ/ 'NGA' in THAENGA 'ങ' in തേങ്ങ	16	1.0	.85, .60	15.470
9	/ḷ/ 'EL' in VAAL 'ൾ' in വാൾ	15	.95	.83, .50	9.414

(1)	(2)	(3)	(4)	(5)	(6)
10	/ḷ/ 'EL' in VAL 'ൾ' in വാൾ	20	.80	.70, .40	6.991
11	/kh/ 'KHA' in NAKHA 'ഖ്' in നഖ	17	.91	.87, .35	7.345
12	/gh/ 'GHA' in MAEGHA 'ഘ്' in മേഘ	17	.98	.85, .43	14.814
13	/ch/ 'CHA' in CHAYA 'ഛ്' in ഛായ	18	.95	.83, .44	11.886
14	/jh/ 'JHA' in JHANCY 'ഝ' in ഝാൻസി	18	1.0	.82, .85	11.641
15	/ṭh/ 'TTA' in PEETTA 'ഠ്' in പീഠ	16	.92	.85, .60	12.102

(1)	(2)	(3)	(4)	(5)	(6)
16	/ḍh/ 'DDA' in THRUDDA 'ഢ' in ദൃഢ	16	.92	.70, .50	9.798
17	/th/ 'THA' in RATHA 'ഥ' in രഥ	17	.98	.85, .49	15.972
18	/dh/ 'DHA' in DHANAM 'ധ' in ധനം	19	.96	.85, .82	14.212

Table 5.5 gives an idea of the different sounds in Malayalam based on the position and manner of articulation. Depending on the place of articulation, there are eight groups. They are bilabial, labiodental, dental, dentialveolar, alveolar, retroflex, palatal and velar. Based on the manner of production of the sound, they are grouped into stops, fricatives, nasals, laterals, trills and approximant.

The results of the analysis of these special sounds are shown in Table 5.6. It can he noted that the nasals /ṇ/ (ണ്‍, ണ) in /tuṇ/ 'thoon' and /kaṇaku/ 'kanaku', /ṅ/ (ങ) in /te:ṅ:a / 'thaenga' and /n̄/ (ഞ) in /n̄aṇṭu/ 'njantu' gave the maximum SNRSEG value, ranging from 15.47 dB to 21.168 dB. The sound /ṇ/ (ണ്‍) with the nasal /n/ following a vowel, gave the highest SNRSEG value, with the maximum going upto 30 dB. They were followed by the dental stops, /th/ (ഥ്) in /ratha/ 'ratha' and /dh/ (ധ്) in /dhanam/ 'dhanam', giving an SNRSEG value of 13.6 dB to 15.97 dB. The retroflex stops, /ṭh/ (ഠ്) in /pi:ṭha / 'peetta' and /ḍh/ (ഢ്) in /driḍha / 'dridda', obtained a gain of 9.798 dB to 12.10 dB, while the retroflex laterals, /ḷ/ (ള്) in /vaḷa/ 'vala' and retroflex approximant /l̲/ (ഴ്) in /maḻa/ 'mazha' gave a gain of 8.395 dB to 11.92 dB. The velar fricatives /ch/ (ഛ്) in /cha:ya/ 'chaya' and /jh/ (ഝ) in /jha:nsi/ 'jhancy' and the velar stops /kh/ (ഖ്) in /nakha/ 'nakha' and /gh/ (ഘ്) in /me:gha / 'maegha' obtained an SNRSEG of 7.345 dB to 14.814 dB. The results obtained, prove the fact that the nasals which are more sonorant than the other consonants exhibit higher gain. Also, among the stops, it is the voiced aspirated stops which show better gain values than the voiceless type.

It has been reported that [8,9,10,64,67] using self–excited vocoders and code excited LPC's, the SNRSEG values obtained are in the range 8.5 dB to 13.8 dB. The computations required for a CELPC is 80 MFLOPs and that for SEV is 4 MFLOPs [8], while the MBAC requires only $O(p^2+p)$ multiples and adds.

To summarise, the MBAC system developed, has comparable quality with that of an SEV or CELPC system in an SNR sense, with much lesser amount of computational burden. This is achieved that the expense of a moderate increase in bit-rate.

Chapter 6

A PROPOSAL FOR A SPEAKER RECOGNITION SYSTEM

6.1 Speaker Recognition

The problem of recognizing speakers from their voices were studied long back in 1937. These works were mainly based on human listening. Later, with the advent of computers, attention was focussed on automatic recognition of speakers (ASR).

When selecting features for the speaker recognition problem, it is essential to restrict the features to those which give discriminatory information about the different speakers. There are different methods for feature evaluation in speaker recognition problems. One among them is the 'knock out' method [89], where we select a subset of features from the main set of features available, by knocking out, one by one, the features which do not contribute to the recognition process. It has been reported by Su et al [90] that co-articulation between /m/ and /v/ have strong speaker dependence. Atal [91] has used 12 coefficients, including filter coefficients, impulse response, its autocorrelation, the area function and the cepstral coefficients of the filter. The essential features required for the speaker recognition problem can be selected by analysing the different salient features of the speech signal.

Speaker identification is the process of determining whether or not an utterance by an unknown speaker corresponds to stored versions of that utterance produced by a number of speakers. Speaker verification is to determine whether a speaker is who he claims he is. A text-dependent knowledge-based speaker recognition method is presented below.

While studying the different parameters evaluated, for voiced speech segments, in the course of the development of the modified block adaptive coder, it was noted that, by just checking the number of samples M, that are present in one pitch period, the speakers can be grouped into male and female. For each group, a range of values can be fixed for M, the normalised correlation coefficients ρ_1 and ρ_M and the signal-to-noise ratio; and this set is unique for each speaker.

6.1.1 Feature Extraction

The speech data corresponding to the phrases given in Table 4.1, were used in the simulation study, for the feature extraction. Voiced regions corresponding to the different vowels /a/, /i/, /I/, /o/, /ɔ/, /ʊ/, etc. contained in the above data files were manually separated by plotting on a graphic VDU, and were stored into different files.

For the "training session" of the system, an acoustic pattern, or a set of acoustic features is to be stored for each speaker. In the present method, the most effective speaker dependent feature chosen, is the value of M. As mentioned in chapters 3 and 4, the data are grouped into segments of length 160 samples and the ACFs R(J) are computed. The value of M is determined by locating the position of the maximum value of the ACF R(J), for lags between 16 and 80. Next, fixing the number of samples in a block as N = 4M, the predictor parameters β and a_k's, and the correlation functions ρ_1 and ρ_M are evaluated. Fixing the order of the predictor as p = 8, the predicted sample values and hence the SNR values G are determined. This is repeated for the different blocks of samples present, and a set of average values is computed. The results of the above evaluations are shown in Table 6.1

On examinining the various parameters of the predictor, shown in Table 6.1, it is observed that they provide good discriminating capacity, particularly between male and female speakers and can assist in speaker recognition. Using the predictor parameters, and also certain standard features of speech signals, certain facts and rules can be formulated to constitute a knowledge - base and this knowledge - base can be used for speaker identification. The values for a large number of phonemes for three speakers- 2 male and 1 female - were studied.

Two methods have been formulated for speaker recognition. In method 1, male/female classification is done, by fixing the range of values for M. It has been quoted that [3,43,44] under natural speaking conditions, the pitch period of humans vary from 2.5 msec. to 20 msec. In the present study, the maximum pitch-period is only around 8 msec, and hence the range of values of M is fixed as explained below. If M is in the range 20 to 40, the input signal belongs to a female speaker; if in the range 41 to 80, it belongs to a male speaker, provided sampling has been done at the Nyquist

Table 6.1

Values of the parameters which are useful in speaker recognition

Speech Sound	M & β			ρ_1, ρ_M, ρ_1/ρ_M			SNR $G = P_s/P_d$ (dB)		
	M1	M2	F1	M1	M2	F1	, M1	M2	F1
/ɑ/ in AFTER	51 .93	57 1.0	35 .98	.705 .868 .812	.820 .957 .888	.677 .837 .809	8.316	8.418	8.402
/ɑ/ in PARTY	47 .94	57 .92	37 .93	.750 .953 .791	.820 .810 1.04	.710 .810 .877	8.903	9.589	8.064
/o/ in DROP	46 .96	46 .98	32 .94	.790 .960 .823	.855 .838 1.02	.710 .823 .863	12.198	8.245	13.051
/ɔ/ in DOOR	45 .97	51 .98	33 .98	.830 .974 .852	.874 .844 1.04	.853 .845 1.01	16.739	10.861	12.877
/o/ in COIN	49 1.0	53 .96	32 .96	.827 .983 .841	.833 .830 1.00	.590 .833 .708	15.667	9.611	9.330

(Contd....)

Speech Sound	$\dot{M}$, β			ρ_1, ρ_M, ρ_1/ρ_M			SNR $G_1 = P_s/P_d$ (dB)		
	M1	M2	F1	M1	M2	F1	M1	M2	F1
/ʊ/ in PUSH	44 .94	46 .96	28 .98	.895 .973 .920	.946 .863 1.10	.898 .854 1.05	11.690	12.690	14.795
/u/ in MOVE	45 1.0	46 .98	32 .98	.883 .983 .875	.895 .843 1.06	.885 .850 1.04	15.503	12.025	16.314
/u/ in BLUE	43 .99	46 .98	32 .96	.886 .984 .900	.904 .848 1.07	.890 .847 1.05	15.331	15.331	14.665
/i/ in SPEECH	42 .97	49 .96	38 .88	.703 .965 .729	.913 .827 1.10	.717 .813 .882	11.552	10.837	8.695

M1 - Male Speaker 1 ; M2 - Male Speaker 2 ; F1 - Female Speaker 1 rate of 8 KHz. Now, if M is from 28 to 38, the speaker is female speaker F1, if M is in the range 42 to 45 the speaker is Male speaker M1; and if it is in the range 52 to 57 it is Male speaker M2. If M is from 46 to 51, check for ρ_M and ρ_1. If $\rho_M > \rho_1$, then the speaker is M1, if not, it is M2. The above algorithm in the form of flow chart is shown in fig.6.1 and the knowledge-base is shown in Table 6.2.

This method works well for a small ensemble of speakers.

In method II, provision for including more speakers and more phonemes is considered. Initially, as in method I, male/female classification is done by noting the value of M. The phonemes are then classified into four groups, as shown in Table 6.3, by fixing up a range of values for ρ_1. The range of values of M, correlation coefficient ratio $\rho' = \rho_1/\rho_M$ and G for different speakers are then used for the final speaker identification.

For group I, consisting of the phoneme /i/, ρ_1 = .70 ± .003, for group II, containing the phoneme /ɑ/, ρ_1 = .73 ±· 02, for group III, with phonemes /o/ and /ɔ/, ρ_1 = .83 ± .04, and for group IV having phonemes /ʊ/ and /u/, ρ_1 = .915 ± .035.

It is seen that in group I, if $\rho' = \rho_1/\rho_M$ = .73 ±.01 and M = 42 ± 1, then it belongs to speaker M1. If ρ' = 1.1 ± .1 and M = 50 ± 1 then, it is male speaker M2. In group II, if ρ' = .80 ± .01 and M = 49 ± 2, then it belongs to speaker M1. If ρ' = 1.03 ± .1 and

Table 6.2

KNOWLEDGE-BASE FOR THE ALGORITHM OF METHOD I

(M (NO. OF SAMPLES IN 1 PITCH PERIOD OF THE INPUT SIGNAL))
(ρ_1, ρ_M (NORMALISED CORRELATION COEFFICIENTS FOR LAGS 1 & M))
(M1, M2 (MALE SPEAKERS 1 & 2))
(F1 FEMALE SPEAKER 1))

10 IF $M \neq 50 \pm 30$
THEN ERROR IN INPUT

12 IF $M = 50 \pm 30$
THEN HUMAN VOICE

14 IF HUMAN VOICE = YES AND $M < 41$
THEN FEMALE SPEAKERS

16 IF FEMALE SPEAKERS = YES AND $28 \leq M \leq 38$
THEN FEMALE SPEAKER F1

18 IF HUMAN VOICE = YES AND $M > 40$
THEN MALE SPEAKERS

20 IF MALE SPEAKERS = YES AND $42 \leq M \leq 45$
THEN MALE SPEAKER M1

22 IF MALE SPEAKERS = YES AND $52 \leq M \leq 57$
THEN MALE SPEAKER M2

24 IF MALE SPEAKERS = YES AND $46 \leq M \leq 51$ AND $\rho_M > \rho_1$
THEN MALE SPEAKER M1

26 IF MALE SPEAKERS = YES AND $46 \leq M \leq 51$ AND $\rho_M < \rho_1$
THEN MALE SPEAKER M2 .

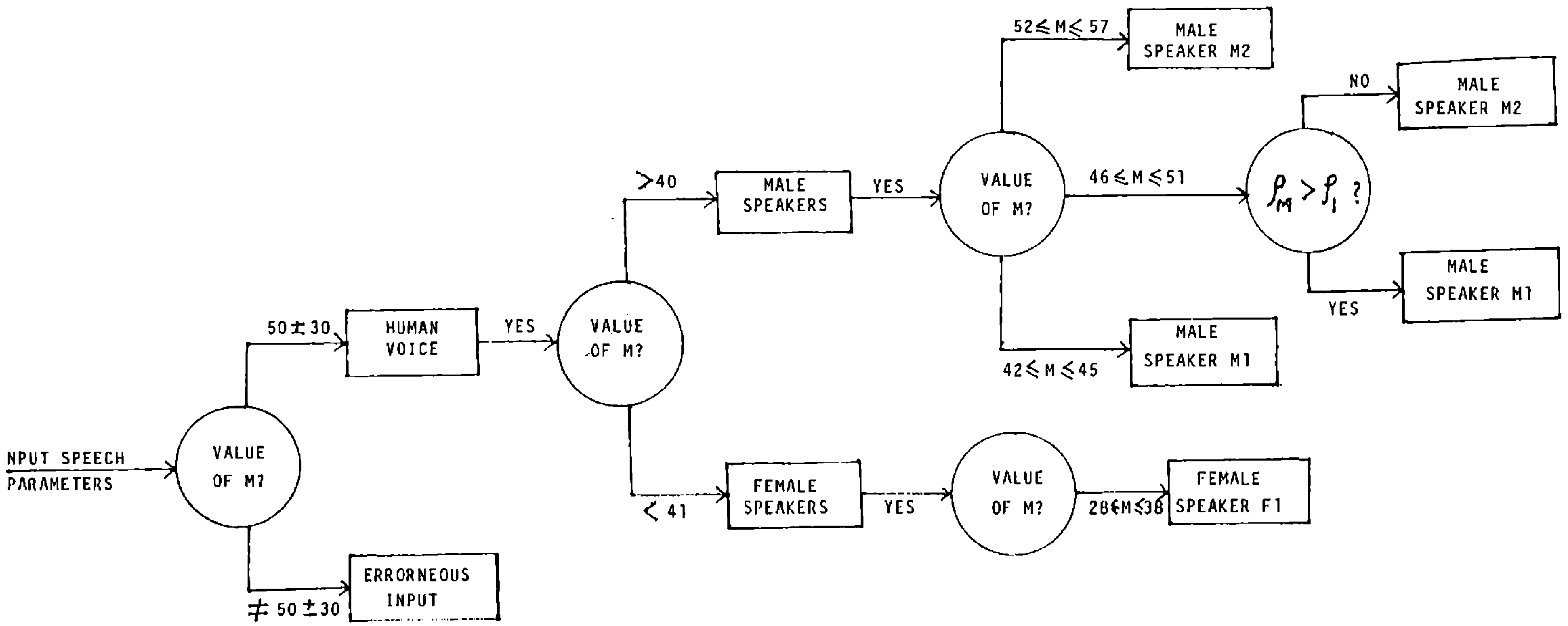


Fig.6.1 Flow-chart for the Speaker Recognition Algorithm of method I

Table 6.3

Parameters for Male Speaker Recognition in method II

Group	ρ_1	For Speakers M1 and M2		
		ρ_1	ρ_1 / ρ_M	M
I /i/	.70±.003	.70±.003	.73±.01 1.10±.10	42±1 50±1
II /ɑ/	.73±.020	.73±.020	.80±.01 1.03±.10	49±2 57±1
III /o/,/ɔ/	.83±.040	.81±.020 .855±.025	.835±.015 1.02±.02	47±2 50±4
IV /ʊ/, /ɥ/	.915±.035	.885±.005 .925±.025	.90±.02 1.08±.02	44±1 46±1

$M = 57 \pm 1$, then it is M2. In group III, if $\rho_1 = .81 \pm {\cdot}02$ and $\rho' = .835 \pm .015$ and $M = 47 \pm 2$, then it is M1. If $\rho_1 = .855 \pm .025$ and $\rho' = 1.02 \pm .02$ and $M = 50 \pm 4$, then it is speaker M2. In group IV, if $\rho_1 = .885 \pm .005$ and $\rho' = .90 \pm .02$ and $M = 44 \pm 1$, then it is M1. If $\rho_1 = .925 \pm .025$ and $\rho' = 1.08 \pm .02$ and $M = 46 \pm 1$, then it is M2.

The flow chart and the facts and rules relating to this identification method are given in Fig.6.2 and Table 6.4 respectively.

In a similar way, if values beyond the above given ranges appear at different stages in the algorithm, while recognising new input signals, then they can also be added to the knowledge base, as shown in Table 6.4, to automatically adapt it to new speakers.

It has been verified that the above mentioned algorithm is very simple and effective when the number of speakers considered is small. Further work has to be done using more speakers and still more phonemes, to prove the efficiency of the system. The author feels that it would also give good recognition score to a wider class.

The above methods have not been tested and verified to a greater extent, and sessions 'outside' the training sequence were not considered. This is beyond the scope of this thesis, because speaker recognition as such, is a wide and deviated field from speech recognition. This is only a

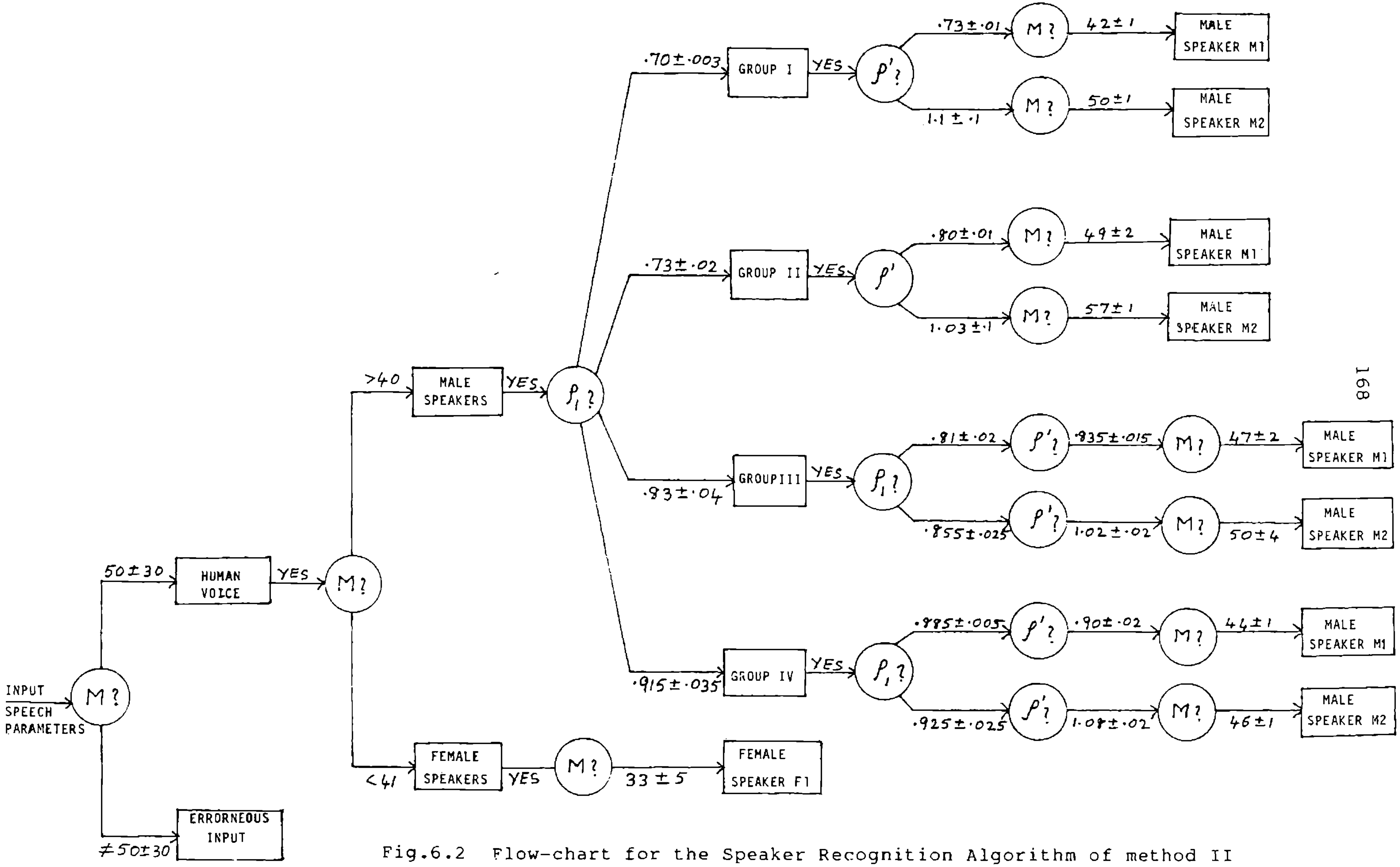


Fig.6.2 Flow-chart for the Speaker Recognition Algorithm of method II

Table 6.4

KNOWLEDGE-BASE FOR THE ALGORITHM OF METHOD II

(M (NO. OF SAMPLES IN 1 PITCH PERIOD OF THE INPUT SIGNAL))
(ρ_1, ρ_M (NORMALISED CORRELATION COEFFICIENTS FOR LAGS 1 & M))
(ρ', (RATIO OF ρ_1 TO ρ_M))
(M1, M2 (MALE SPEAKERS 1 AND 2))
(F1 (FEMALE SPEAKER 1))

(n=2)

10 IF M $\neq$ 50 ± 30
THEN ERROR IN INPUT

12 IF M = 50 ± 30
THEN HUMAN VOICE

14 IF HUMAN VOICE = YES AND M $<$ 41
THEN FEMALE SPEAKERS

16 IF FEMALE SPEAKERS = YES AND M = 33 ± 5
THEN FEMALE SPEAKER F1

18 IF HUMAN VOICE = YES AND M $>$ 40
THEN MALE SPEAKERS

20 IF MALE SPEAKERS = YES AND ρ_1 = .70 ± .003
THEN GROUP I

22 IF MALE SPEAKERS = YES AND ρ_1 = .73 ± .02
THEN GROUP II

24 IF MALE SPEAKERS = YES AND ρ_1 = .83 ± .04
THEN GROUP III

```
26    IF MALE SPEAKERS = YES AND $\rho_1$ = .915 ± .035
      THEN GROUP IV

28    IF GROUP I = YES AND $\rho'$ = .73 ± .01 AND M = 42 ± 1
      THEN MALE SPEAKER M1

30    IF GROUP I = YES AND $\rho'$ = 1.1 ± .1 and M = 50 ± 1
      THEN MALE SPEAKER M2

32    IF GROUP I = YES BUT MALE SPEAKER $\neq$ M1 TO Mn
      THEN MALE SPEAKER $M_{n+1}$

34    IF MALE SPEAKER $M_{n+1}$
      THEN ADD Mn+1 (M, $\rho_1$, $\rho'$ ) TO THE DATA BASE AND I=I-1

36    IF GROUP II = YES AND $\rho'$ = .80 ± .01 AND M = 49 ± 1
      THEN MALE SPEAKER M1

38    IF GROUP II = YES AND $\rho'$ = 1.03 ± .1 AND M = 57 ± 1
      THEN MALE SPEAKER M2

40    IF GROUP II = YES BUT MALE SPEAKER $\neq$ M1 TO Mn
      THEN MALE SPEAKER $M_{n+1}$

42    IF MALE SPEAKER $M_{n+1}$
      THEN ADD $M_{n+1}$ (M, $\rho_1$ ,$\rho'$) TO THE DATA BASE AND I=I-1

44    IF GROUP III = YES AND $\rho_1$ = .81 ± .02
      AND $\rho'$ = .835 ± .015 AND M = 47 ± 2
      THEN MALE SPEAKER M1
```

46 IF GROUP III = YES AND ρ_1 = .855 ± .025
AND ρ' = 1.02 ± .02 AND M = 50 ± 4
THEN MALE SPEAKER M2

48 IF GROUP III = YES BUT MALE SPEAKER ≠ M1 TO Mn
THEN MALE SPEAKER M_{n+1}

50 IF MALE SPEAKER M_{n+1}
THEN ADD M_{n+1} (M, ρ_1, ρ') TO THE DATA BASE AND n=n÷1

52 IF GROUP IV = YES AND ρ_1 = .885 ± .005
AND ρ' = .90 ± .02 AND M = 44 ± 1
THEN MALE SPEAKER M1

54 IF GROUP IV = YES AND ρ_1 = .925 ± .025
AND ρ' = 1.08 ± .02 AND M = 46 ± 1
THEN MALE SPEAKER M2

56 IF GROUP IV = YES BUT MALE SPEAKER ≠ M1 TO Mn
THEN MALE SPEAKER M_{n+1}

58 IF MALE SPEAKER M_{n+1}
THEN ADD M_{n+1} (M, ρ_1, ρ') TO THE DATA BASE AND n=n÷1 .

side result obtained from the speech recognition system developed.

6.2 Phoneme Identification

Phoneme is the basic unit which describes how speech conveys a linguistic meaning. Or, in different terms, the set of phonemes in any language is the set of units that are required to represent utterances in an unambiguous manner. Roughly speaking, a phoneme is a group of similar, but not identical, sounds that differ from one another in accordance with the context in which each occurs [92].

Based on the mode of vibration of the source of excitation, speech sounds are broadly classified as voiced and unvoiced. Depending on the position of the different articulators in the process of the production of the various sounds, they are again subdivided to form ten groups containing altogether 36 phonemes, excluding the 9 dipthongs (in British English) [73]. Fig.6.3 shows the chart of the various phonemes in English.

The same phoneme, judged by phoneticians to be the 'same' at a phonetic detail level, when spoken by a male and a female speaker vary much in their spectrogram and hence it is difficult for a machine to recognize it, though a human brain is able to recognize it correctly. Hence transformation of the spectrum has to be done using current knowledge gleaned from psychoacoustic experiments and from electrophysiological investigations of the response patterns of the auditory system [92]. Such transformed spectra can offer a superior input to a speaker-independent

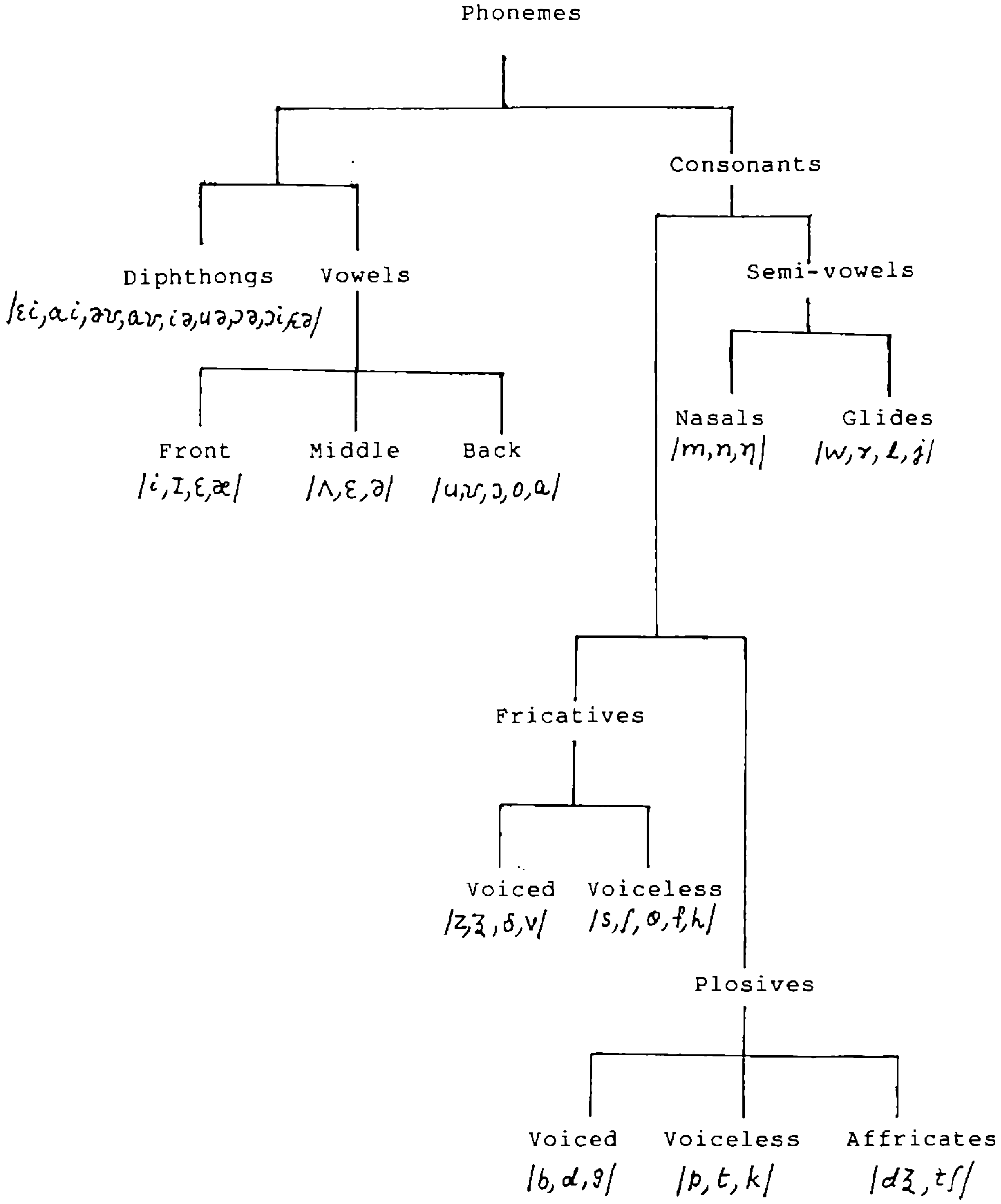


Fig.6.3 The Phonemes of British English [73]

speech system.

Recently, Ljolje and Levison [93] have developed a single ergodic hidden Markov model to represent the acoustic-phonetic structure of English language. The inherent variability of each phoneme is modelled as the observable random process of the Markov chain, while the phonotactive model of the unobservable phonetic sequence is represented by the state transition matrix of the hidden Markov model. The recognition score was high enough, using 43 states.

A speaker-dependent phoneme identification approach is presented below. The salient features of the speech signal, like short time energy (STE) short-time zero crossing rate (STZCR), correlation coefficients, pitch period, etc. have been made use of, for the identification. The method can be represented as an energy-based recognition, as the first parameter considered in the knowledge base is the energy content in a particular phoneme envelope. The fact that voiced phonemes corresponding to vowels and nasals have higher energy than other phonemes forms the basis for the main classification. Based on the average STE content in the phonemes, the 36 phonemes (in British English) are first categorised into three groups. Within each group, a range of values are fixed up for the various parameters and a final identification is done.

6.2.1 Feature Extraction for Creating a Knowledge Base

Three to four words each, were chosen for each of the 36

Table 6.5

Analysis results of all the 36 phonemes in British English

Phonere	STE	STZCR	ZEP	ZER	PP	ρ_1	ρ_2	ρ_1/ρ_2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
/b/	.13-.15	1140-1458	170-187	7670-11,390	4.875	.83-.865	.45-.65	1.3-1.9
	.19-.21	792-1150	151-180	4047-4146	4.625-5.00	.91-.94	.78-.83	1.1-1.3
/d/	.14-.16	1479-1536	210-240	7800-10300	5.0	.75-.80	.34-.80	1.0-1.013
/g/	.15-.157	1020-1220	145-190	6800-7750	4.875	.89-.90	.67-.75	1.02
/p/	.14-.163	1075-1350	150-220	7675-8310	--	.82-.85	.47-.52	1.65-1.90
/t/	.14	1321	180	9645	--	.87	.63	1.16
	.18	979	174	5500	--	.905	.78	1.40
/k/	.17-.183	658-792	116-145	4050-4350		.92-945	.80-.82	1.14-1.15
/w/	.13-.166	946-1016	127-170	6120-7035	4.3-4.625	.91-.915	.71-.72	1.2-1.3
/r/	.11-.125	670-830	71-104	6280-6650	4.75-4.875	.92-.94	.785-.81	1.15-1.2
/l/	.14-.15	1140-1625	160-250	7979-10,690	4.75-4.875	.79-.88	.72-.73	1.1-1.25
/j/	.196	875	171.5	4465	4.75	.913	.81	1.127

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
/h/	.09-.13	1300-1500	122-201	11194-13873	--	.63-.93	.77	1.19
/f/	.19-.20	645-750	125-150	3329-3740	--	.86-.89	.76-.78	1.10-1.2
/θ/	.194	1033	210-210	5574	--	.84	.74	1.127
/s/	.23-.29	580-710	160-163	1960-3080	--	.91-.93	.80-.82	1.14-1.16
/ʃ/	.17-.24	571-1812	94-193	3420-3460	--	.90-.93	.80-.86	1.09-1.13
/v/	.12-.13	859 -1281	100-172	7367-9540	4.875-5.25	.81-.93	.78	1.2-1.33
/ð/	.205-.207	628-930	125-190	2930-4540	4.375-4.875	.94-.95	.81-.84	1.13-1.15
/z/	.14	834	113	5734	4.625	.84	.71	1.377
	.199	518	103	2603	4.625-4.75	.95	.82	1.48
	.234-258	429-562	110-129	1663-2444	4.5-4.625	.955-.965	.89-.90	1.16-1.26
/ʒ/	.200	953	190.6	4765	4.375-4.5	.925	.858	1.082
/dʒ/	.24-.34	518-641	156-177	1514-2630	4.125-4.875	.93-.952	.87-.91	1.05-1.07
	.185	636	127	3700	4.75	.935	.843	1.05
/tʃ/	.24-.325	500	120-163	1540-2075	4.125	.94-.96	.865-.91	1.05-1.09

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
/i/	.28-.29	530-535	150-190	1870-2320	4.125	.87	.825	1.055
/I/	.2-.227	525-646	119-131.14	2312-3185	4.125	.89-895	.80-.83	1.08-1.12
/ɛ/	.18-.22	833-875	155-189	4050-4450	4.625	.91	.73-.75	1.22-1.25
/æ/	.097-.14	1500-2375	204-230	10,980-24,540	5.0	.37-.60	.62	1.3-2.14
/ʒ/	.18-.22	840-857	157-186	3958-4670	4.375-4.75	.93	.77	1.2-1.21
	.24-.25	911-917	224-230	3615-3750	4.125-4.5	.93-.935	.77	1.21
/ʌ/	.11-.115	1438-1650	157-190	13,192-14410	4.375-4.875	.67-.69	--	--
/ə/	.31	708	220.72	2270	4.125	.94	.905	1.112
/a/	.095-.13	1060-1150	132-150	8496-8850	4.625-4.75	.635-.80	.475-.538	1.25-1.5
/o/	.162	857	138.62	5297	4.625	.923	.745	1.238
/ɔ/	.25	1521	380.63	6076.7	4.5	.838	.458	1.06
/ʊ/	.25-.45	479	119-213	1076-1925	4.0	.948-.978	.84-.93	1.06-1.13
/u/	.23-.274	429-458	105-117	1991-2045	4.125-4.375	.958-.97	.855-.90	1.08-1.12
/m/	.26-.46	480-489	189-221	1050-1275	4.875-5.0	.94-.97	.85-.92	1.05-1.09
/n/	.232-.42	446-460	187-195	1060-1955	4.375-4.75	.89-.99	.72-.92	1.05-1.3
/ŋ/	.34-.47	450-464	150-220	987-1340	4.125	.97	.91-.92	1.05-1.06

phonemes. The data corresponding to these words were obtained as in the earlier cases, using a speech digitizer. The region corresponding to the particular phoneme, was manually separated from each word, by plotting on a VDU, and then re-checking it by reconstruction and listening. The data corresponding to each phoneme were stored in different files.

The data were time-normalised and grouped into segments of length 80 samples (since the vocal tract 'rings' for 10 msec and this corresponds to 80 samples at a sampling frequency of 8 kHz). The parameters like STE, STZCR, ZCR-energy product (ZEP), STZCR to STE ratio (ZER), normalised correlation coefficients for lags 1 and 2, ρ_1 and $\rho_{2,}$ predictor coefficients and reflection coefficients were evaluated for each segment and a set of average values were computed for each phoneme, in the different words considered. From these a range of values were obtained for each parameter, for each of the 36 phonemes. At the outset, the parameters like predictor coefficients and reflection coefficients were 'knocked out' or eliminated, as they were not showing much of discriminative capacity among the phonemes. A list of the values of the different parameters useful in the phoneme identification process is given in Table 6.5.

Studying the energy ranges of the different phonemes, three different ranges were fixed and the phonemes falling in a particular range was grouped together. If a particular phoneme has STE values falling in two different ranges, they are included in both the ranges and are finally identified based on the parameters chosen in that particular group.

The energy limits obtained in the study were 0.09 to 0.47. Three different ranges - 0.09 to 0.17, .17 to .23, .23 to .47 - were fixed within the extreme limits. The phonemes that come under the topmost energy group (group I), that is, .23 $\leqslant$ STE $\leqslant$ .47 are the vowels, nasals, and some of the fricatives and affricates. They are /n/, /m/, /ŋ/, /i/, /ʒ/, /ə/, /ɔ/, /ɚ/, /u/, /S/, /ʃ/, /tʃ/, /dʒ/ and /z/. Similarly, the second energy group (group II), .17 $\leqslant$ STE $<$ .23, contains the higher energy plosives, glides, low energy vowels and a few of the fricatives. They are /b/, /t/, /k/, /j/, /f/, /θ/, /ð/, /z/, /ʒ/, /dʒ/, /l/, /ɛ/ and /ɜ/. These phonemes are comparatively low energy phonemes with lower correlation than the high energy phonemes of group I. (/l/ and /i/ have been proved to show low correlation among the samples themselves and have proved to produce low SNR value of the order of 5 dB, when used in the modified coder developed earlier). The third energy group (group III), .09 $\leqslant$ STE $<$ .17, contains almost all of the plosives, glides and few of the very low energy vowels.They are /b/, /d/, /g/, /p/, /t/, /w/, /l/, /r/, /h/, /v/, /æ/, /ʌ/, /ɑ/, /o/. It can be noted that the phoneme /z/ overlaps all the three energy groups, while /dz/ and /ʒ/ overlap groups I and II and /b/ and /t/ overlaps energy regions of group II and III. Within each group, various sub-groups are formed based on the zero crossing rate and a final identification algorithm is developed using the different ranges of values of ZEP, ZER, ρ_1, and the ratio ρ_1/ρ_2. The detailed algorithm developed is shown in Table 6.6.

If more and more words are considered for each phoneme, then

Table 6.6

KNOWLEDGE-BASE FOR THE PHONEME IDENTIFICATION METHOD

(STE, STZCR (SHORT TIME ENERGY AND SHORT TIME ZEROCROSSING RATE OF THE INPUT SIGNAL))
(ZEP, ZER (ZEROCROSSING RATE. ENERGY PRODUCT, AND ZEROCROSSING RATE BY ENERGY))
(K_1 (FIRST REFLECTION COEFFICIENT))
(PP (PITCH PERIOD OF THE INPUT SIGNAL))
(ρ_1, ρ_2 (NORMALISED CORRELATION COEFFICIENTS FOR LAGS 1 AND 2)
(ρ' (RATIO OF ρ_1 to ρ_2)

1 IF .23 ≤ STE ≤ .47
THEN GROUP I (ʒ, ɔ, ʃ, ə, ʂ, tʃ, m, n, ŋ, ʋ, u, dʒ, z, i)

2 IF GROUP I = YES AND STZCR = 1515 ± 10 OR ZER = 6000 ± 100
OR ZEP = 370 ± 10
THEN PHONEME / ɔ /

3 IF GROUP I = YES AND STZCR = 910 ± 100
OR ZEP = 225 ± 5
THEN PHONEME / ʒ /

4 IF GROUP I = YES AND STZCR = 810 ± 10
OR ZER = 3400 ± 100 OR ZEP = 195 ± 10
THEN PHONEME / ʃ /

5 IF GROUP I = YES AND STZCR = 710 ± 10
OR ZER = 2200 ± 100 OR ZEP = 215 ± 5
THEN PHONEME / ə /

6 IF GROUP I = YES AND STZCR = 630 ± 70
AND ρ_2 = .81 ± .01
THEN PHONEME / s /

7 IF GROUP I = YES AND STZCR = 580 ± 70
AND ρ_2 = .84 ± .01
THEN PHONEME / dʒ /

8 IF GROUP I = YES AND STZCR = 530 ± 5
AND ρ_1 = .87 ± .01
THEN PHONEME /i/

9 IF GROUP I = YES AND STZCR = 500 ± 5
AND ρ_1 = 0.93 ± .01
THEN PHONEME /tʃ/

10 IF GROUP I = YES and STZCR = 480 ± 5
AND PP = 4.875 ± .125 AND ZEP = 200 ± 20
THEN PHONEME / m /

11 IF GROUP I = YES and STZCR = 480 ± 2
AND PP = 4.0 ± .125 AND ZEP = 165 ± 50
THEN PHONEME / ʊ /

12 IF GROUP I = YES and STZCR = 450 ± 10
AND PP = 4.25 ± .125
AND K_1 = -.96 ±·05
THEN PHONEME / u /

13 IF GROUP I = YES and STZCR = 490 ± 70
AND PP = 4.5 ± .125 AND K_1 = -.945 ± .05
THEN PHONEME / z /

14 IF GROUP I = YES AND STZCR = 455 ± 10
AND PP = 4.5 ±.25
THEN PHONEME /η /

15 IF GROUP I = YES AND STZCR = 455 ± 10
AND PP = 4.125 ± 0
THEN PHONEME / η /

16 IF .17 ⩽ STE < .23
THEN GROUP II (t , θ , ʓ , j , ʒ , δ , k , f , ɪ , ɛ , z , dʓ)

17 IF GROUP II = YES AND STZCR = 1310 ± 20
OR ZER = 9650 ± 10 AND ρ_1 = .905 ± .05
THEN PHONEME /t/

18 IF GROUP II = YES AND STZCR = 1070 ± 20
OR ZER = 5550 ± 100 AND ρ_1 = .84 ± .05
THEN PHONEME / θ /

19 IF GROUP II = YES AND STZCR = 950 ± 20
OR ZER = 4750 ± 50 AND ρ_1 = .925 ± .05
THEN PHONEME / ʓ /

20 IF GROUP II = YES AND STZCR = 875 ± 20
AND ZER = 4460 ± 50 AND ρ_1 = .915 ± .05
THEN PHONEME /j/

21 IF GROUP II = YES AND STZCR = 850 ± 20
AND ZER = 4300 ± 350 AND ρ_1 = .930 ± .05
THEN PHONEME /ʒ /

22 IF GROUP II = YES AND STZCR = 815 ± 25
AND ZER = 4100 ± 50 AND ρ_1 = .925 ± .15
THEN PHONEME /b/

23 IF GROUP II = YES AND STZCR = 770 ± 160
AND ZER = 3750 ± 750 AND ρ_1 = .945 ± .05
THEN PHONEME / ð /

24 IF GROUP II = YES AND ZCR = 740 ± 60
AND ZER = 4150 ± 100 AND ρ_1 = .93 ± .15
THEN PHONEME /k/

25 IF GROUP II = YES AND STZCR = 700 ± 60
AND ZER = 3540 ± 200 AND ρ_1 = .875 ±.15
THEN PHONEME / f /

26 IF GROUP II = YES AND STZCR = 680 ± 20
AND ZER = 3700 ± 50 AND ρ_1 = .935 ± .05
THEN PHONEME / dʒ /

27 IF GROUP II = YES AND STZCR = 585 ± 60
AND ZER = 2750 ± 450 AND ρ_1 = .89 ± .05
THEN PHONEME /ɪ/

28 IF GROUP II = YES AND STZCR = 530 ± 20
AND ZER = 1870 ± 20 AND ρ_1 = .87 ± .05
THEN PHONEME / ɛ /

29 IF GROUP II = YES AND STZCR = 510 ± 10
AND ZER = 2600 ± 20 AND ρ_1 = .925 ± .10
THEN PHONEME /z /

30 IF .09 $\leq$ STE $<$ .17
THEN GROUP III (t, w, r, v, z, o, b, d, g, p, l, h, æ , ʌ , ɑ)

31 IF GROUP III = YES AND STZCR = 1940 ± 540 AND ZEP = 220 ± 15
OR ZER $>$ 18000 AND ρ' = 1.85 ±.45
THEN PHONEME /æ/

32 IF GROUP III = YES AND STZCR = 1550 ± 100 AND ZEP = 175 ± 15
OR ZER = 13800 ± 600, AND ρ' = 2.75 ± .25
THEN PHONEME / ʌ /

33 IF GROUP III = YES AND STZCR = 1510 ± 30 AND ZEP = 225 ± 15
AND ZER = 10050 ± 2500 AND ρ' = 1.01 ± .01
THEN PHONEME / *d* /

34 IF GROUP III = YES AND STZCR = 1400 ± 100 AND ZEP = 160 ± 40
AND ZER = 12500 ± 1400 AND ρ' = 1.8 .1
THEN PHONEME /h/

35 IF GROUP III = YES AND STZCR = 1400 ± 250 and ZEP = 205 ± 45
AND ZER = 9300 ± 1300 AND ρ' = 1.15 ± .075
THEN PHONEME / /

36 IF GROUP III = YES AND STZCR = 1260 ± 200 AND ZEP = 140 ± 10
AND ZER = 11500 ± 3500 AND ρ' = 1.45 ± .02
THEN PHONEME / ɑ /

37 IF GROUP III = YES AND STZCR = 1300 ± 160 AND ZEP = 180 ± 10
AND ZER = 9500 ± 2000 AND ρ' = 1.6 ± .3
THEN PHONEME / b /

38 IF GROUP III = YES AND STZCR = 1200 ± 150 AND ZEP = 185 ± 35
AND ZER = 8000 ± 330 AND ρ' = 1.8 ± .15
THEN PHONEME /p/

39 IF GROUP III = YES AND STZCR = 1120 ± 100 AND ZEP = 170 ± 25
AND ZER = 7300 ± 500, AND ρ' = 1.02 ± .01
THEN PHONEME /ɡ/

40 IF GROUP III = YES AND STZCR = 1080 ± 220 AND ZEP = 135 ± 35
AND ZER = 8500 ± 1050 AND ρ' = 1.95 ± .25
THEN PHONEME /v/

41 IF GROUP III = YES AND STZCR = 980 ± 35 AND ZEP = 130 ± 44
AND ZER = 6570 ± 450 AND ρ' = 1.45 ± .15
THEN PHONEME /ɯ/

42 IF GROUP III = YES AND STZCR = 980 ± 10 AND ZEP = 175 ± 5
AND ZER = 5500 ± 100 AND ρ' = 1.16 ± .05
THEN PHONEME /t/

43 IF GROUP III = YES AND STZCR = 855 ± 10 AND ZEP = 140 ± 5
AND ZER = 5300 ± 100 AND ρ' = 1.2 ± .05
THEN PHONEME /o/

44 IF GROUP III = YES AND STZCR = 805 ± 10 AND ZEP = 115 ± 5
AND ZER = 5750 ± 100 AND ρ' = 1.18 ± .05
THEN PHONEME /z/

45 IF GROUP III = YES AND STZCR = 750 ± 80 AND ZEP = 90 ± 20
AND ZER = 6450 ± 200 AND ρ' = 1.175 ± .025
THEN PHONEME /ɣ/

the limiting ranges of the different parameters might change. Then, the number of groups into which the phonemes are initially divided, based on STE, can be increased, and the same procedure can be applied effectively.

Increasing the size of the vocabulary requires more computation, storage and time-consuming training sessions. The response time will also increase linearly with the size of the vocabularies and the error rate tends to become larger.

To summarise, a simple and efficient knowledge-based approach to a speaker recognition system and a phoneme identification system have been presented.

Chapter 7

THE CONCLUSIONS

The underlying thread which runs throughout this thesis is the continuing and the increasing need for effective communication systems in the field of speech processing. A low bit rate coder which is simple, but maintains a good level of speech quality, is a must in the present-day man-machine communication, on a global basis.

This thesis presents a modified block adaptive predictive coder (MBAC) which reduces the computational burden and complexity of the coder by introducing certain changes in the evaluation and transmission techniques of the predictor parameters. The difference between the actual and predicted values of the speech samples are not transmitted. Only the predictor parameters and a difference signal which is a coded version of the first few sample values, is transmitted to the decoder. This makes the system more real time. At the decoder, the original samples are reconstructed based on the earlier predicted values. The predictor at the receiver, is updated by transmitting the predictor parameters afresh every block. The block length for processing is made relative to the signal being processed (that is, block length N is taken as equal to 4M, where M is the number of samples within 1 pitch period of the input signal) rather than choosing it to be a constant frame length. The aim was to achieve a reduced band width requirement and an SNR around 10 dB. This was achieved as shown in the results section.

At the outset of speech data processing, using the coder, one has to detect whether the block of data under consideration is voiced/unvoiced/silent/transition. For this, three classification algorithms have been developed. Of these, the third method, which gives the minimum error probability of 2.6% was used to design the modified coder.

The results of the comparative study between the autocorrelation method and covariance method, in the evaluation of the predictor parameters, revealed that the covariance method tends to make the filter unstable. Also the SNRSEG value obtained was lesser than those obtained by using autocorrelation method, though only less by 0.2 to 0.5 dB. Another point is that the addition of a pitch predictor to the spectrum predictor showed a marked improvement in the quality of the reconstructed speech. The SNRSEG value increased by 8 to 9 dB. The unvoiced segments gave a maximum SNRSEG of 5.11 dB using spectrum prediction.

Computer simulation of the MBAC system showed that for different sounds and different speakers, an SNRSEG value of 7 to 18 dB can be obtained in the voiced/transition region, while it is only 2 to 5 dB in the unvoiced region. On the whole, the system gain obtainable is 8 to 12 dB. The applicability of the MBAC on certain special sounds in Malayalam was also studied. The coder performance on these special sounds was very good. The sound /ṇ/ ENN in 'thoon', with the nasal /n/ following a vowel gives the maximum average SNRSEG of 21.96 dB, with the segment maximum going up to 30 dB.

Compared to the SEV and CELPC coders [8, 9, 10, 64, 67] which produce good quality speech with an average SNRSEG ranging from 8.5 dB to 13.8 dB, the performance of the MBAC system is good. Computationally, the MBAC system is very simple. Whereas an SEV requires 40,000 multiples/adds per frame (or 4 MFLOPs) and a CELPC requires 800,000 multiples/adds per frame (or 80 MFLOPs) [8], the MBAC requires just an order of (p^2+p) multiples and adds, where p is the order of the predictor. The bit rate achieved is 11.8835 kb/sec. An adaptive knowledge-based speaker recognition system has been developed based on the various parameter values obtained during the processing of the MBAC, on voiced speech samples. A proposal for a phoneme identification system has also been outlined.

Using the results of the study on speech sounds (English) the methods developed has helped in studying the behaviour of these parameters for Malayalam sounds. An expert system for this analysis will be a worthwhile study.

Further improvement in the performance of the MBAC system can be realized by increasing the number of bits per sample used for encoding the first few sample values that are to be transmitted to the decoder. But this increases the transmission rate and hence a compromise has to be done between the two. Hardware implementation of the coder can be done to prove its performance and hence its credibility for use in communication systems. The performance of the system developed, can be tried on other signals like music, sounds of children and animals. The feasibility of an implementation of the system using Artificial Neural Network would be a desirable task.

APPENDIX I

PREDICTOR PARAMETER ESTIMATION

(i) Linear Predictive Coder--Using Spectrum Prediction

Gain Computation

In an all-pole model, the nth sample S_n is represented by [2,3,11,14]:

$$S_n = \sum_{k=1}^{p} a_k S_{n-k} + GU_n \tag{1}$$

where,

U_n is the input or excitation,

G is the gain factor,

a_k's are the filter coefficients, and

p is the order of the filter.

If the input U_n is totally unknown, then the estimated value of S_n is:

$$\hat{S}_n = \sum_{k=1}^{p} a_k S_{n-k} \tag{2}$$

The error in prediction is

$$E_n = S_n - \hat{S}_n = S_n - \sum_{k=1}^{p} a_k S_{n-k}$$

or,

$$S_n = \sum_{k=1}^{p} a_k S_{n-k} + E_n \qquad (3)$$

Comparing equations (1) and (3), it can be seen that the only input signal U_n, that will result in the same signal S_n as the output, is where $GU_n = E_n$. Since the filter is fixed, the total energy in the input signal GU_n must be equal to the total energy E_p, in the error signal. Thus, it can be shown that [11], the total energy in the input is given as

$$G^2 = E_p = R(0) + \sum_{k=1}^{p} a_k R(k) \qquad (4)$$

From the above equation, gain G can be evaluated.

(ii) Linear Predictive Coder -- Using Pitch Prediction - Parameter Estimation

The predictor parameters are determined by minimizing the mean square error between the true and the predicted values of the speech samples.

The predicted value of the nth sample is [2, 44]:

$$\hat{S}_n = \beta S_{n-M} + \sum_{k=1}^{p} a_k(S_{n-k} - \beta S_{n-k-M}) \qquad (5)$$

Hence the prediction error for the nth sample is

$$E_n = S_n - \hat{S}_n$$

$$E_n = (S_n - \beta S_{n-M}) - \sum_{k=1}^{p} a_k (S_{n-k} - \beta S_{n-k-M}) \qquad (6)$$

The mean-square error in prediction is given by

$$\langle E_n^2 \rangle_{av} = \frac{1}{N} \sum_n E_n^2, \qquad (7)$$

where, the summation extends over all the samples in the interval, during which the predictor is to be optimum.

The total error minimisation is done in two steps. The parameters β and M are first determined to minimise the error,

$$E_1 = \frac{1}{N} \sum_n (S_n - \beta S_{n-M})^2$$

$$= \langle (S_n - \beta S_{n-M})^2 \rangle_{av} \qquad (8)$$

Thus, setting $\frac{\partial E_1}{\partial \beta} = 0$, we obtain,

$$\beta = \langle s_n s_{n-M} \rangle_{av} \Big/ \langle s_{n-M}^2 \rangle_{av} \qquad (9)$$

Substituting the above value of β in equation (8) it can be seen that the optimum value of M can be determined by locating the position of the maximum of the normalized correlation coefficient ρ, given by

$$\rho\,(M) = \langle s_n s_{n-M} \rangle_{av} \Big/ \left\{ \langle s_n^2 \rangle_{av} \cdot \langle s_{n-M}^2 \rangle_{av} \right\}^{1/2}, \quad M \quad 0 \qquad (10)$$

Next, using these optimum values of M and β, the total error $\langle E_n^2 \rangle_{av}$ is minimised, with respect to each of the coefficients $a_1, a_2, \ldots\ldots a_k$, to obtain the optimum predictor parameters.

Let, $\quad v_n = s_n - \beta\, s_{n-M} \qquad (11)$

Then,

$$\langle E_n^2 \rangle_{av} = \langle (v_n - \sum_{k=1}^{p} a_k v_{n-k})^2 \rangle_{av} \qquad (12)$$

Setting $\dfrac{\partial E_n^2}{\partial aj} = 0$, for $j=1,2,\ldots,p$, $\qquad (13)$

we get a set of p equations in p unknowns, which can be written in matrix notation as:

$$\emptyset a = \psi \qquad (14)$$

where,

$\emptyset$ is a p by p matrix, with its ijth term given by

$$\varphi_{ij} = \langle V_{n-i} V_{n-j} \rangle_{av} \qquad (15)$$

ψ is a p-dimensional vector with its jth component given as

$$\psi_{j} = \langle V_{n} V_{n-j} \rangle_{av} \qquad (16)$$

'a' is the p-dimensional vector which directly gives the values of the optimum predictor coefficients.

Solving for 'a' in equation (14), the optimum predictor coefficients a_1, a_2, ...a_p are obtained.

APPENDIX II

COMPUTATIONAL SAVINGS IN THE PARAMETER ESTIMATION

(i) Pitch Period Determination

The normalised correlation coefficient of the data $\{s_n\}$ is given as

$$\rho(J) = \langle s_n s_{n-J} \rangle_{av} \Big/ \left\{ \langle s_n^2 \rangle_{av} \cdot \langle s_{n-J}^2 \rangle_{av} \right\}^{1/2} \qquad (17)$$

The autocorrelation function of the clipped samples $\{s'_n\}$ is

$$R_c(J) = \langle s'_n s'_{n-J} \rangle_{av} \qquad (18)$$

The autocorrelation function of the data samples $\{s_n\}$ is expressed as

$$R(J) = \langle s_n s_{n-J} \rangle_{av} \qquad (19)$$

In all the above cases, $\langle \; \rangle_{av}$ denotes the averaging over all the 'N' samples in the block under consideration. Also, the number of samples, M, in one pitch period of the signal (which gives a measure of the pitch period) is determined by locating the position of the maximum of the correlation coefficients.

For each value of J, computation of ρ (J) requires, (3N+1) multiplications, (3N-3) additions, 1 square-rooting and 1 division, while R(J) requires only N multiplications (N-1) additions.

In the present work, the correlation coefficient chosen for the determination of M is R(J). Hence, comparing between equations (17) and (19), the computational savings in going from ρ (J) to R(J), is (2N+1) multiplications, (2N-2) additions, 1 square-rooting and 1 division per coefficient or lag 'J'. This gives a good amount of computational savings on the whole.

In the centre-clipping method, an additional amount of computation, on the order of N additions or subtractions, is needed at the beginning of each block.

(ii) Covariance/Autocorrelation Matrix Evaluation

The covariance matrix is a symmetric matrix, while the autocorrelation matrix is a symmetric Toeplitz matrix.

Hence the elements along each of the diagonals, are the same in an autocorrelation matrix, while they are not, in a covariance matrix.

The autocorrelation matrix is expressed as

$$\phi = \begin{bmatrix} R_0 & R_1 \cdots\cdots\cdots R_{p-1} \\ R_1 & R_0 \cdots\cdots\cdots R_{p-2} \\ \cdots\cdots & \cdots\cdots\cdots\cdots \\ R_{p-1} & R_{p-2} \cdots\cdots\cdots R_0 \end{bmatrix} \qquad (20)$$

and the covariance matrix is given by

$$\phi = \begin{bmatrix} \phi_{11} & \phi_{12} \cdots\cdots\cdots \phi_{1p} \\ \phi_{12} & \phi_{22} \cdots\cdots\cdots \phi_{2p} \\ \cdots\cdots & \cdots\cdots\cdots\cdots \\ \phi_{p1} & \phi_{p2} \cdots\cdots\cdots \phi_{pp} \end{bmatrix} \qquad (21)$$

where, $\phi_{ij} = \phi_{ij}$. p is the order of the filter considered.

Hence the number of matrix elements to be computed is, p in autocorrelation matrix and p(p+1)/2 in covariance matrix. If N is the number of samples in the interval considered, then each matrix element needs nearly N multiplications and N additions. Or, the autocorrelation matrix elements need, on the whole, pN multiplications and additions while covariance matrix elements need $\frac{p(p+1)}{2}$ N operations.

Taking into consideration the relationship that exists between elements along each diagonal, it is noted that each element is obtained from its preceding element, by adding and subtracting one product term each, as given by

$$\emptyset_{i+1,j+1} = \emptyset_{i,j} + V_{-i}V_{-j} - V_{N-i}V_{N-j} \qquad (22)$$

where

$$\emptyset_{i,j} = \sum_{n=1}^{N} V_{n-i}V_{n-j} \qquad (23)$$

Hence, from the p elements ($\emptyset_{11}$ to $\emptyset_{pp}$) in the first row, the rest of the $(p^2 - p)/2$ elements (that is, $\emptyset_{23}$ to $\emptyset_{p-1,p}$, $\emptyset_{24}$ to $\emptyset_{p-2,p}$, ..., $\emptyset_{2,p-1}$ to $\emptyset_{3,p}$ and $\emptyset_{2,p}$) can be evaluated using equation (22).

The computations needed for the elements $\emptyset_{11}$ to $\emptyset_{pp}$ is

$= pN$ multiplications and $p(N-1)$ additions

$\approx pN$ multiplications and additions.

The computations required for the other $(p^2 - p)/2$ elements

$= (p^2 - p)$ multiplications and additions.

Therefore the total number of operations required

$= (pN + p^2 - p)$ multiplications and additions

$\approx (pN + p^2)$ multiples and adds.

In this case, the reduction in computational load is appreciable only if N is low.

(iii) Evaluation of the Predictor Coefficients

In the Gauss-Jordan elimination method, to solve for the predictor coefficient matrix **a** from the matrix equation $\emptyset a = \psi$, the matrix $\emptyset\psi$ is first reduced to an upper triangular matrix and then the coefficients 'a' are evaluated. The computations required for the formation of the triangular matrix are:

$(p+1) + p + (p-1) + \ldots.+3+2$ divisions

$= 2p + \frac{p(p-1)}{2} = [(p^2/2) + (3/2)p]$ divisions

and $(p+1) + p + (p-1) + \ldots.. + 3$ multiples and subtracts

$= 3(p-1) + [(p-1)(p-2)/2]$

$= [(p^2/2) + (3/2)p-4]$ multiples and subtracts.

Hence, on the whole, it requires on the order of $[p^2+O(p)]$ multiples and adds.

Makhoul [11] has reported that the square-root or Cholesky decomposition method requires $p^3/6 + O(p^2)$ operations, while Durbin's method requires $p^2+O(p)$ operations. The Gauss-Jordan elimination method is thus as effective as the Durbin's method, in a computational sense.

APPENDIX III

RATE OF TRANSMISSION

(a) Predictor Parameter Transmission

(i) Voiced region

The parameters to be transmitted are β , M, voiced/unvoiced/silent/transition parameter and a_k's.

According to Atal et al [88], using predictor polynomial roots, 5 bits per root are adequate to preserve the quality of the synthesised speech. Hence, the total number of bits needed per frame is

a_k's (4 in number)	20 bits
V/U/S/T parameter	2 bits
β	5 bits
M	6 bits
Total bits per frame	33 bits

Taking an average value of M as 50 for male speakers and 35 for female speakers, the average value of M on the whole will be around 45, so that the block length $N = 4M$ is equal to 180. Since the sampling frequency $f_s = 8$ KHz, the number of blocks per second is 44 and the number of bits per second is 1.452 kb/s.

(ii) Transition region

Here, the optimum value of $N = 2M$ and $p = 8$. Hence the number of bits per frame is 53 (a_k's → 40 bits, M → 6 bits, β → 5 bits, V/U/S/T parameter → 2 bits), and the number of blocks per second is 89 and the number of bits required per second is 4.717 kb/s.

(iii) Unvoiced region

The optimum values of $N = 40$ and $p = 12$ and hence the number of bits per frame is 62 and the number of bits per second is 12.4 kb/s.

(iv) Silent region

The silent region needs no processing and only the code for the region need be transmitted to the receiver. Thus taking the frame length same as the initial block length of 160, the number of blocks per second is 50 and the number of bits per second is 100.

From Table 4.7, it can be noted that on an average, in the phonetically balanced sentences chosen for the simulation work, the percentage of occurrences of the various regions are approximately, 40% for voiced, 10% for unvoiced, 20% for silent and 30% for transition regions. Hence, the total number of bits required to transmit the predictor parameters is

Voiced region → 1.452x.4 = 0.5808 kb/s
Unvoiced region → 12.40x.1 = 1.240 kb/s
Silent region → 0.10x.2 = 0.002 kb/s
Transition region → 4.717x.3 = 1.4151 kb/s

Total = 3.2379 kb/s .

(b) Side Informations

(i) Voiced region

N = 4M and p = 4. As explained above, taking M as equal to 45, the number of d_n's required to be transmitted per block is 49. Taking 3 bits per difference sample d_n and taking 12 bits for the first sample S_1 of the block and 5 bits for the standard deviation, the total number of bits per frame

= 49 x 3 + 12 + 5 = 164 bits/frame.

Taking 44 voiced frames per second, number of bits per second with respect to voiced frames

= 164 x 44 = 7.216 kb/s

(ii) Transition region

Taking M = 45, N = 2M and p = 8, the total number of bits per frame is 176 (ie., 53 x 3 + 12 + 5), and the number of bits per second = 176 x 89 = 15.664 kb/s.

(iii) Unvoiced region

Here N = 40 and p = 12, and the number of bits per frame is 53 (ie., 12 x 3 + 12 + 5). Hence the number of bits per second = 53 x 200 = 10.60 kb/s.

(iv) Silent region

No side information is needed.

Thus the total number of bits required per second to transmit the side informations is given as

Voiced region	7.216 x .4	= 2.8864 kb/s
Unvoiced region	10.60 x .1	= 1.060 kb/s
Transition region	15.664 x .3	= 4.6992 kb/s
Total		= 8.6456 kb/s

The overall number of bits per second, required for the transmission of both the predictor parameters and the side informations is 11.8835 kb/s (that is,(3.2379 + 8.6456) kb/s).

APPENDIX IV

SPEECH WAVEFORMS

Speech Waveforms of the utterances used in the present work
(a) - male, (b) - female

Fig.1 The pipe began to rust while new.

Fig.2 Cats and dogs hate each the other.

Fig.3 Oak is strong and also gives shade.

Fig.4 Thieves who rob friends deserve jail.

Fig.5 Open the crate but do not break the glass.

Fig.6 Add the sum to the product of these three.

Fig.7 Joe brought a young girl.

Fig.8 Drop coin after tone.

Fig.9 Push blue after speech.

Fig.10 Close door after party.

Fig.11 Right move close lock.

Fig.12(a-j)Waveforms of the different words used for phoneme identification.

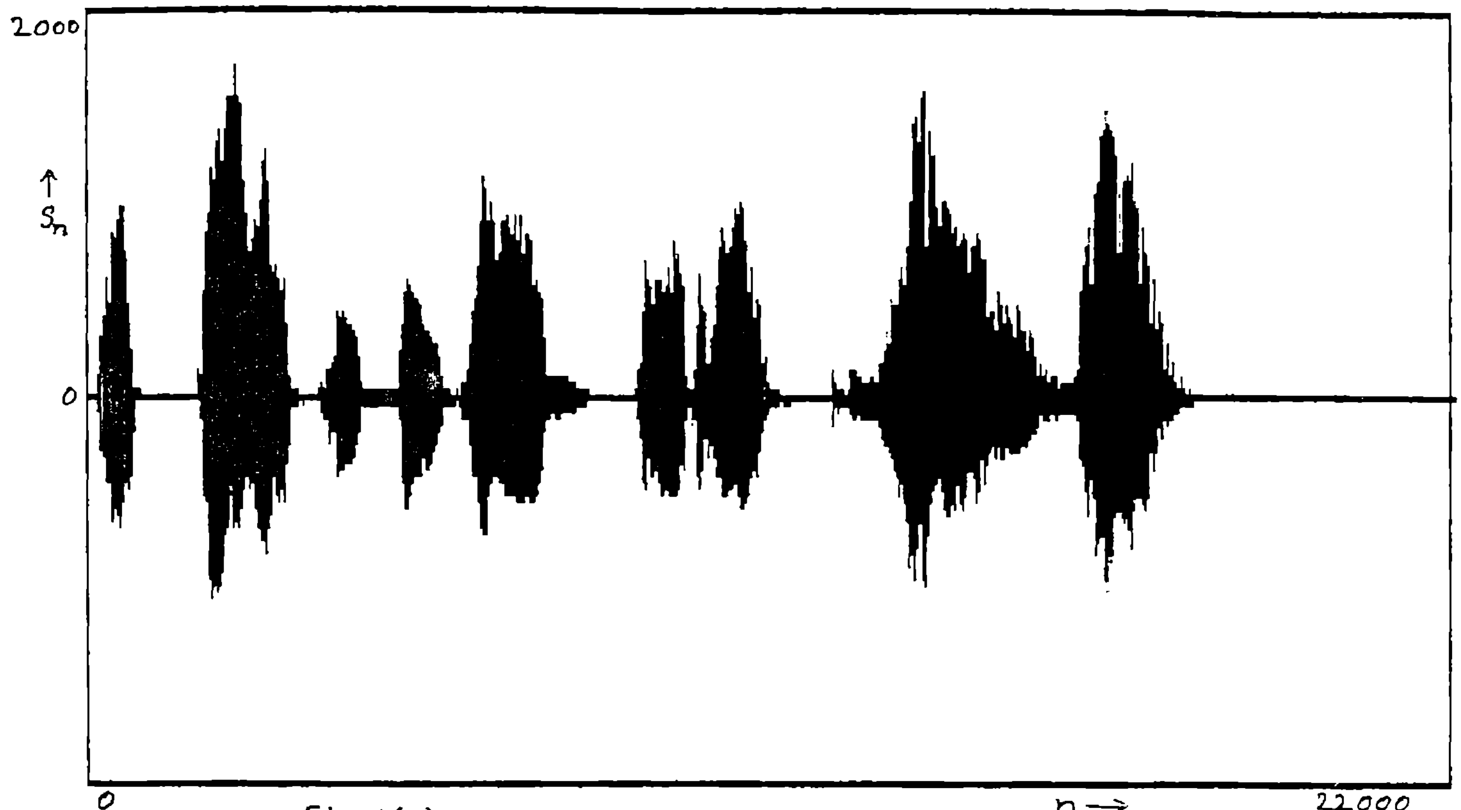


Fig 1(a)

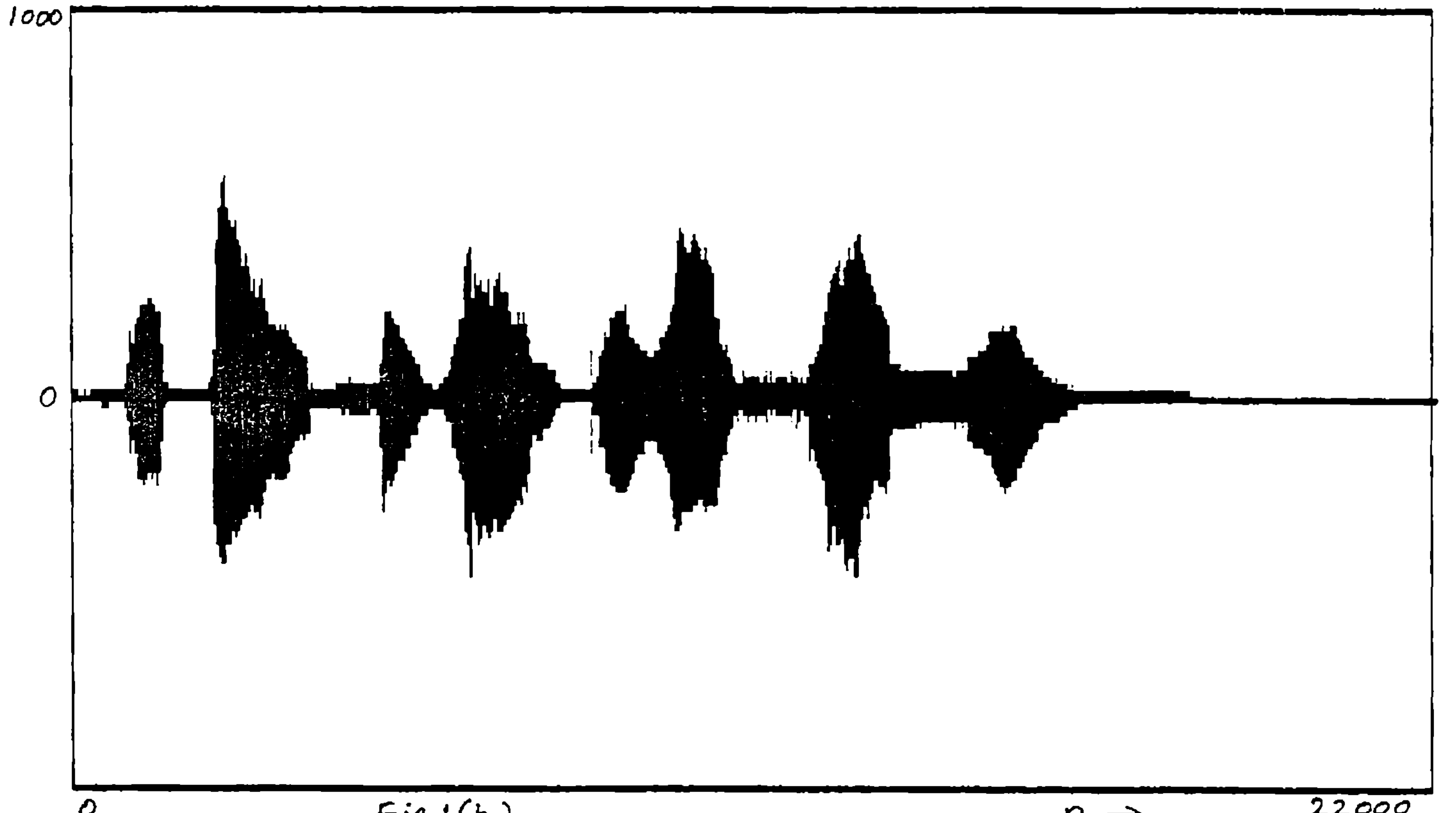


Fig 1(b)

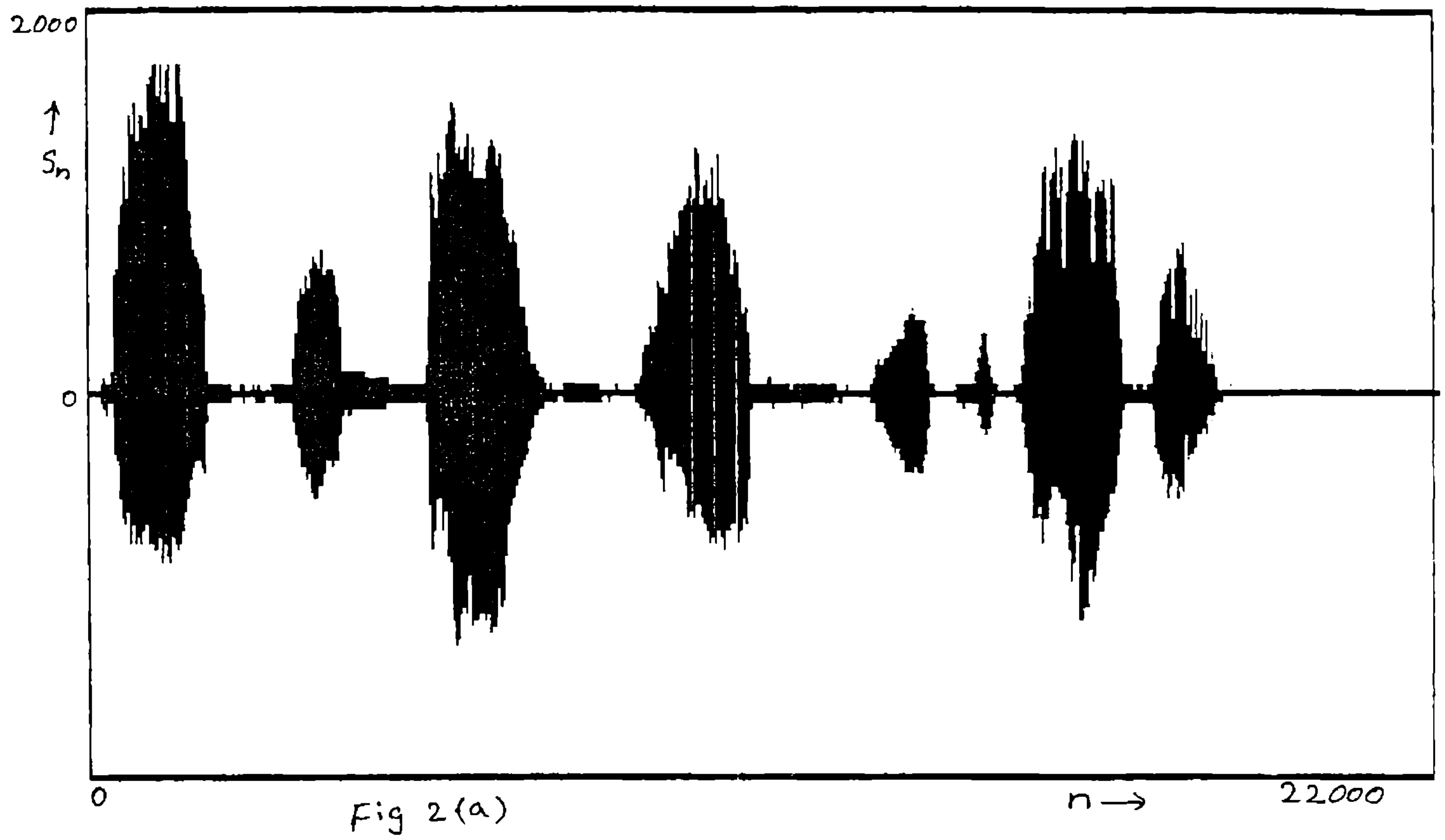


Fig 2(a)

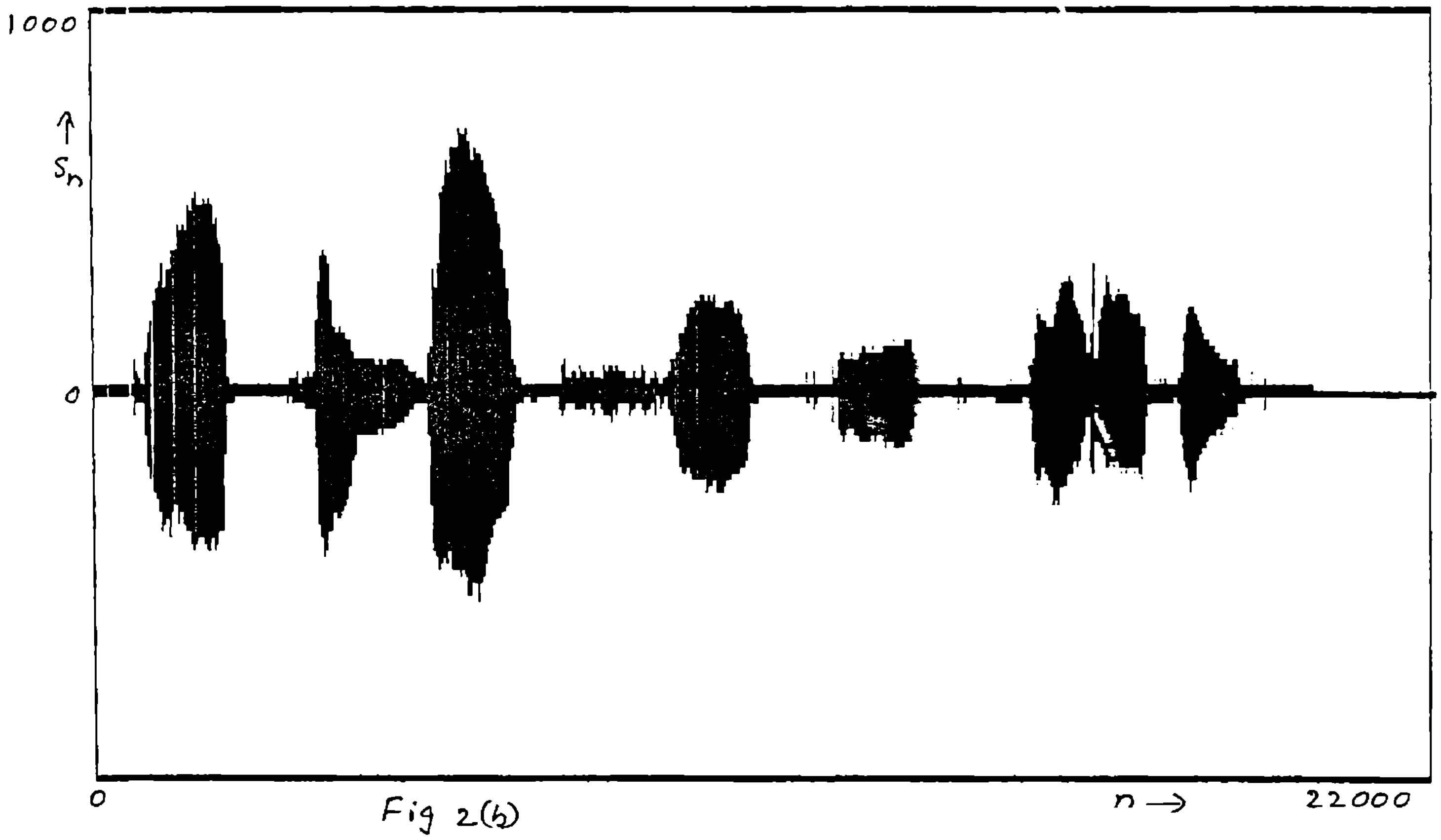


Fig 2(b)

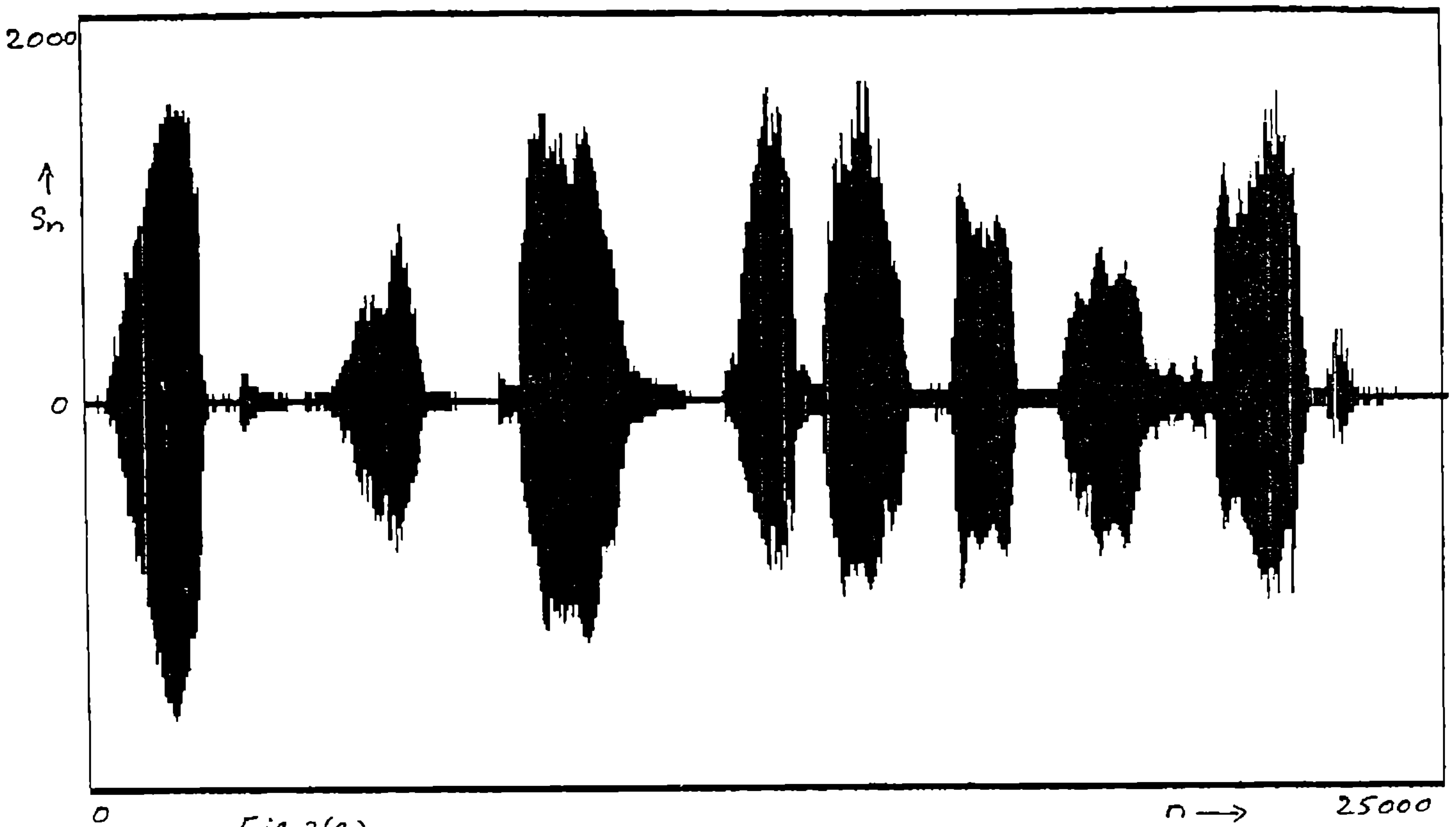


Fig 3(a)

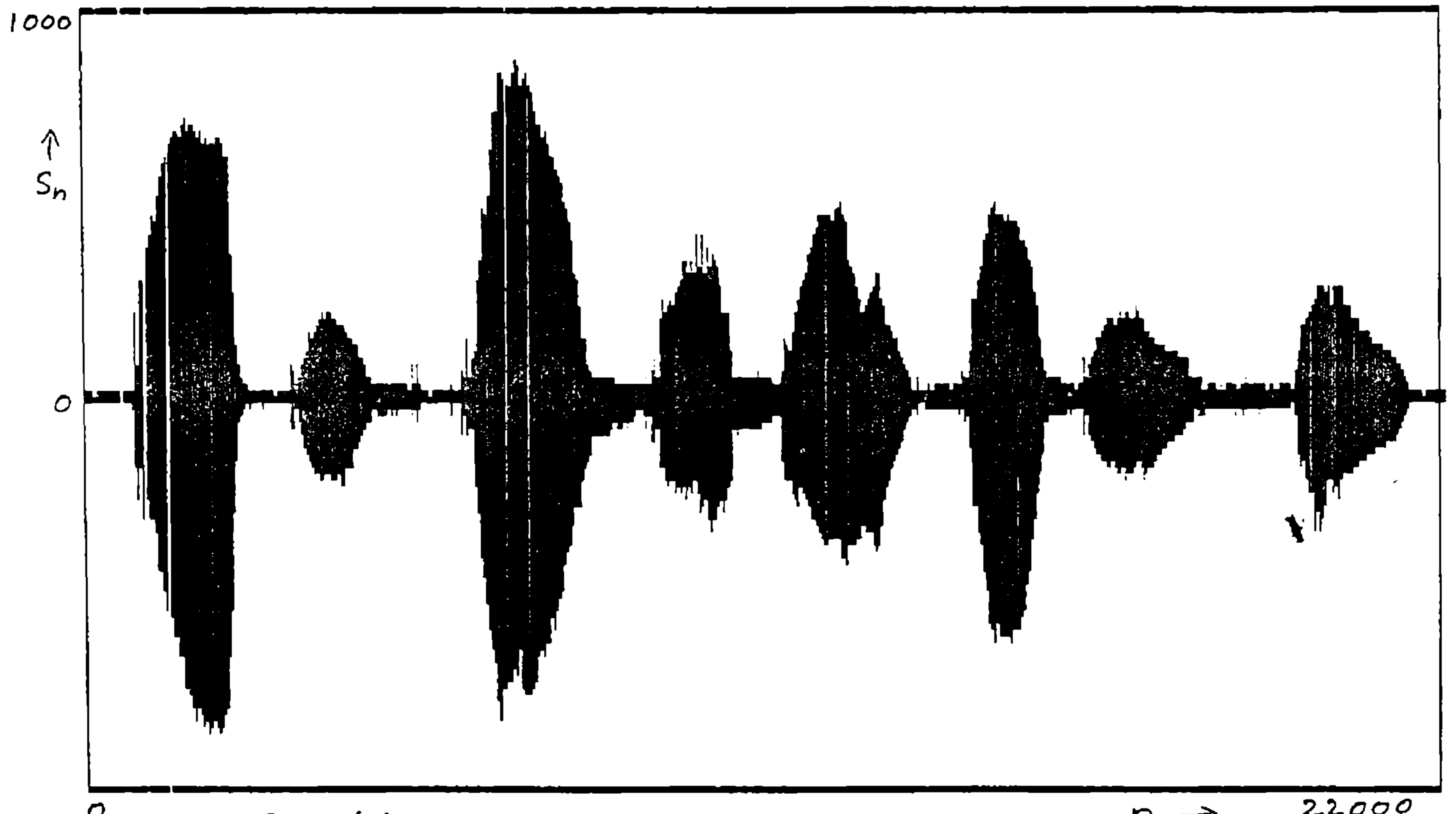


Fig 3(b)

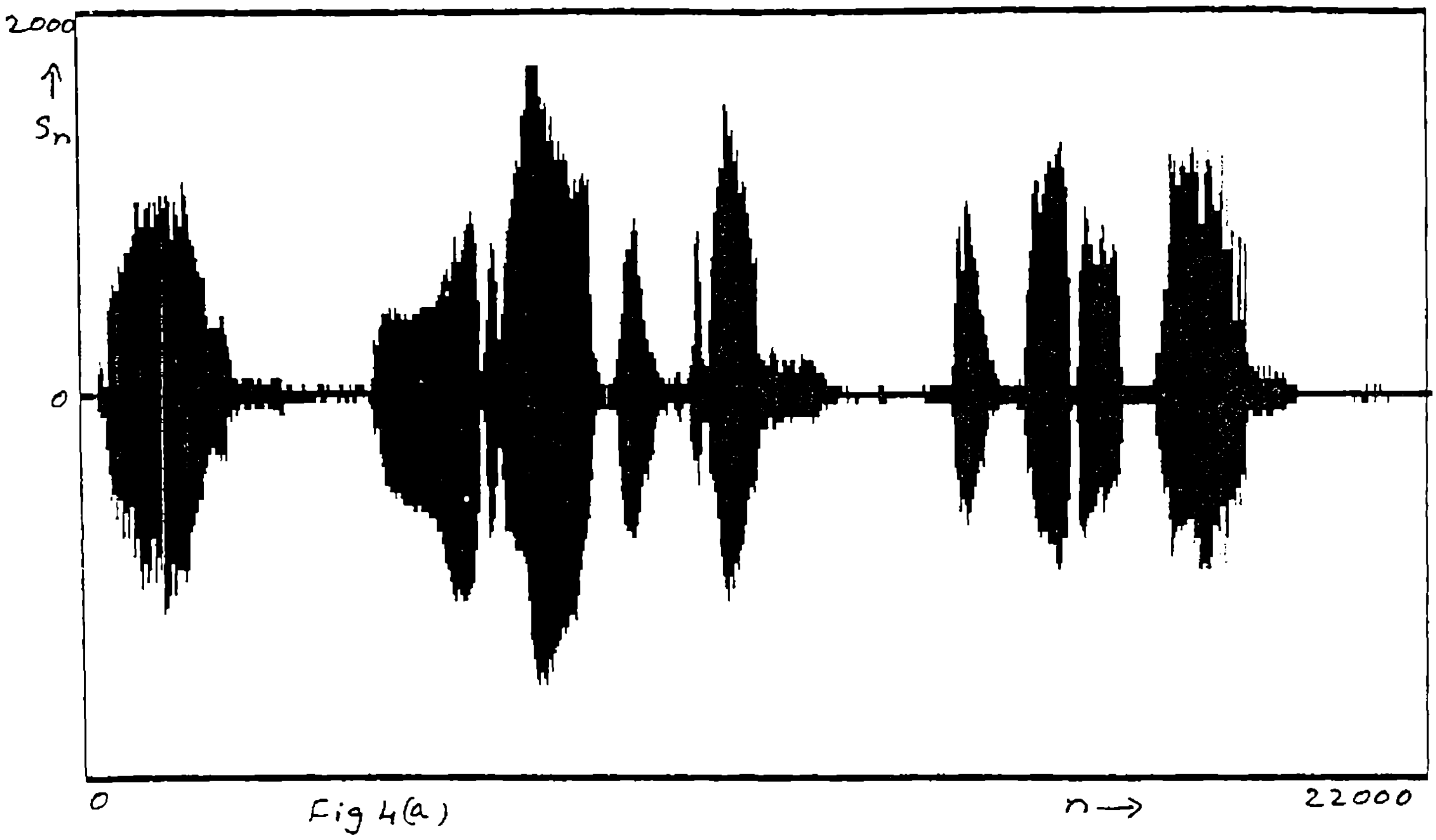


Fig 4(a)

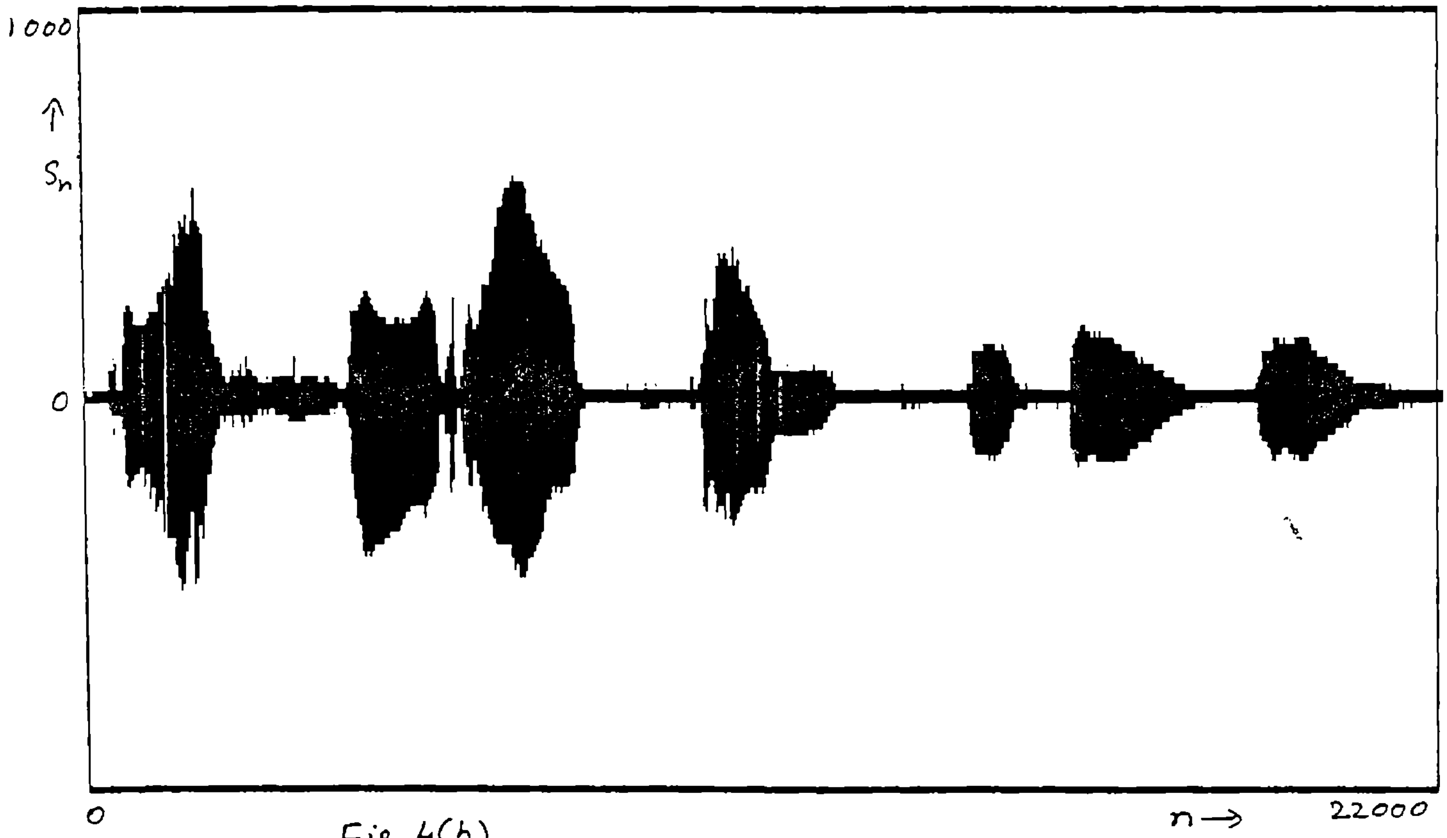


Fig 4(b)

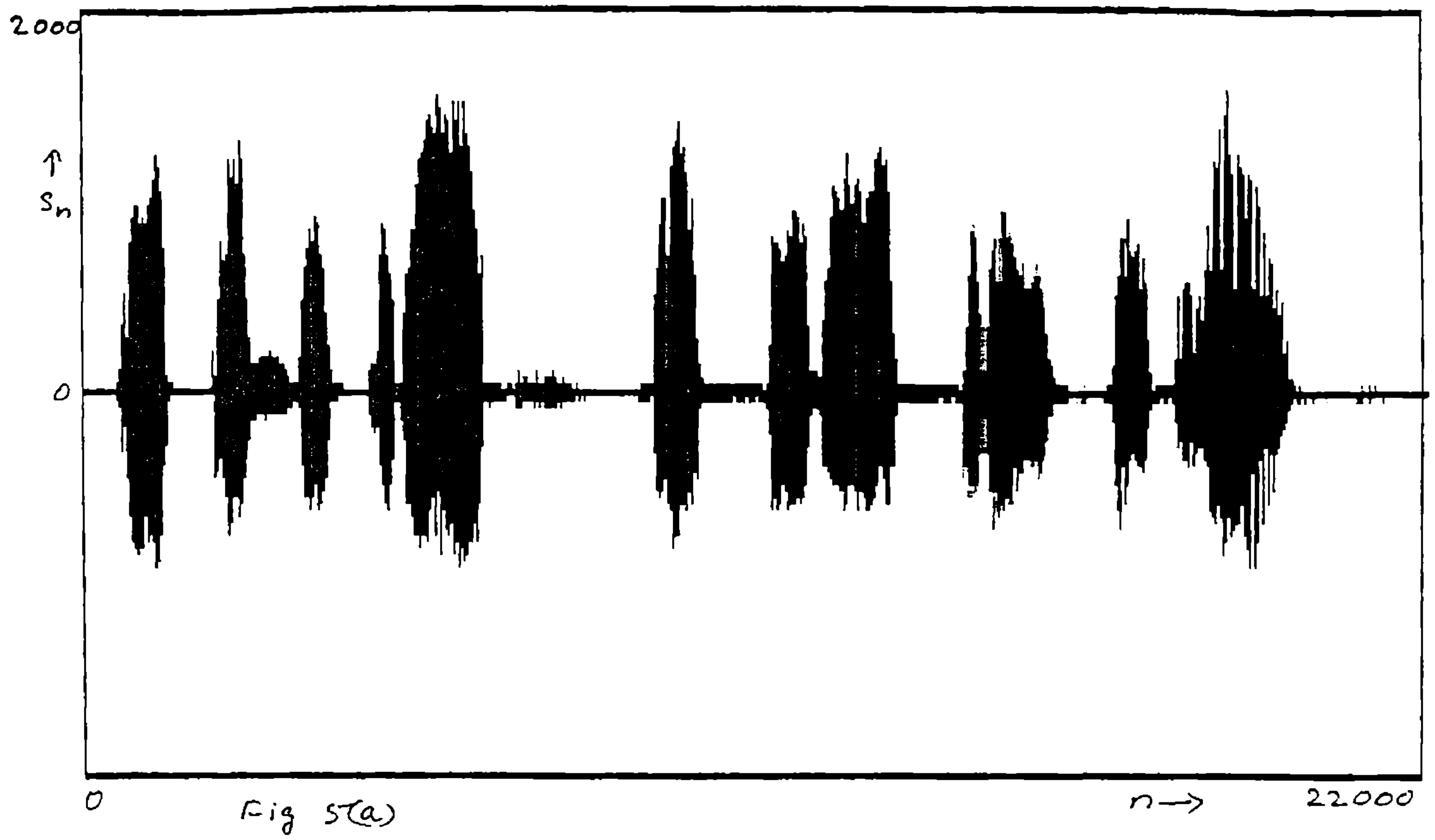


Fig 5(a)

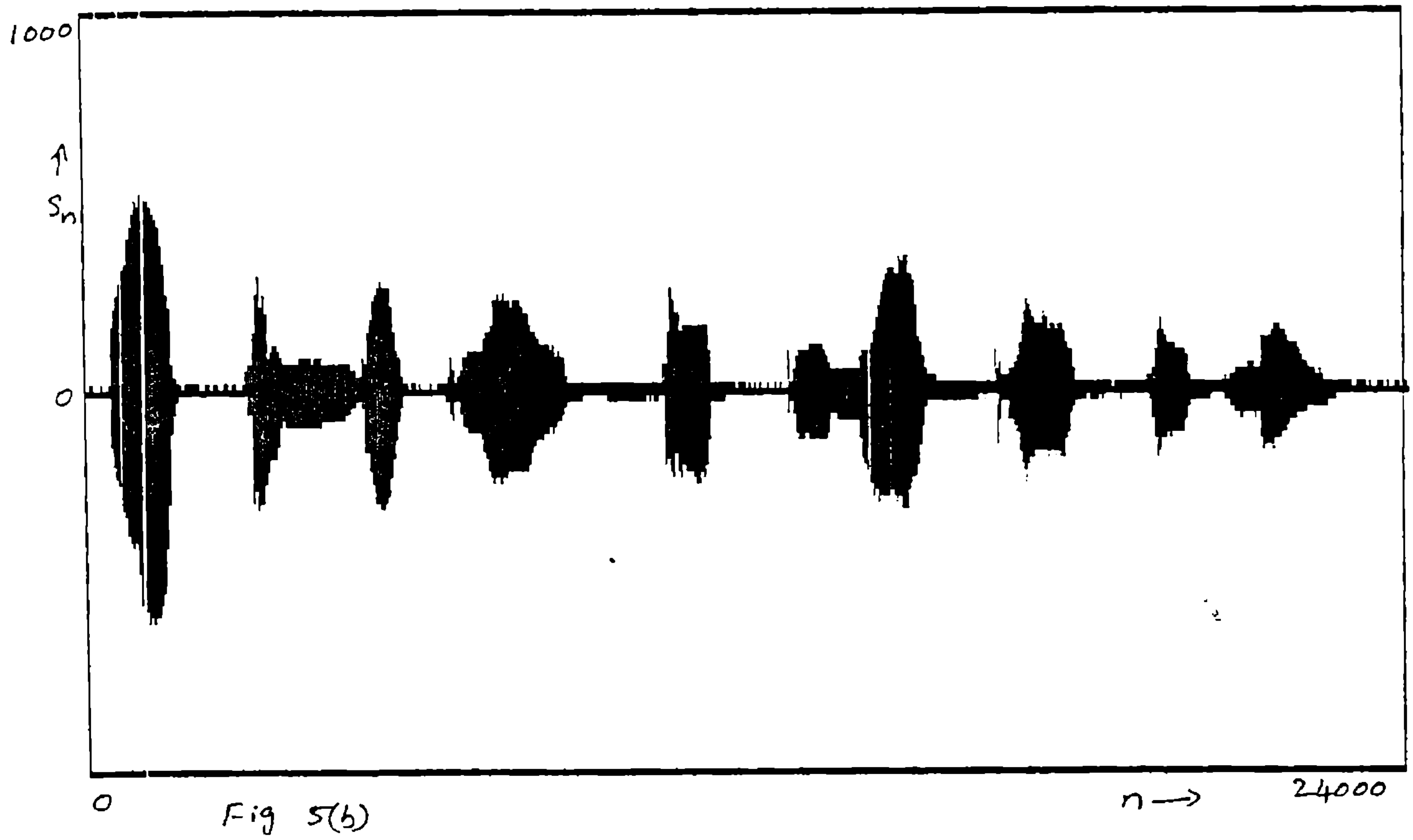


Fig 5(b)

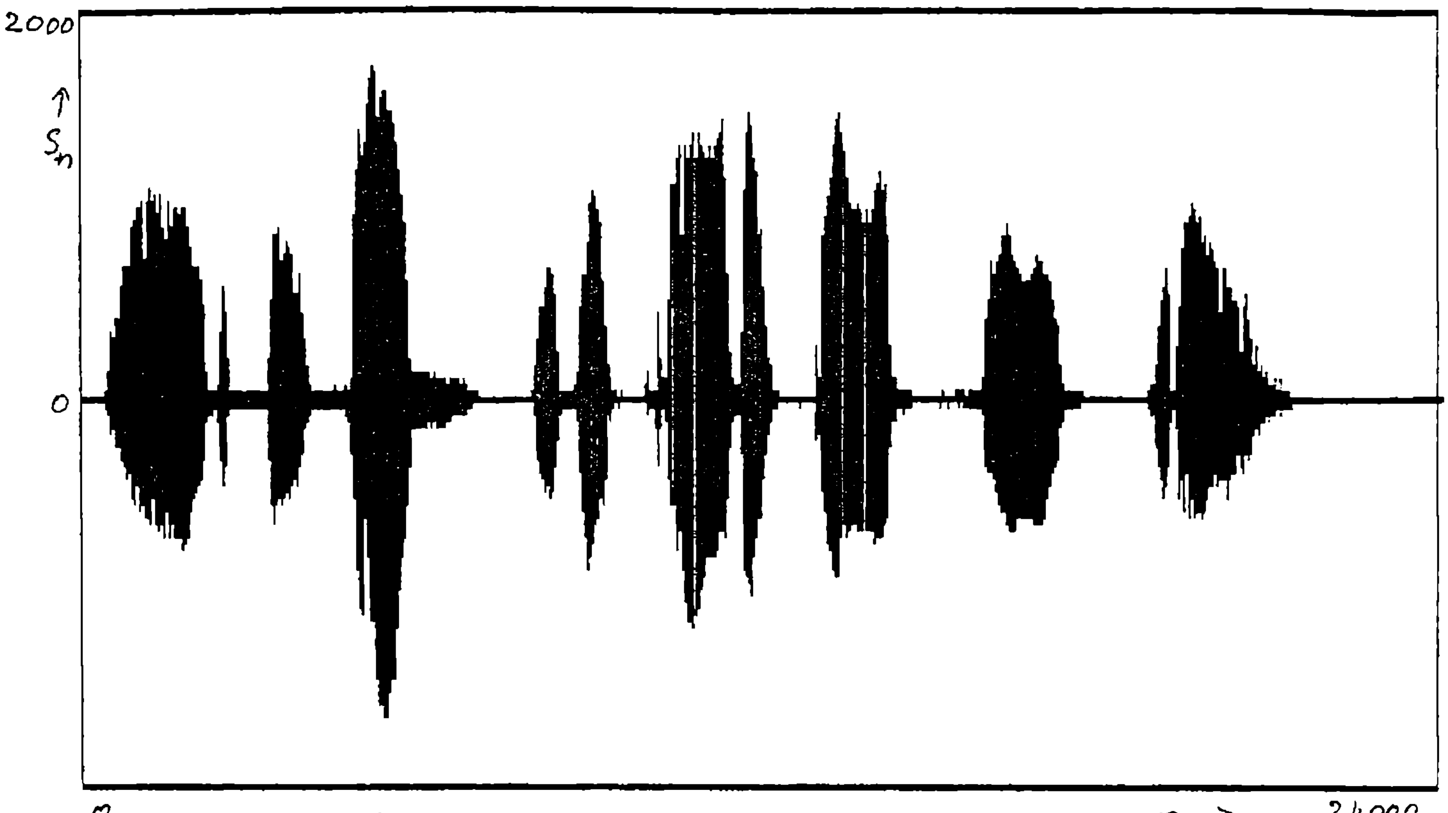


Fig 6(a)

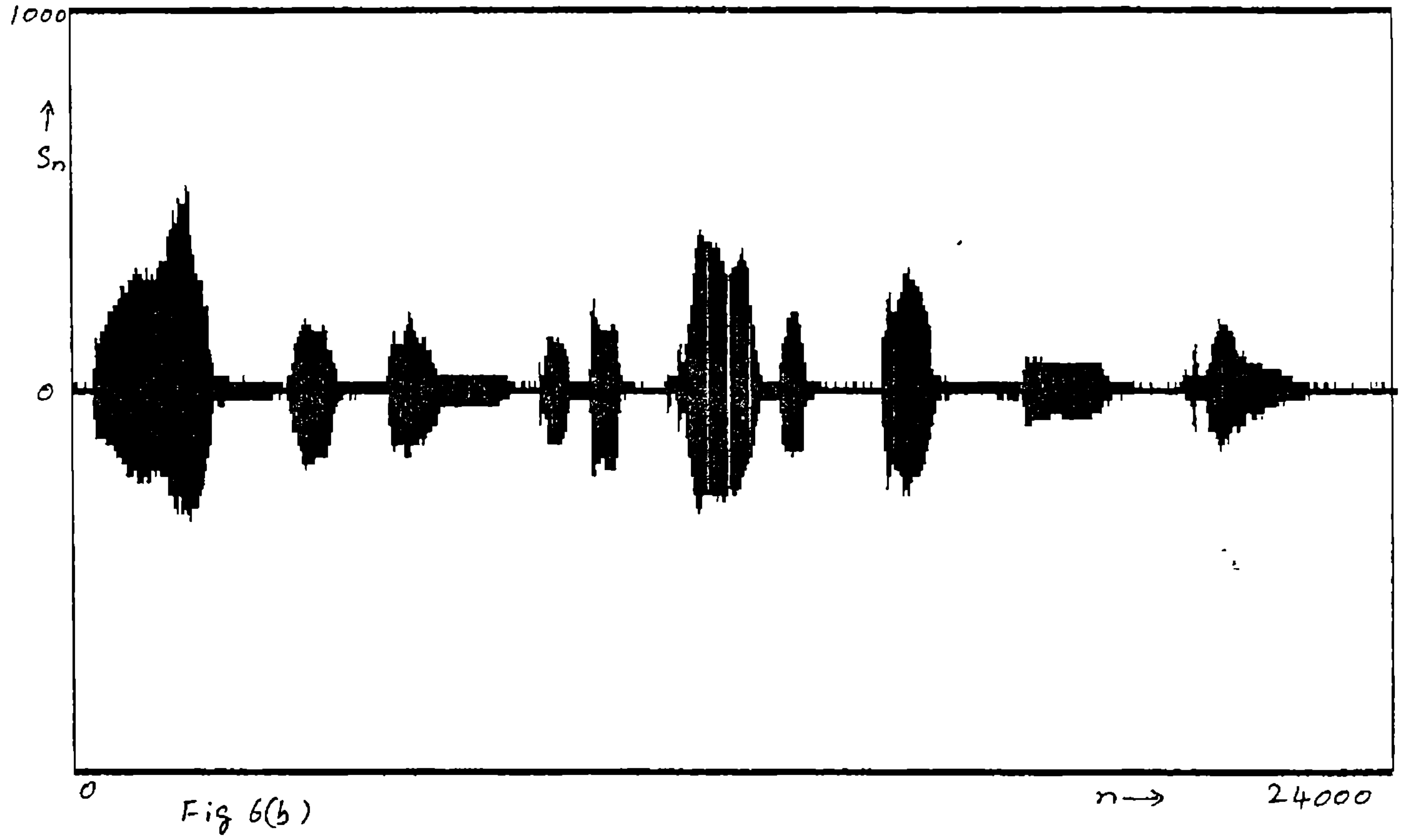


Fig 6(b)

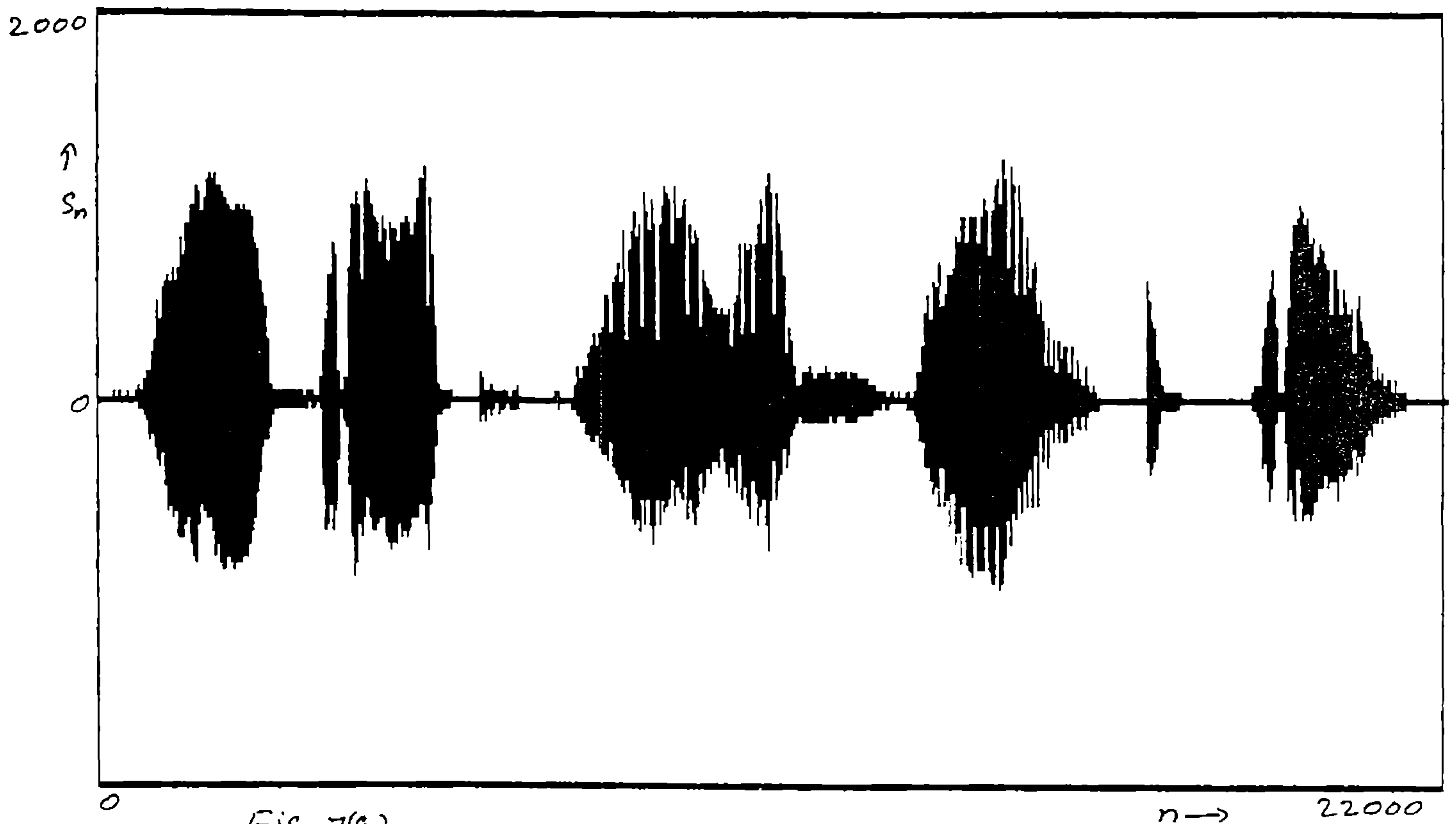


Fig 7(a)

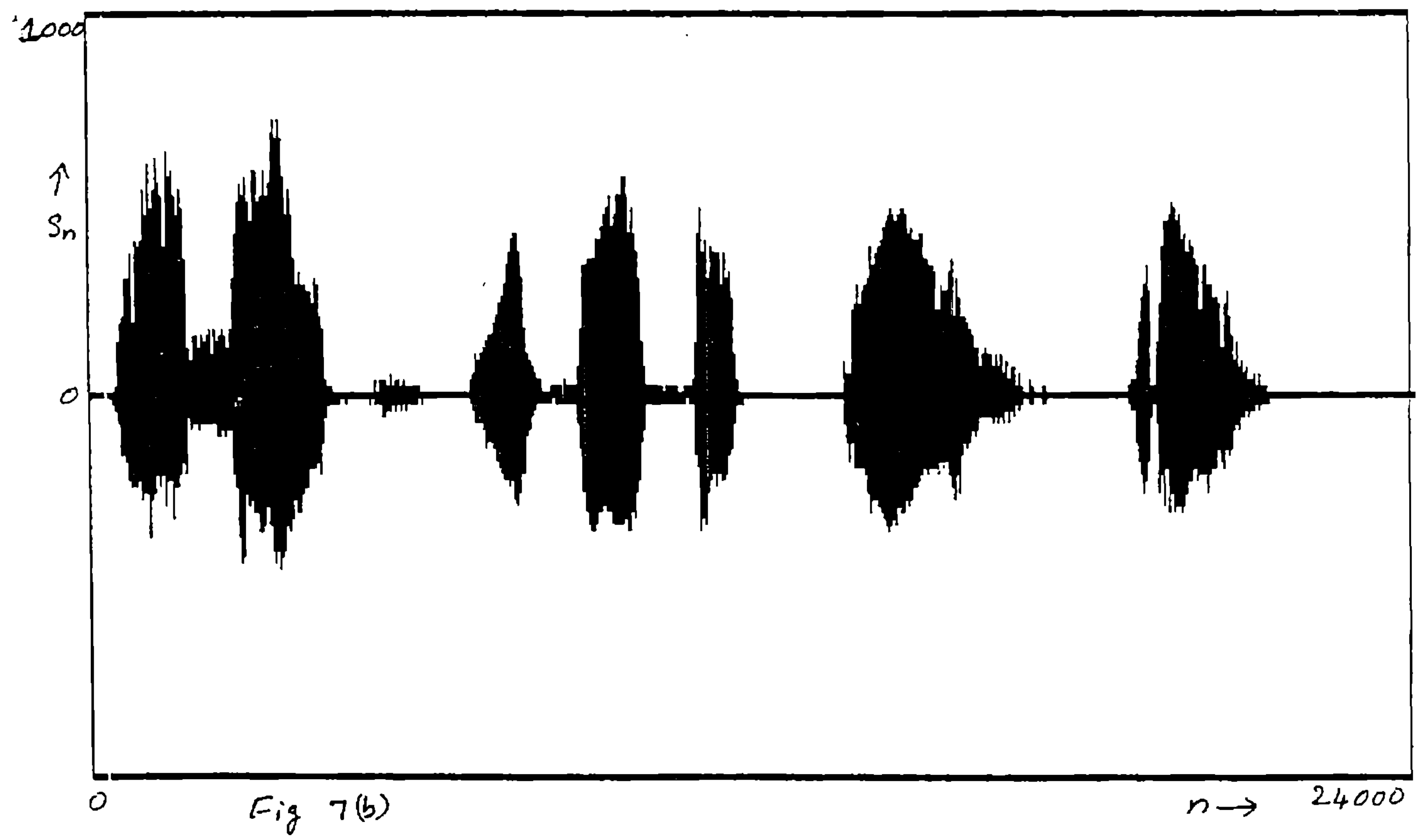


Fig 7(b)

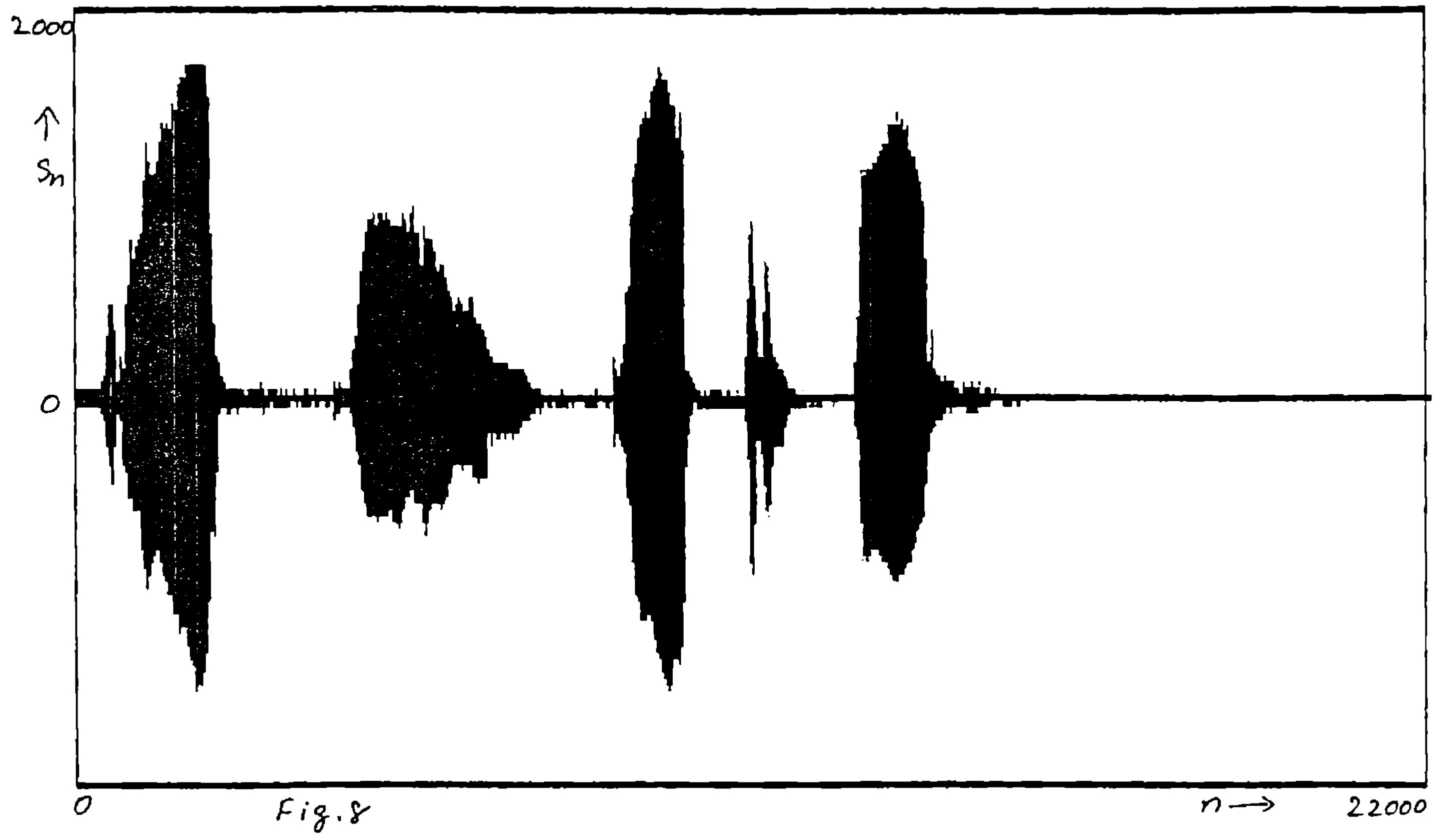


Fig. 8

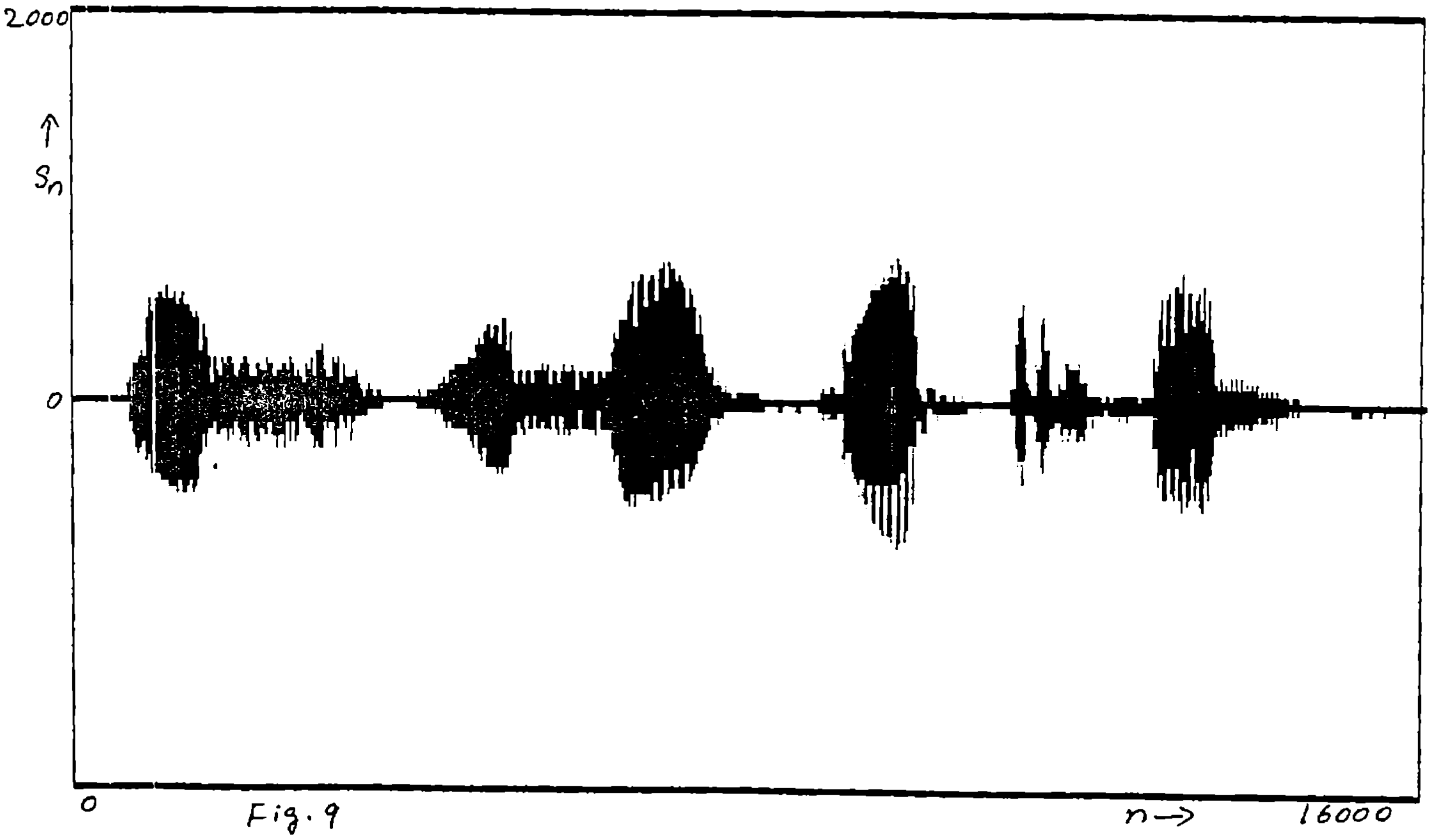


Fig. 9

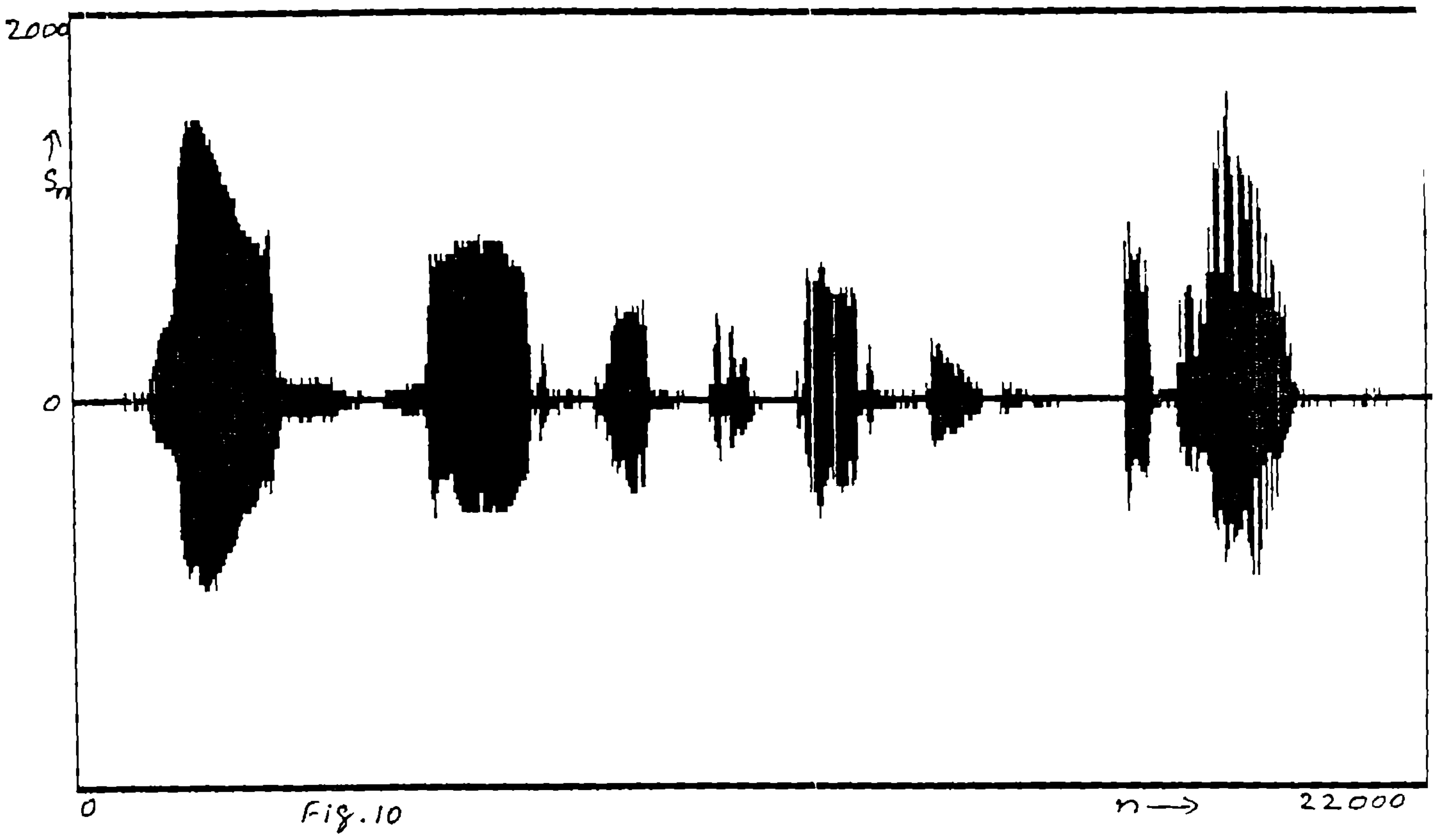


Fig. 10

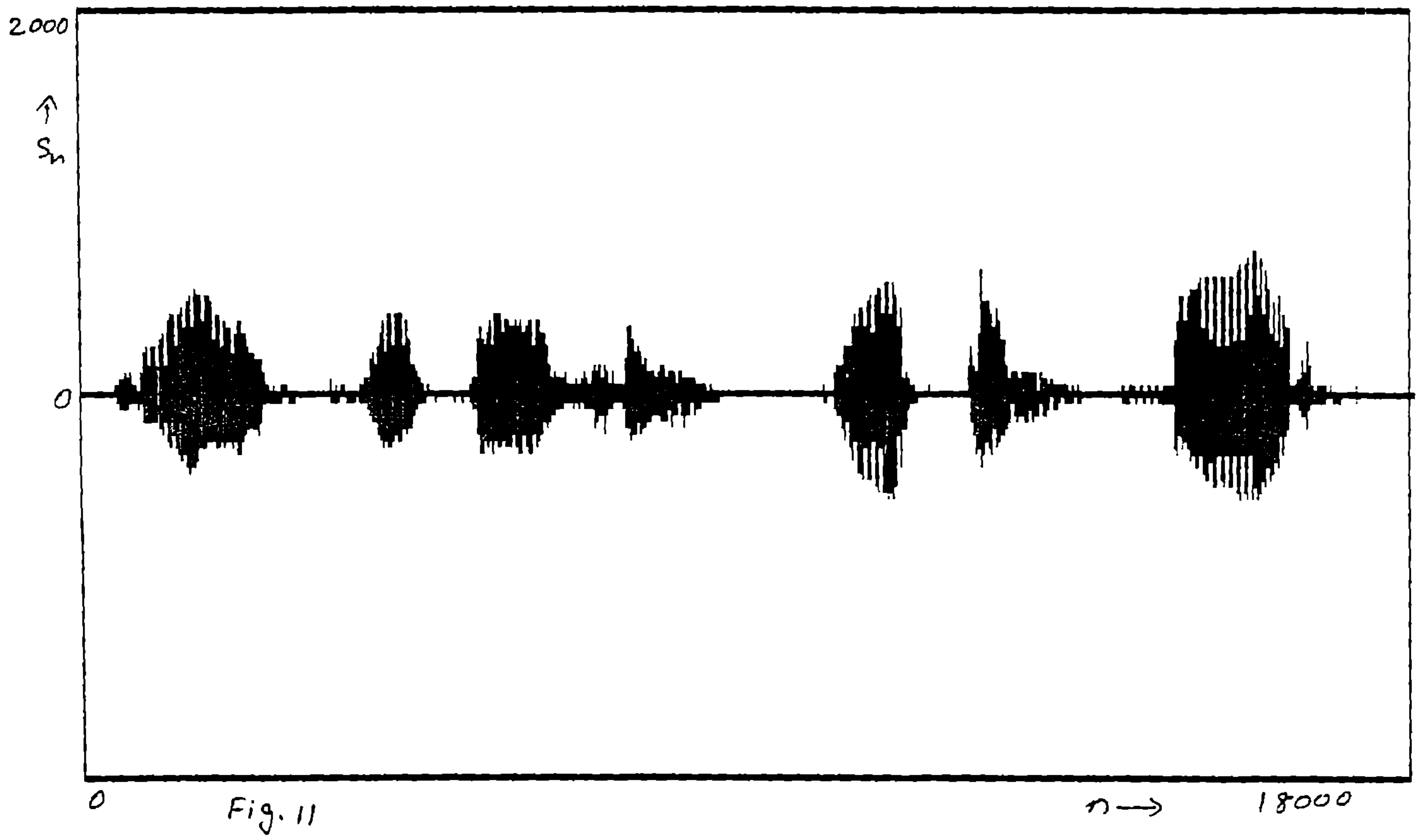


Fig. 11

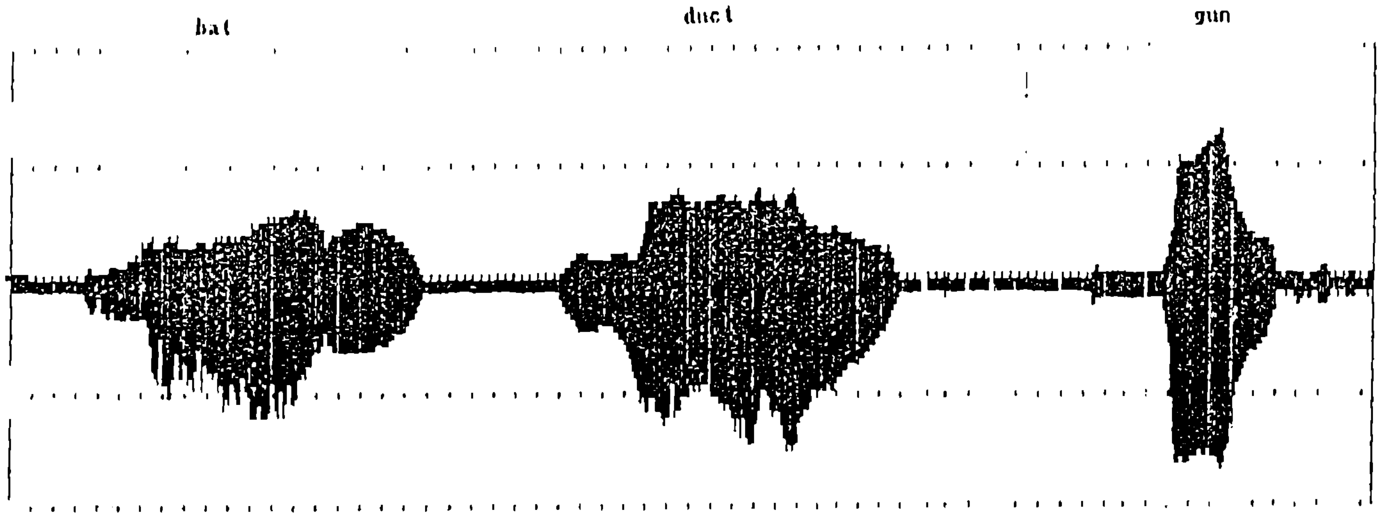


Fig 12(a)

VOICELESS PLOSIVES /b/, /d/, /g/

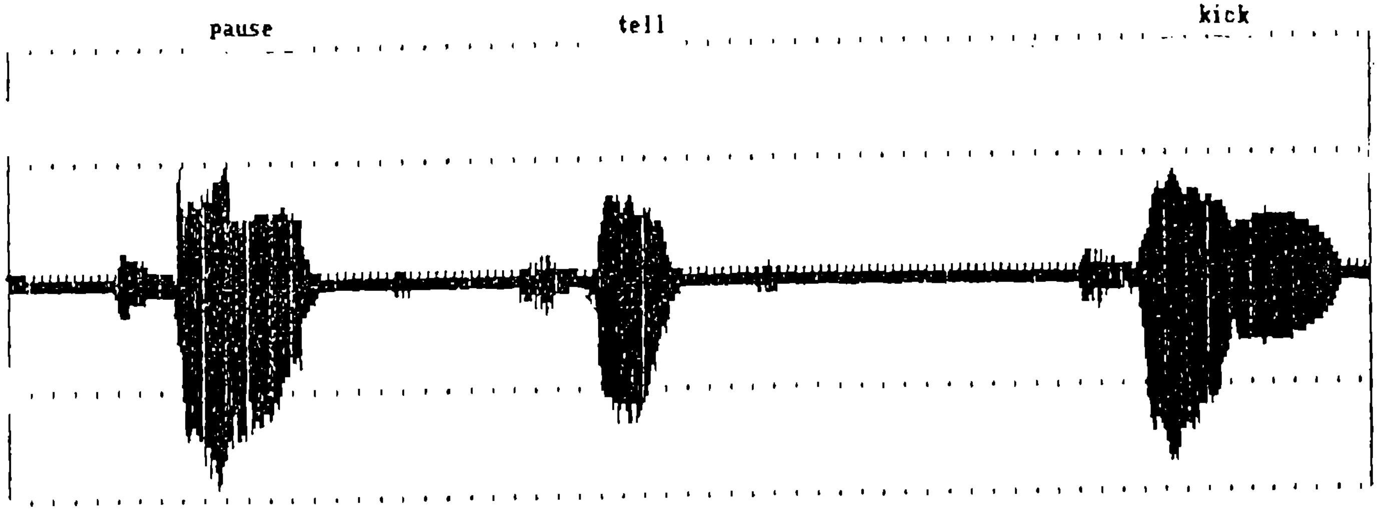
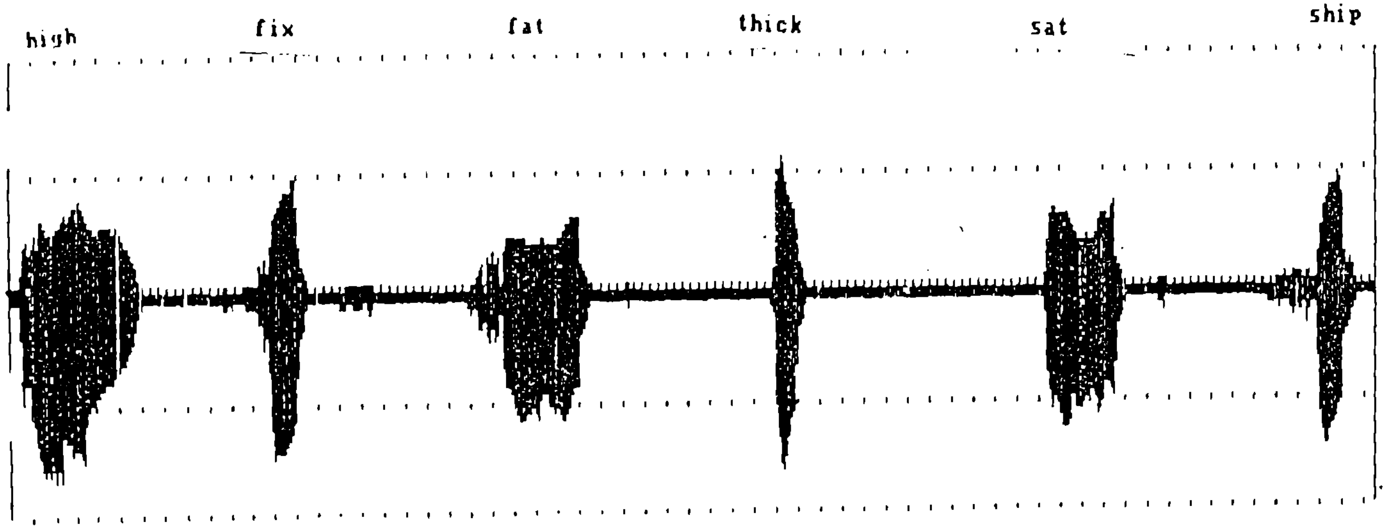


Fig. 12(b)

VOICED PLOSIVES /p/, /t/, /k/



VOICELESS FRICATIVES /f/, /θ/, /h/, /s/, /ʃ/.

Fig. 12(e)

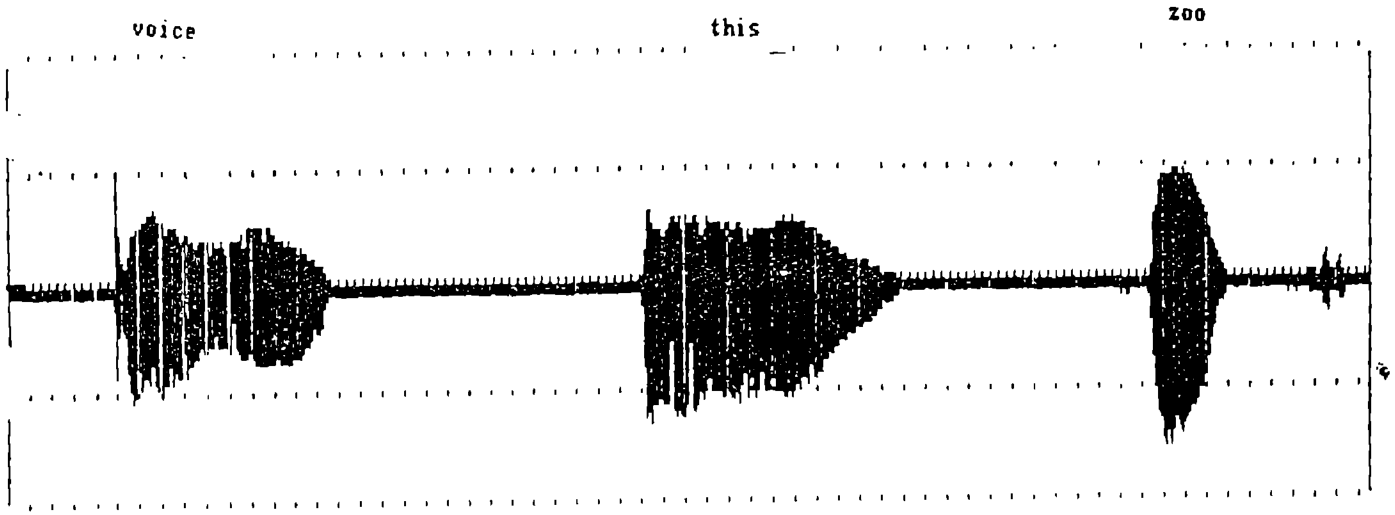
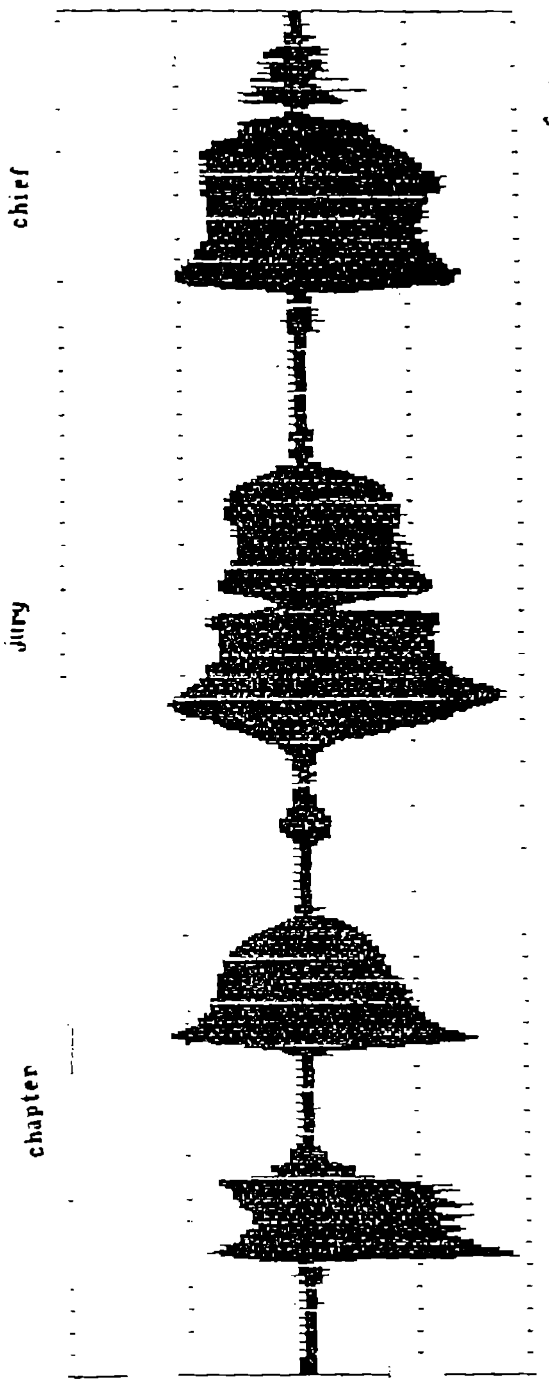


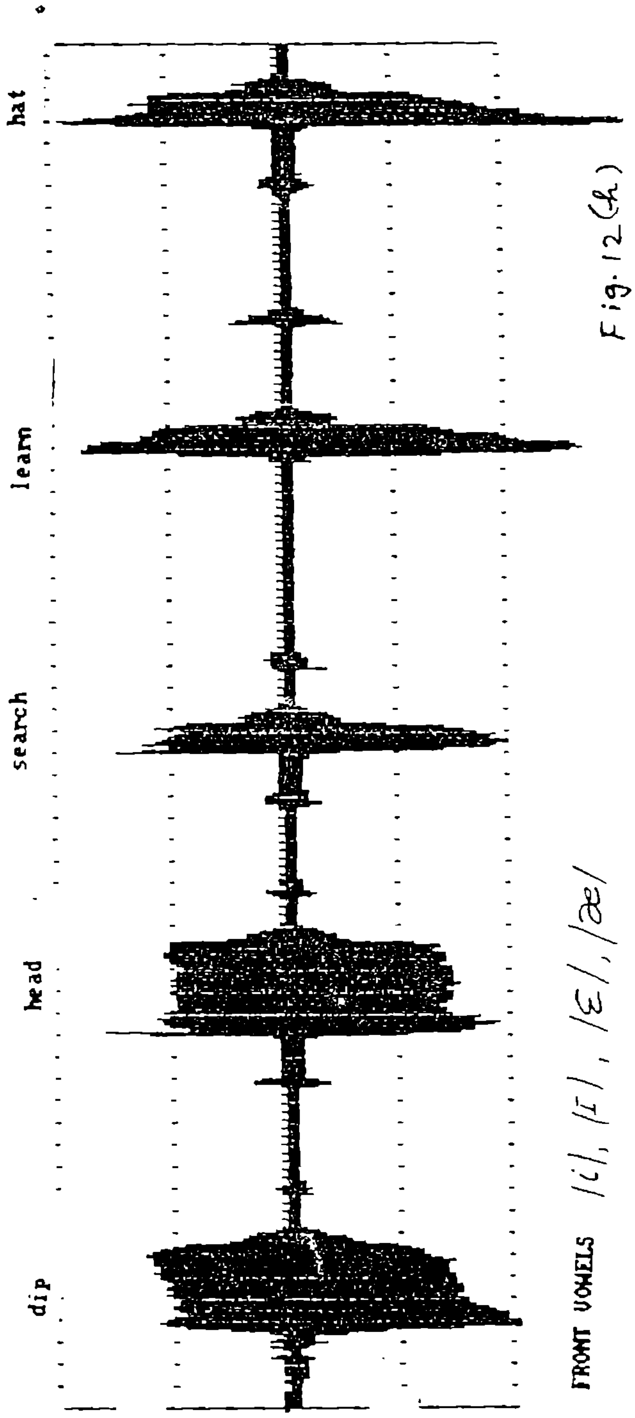
Fig. 12 (f)

VOICED FRICATIVES /v/, /ð/, /z/



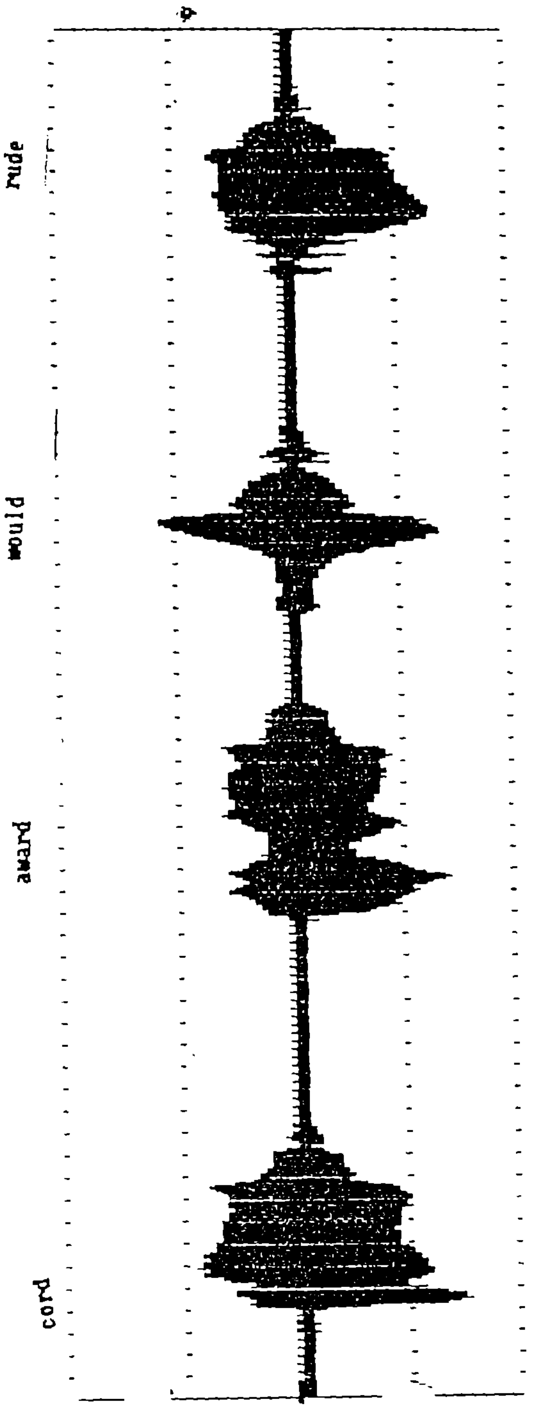
AFFRICATES /dʒ/, /tʃ/

Fig. 12(g)



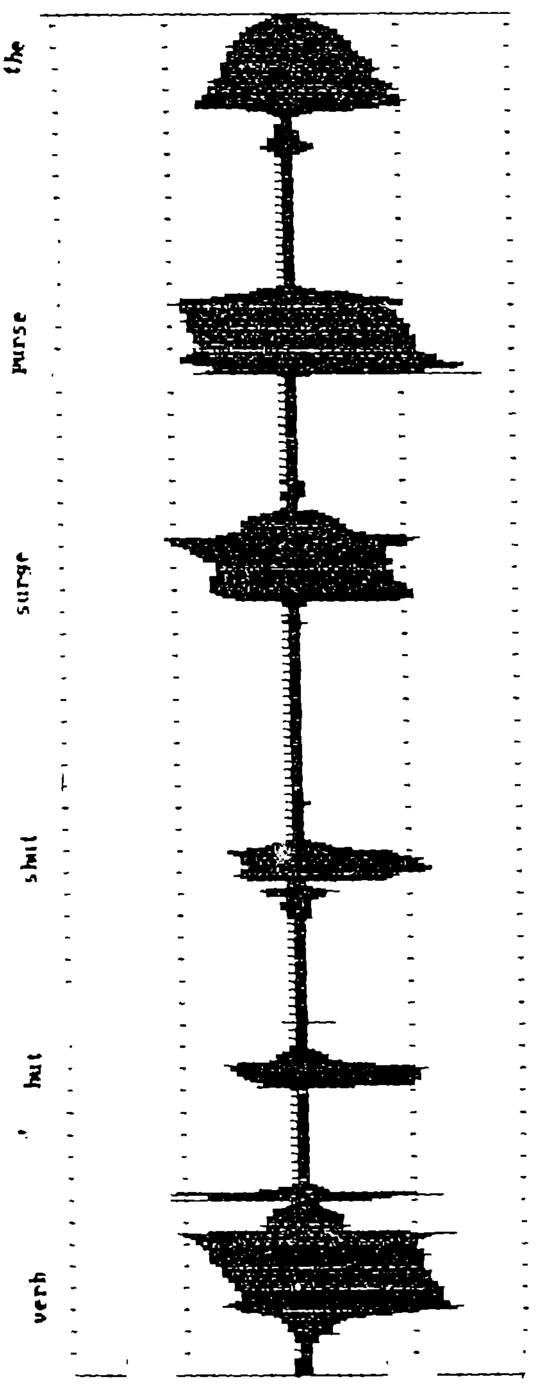
FRONT VOWELS /i/, /ɪ/, /ɛ/, /æ/

Fig. 12(h)



BACK VOWELS /ɑ/,/ɒ/,/ʊ/,/u/

Fig. 12(i)



MIDDLE VOWELS /ɜ/,/ʌ/,/ə/

Fig. 12(j)

REFERENCES

1. C.Wheddon and R.Linggard (Eds.), "Speech and Language Processing", Chapman & Hall, London, 1990.

2. N.S.Jayant and Peter Noll, "Digital Coding of Waveforms: Principle and Applications", Prentice-Hall, Inc., New Jersey, 1984.

3. L.R.Rabiner and R.W. Schafer, "Digital Processing of Speech Signals", Prentice Hall, Inc., New Jersey, 1978.

4. Ian H.Witten, "Principles of Computer Speech", Academic Press, London, 1982.

5. J.L.Flanagan, "Computers that Talk and Listen: Man-Machine Communication by Voice", Proc. IEEE, Vol.64, No.4, pp.416-432, 1976.

6. L.R.Rabiner and R.W.Schafer, "Digital Techniques for Computer Voice Response: Implementations and Applications", Proc. IEEE, Vol.64, pp.516-533, 1976.

7. J.L.Flanagan, M.R.Schroeder, B.S.Atal, R.E.Crochiere, N.S.Jayant and J.M.Tribolet, "Speech Coding", IEEE Trans. Commun., Vol.COM-27, pp.710-737, 1979.

8. Richard C.Rose and Thomas P.Barnwell, "Design and Performance of an Analysis-by-Synthesis Class of Predictive Speech Coders", IEEE Trans. Acoustics, Speech, and Signal Processing, Vol.38, No.9, pp.1489-1503, 1990.

9. Maurizio Copperi, "Rule-Based Speech Analysis and Application to CELP Coding", IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Vol.1, pp.143-146, 1988.

10. I.S.Dedes, D.R.Vaman and C.V.Chakravarthy, "Variable Bit Rate Adaptive Predictive Coder", IEEE Trans. Signal Processing, Vol.40, No.3, pp.511-517, 1992.

11. J.Makhoul, "Linear Prediction: A Tutorial Review", Proc. IEEE, Vol.63, No.4, pp.561-580, 1975.

12. B.S.Atal, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", J. Acoust. Soc. Amer., Vol. 47, 65(A), 1970.

13. B.S.Atal, "Characterisation of Speech Signals by Linear Prediction of the Speech Wave", Proc. IEEE Symp. on Feature Extraction and Selection in Pattern Recognition, pp.202-209, 1970.

14. B.S.Atal and S.L.Hanauer, "Speech Analysis and Synthesis by

Linear Prediction of the Speech Wave", J. Acoust. Soc. Amer., Vol.50, No.2, pp.637–655, 1971.

15. P.B. Denes, "Automatic Speech Recognition, Old and New Ideas", in 'Speech Recognition', Ed. D.Raj Reddy, Academic Press, 1975.

16. A.H.Frei, H.R.Schindler, P.Vettiger and E.Von Fellen, "Adaptive Predictive Speech Coding Based on Pitch-Controlled Interruption/Reiteration", in 'Waveform Quantisation and Coding', Ed.N.S.Jayant, IEEE Press, New York, 1976.

17. P.Cummiskey, N.S.Jayant and J.L.Flanagan, "Adaptive Quantisation in Differential PCM Coding of Speech", BSTJ, pp.1105–1118, 1973.

18. N.S.Jayant, "Digital coding of Speech Waveforms: PCM, DPCM and DM Quantizers", Proc. IEEE, Vol.62, pp.611–632, 1974.

19. N.S. Jayant, "Pitch-Adaptive DPCM Coding of Speech with Two-bit Quantisation and Fixed Spectrum Prediction", BSTJ., Vol.56, No.3, pp.439–454, 1977.

20. P.Noll, "A Comparative Study of Various Schemes for Speech Encoding", BSTJ, Vol.54, No.9, pp.1597–1614, 1975.

21. C.S.Xydeas, C.C.Evci and R.Steele, "Sequential Adaptive Predictors

for ADPCM Speech Encoders", IEEE Trans. Commun., Vol.COM-30, No.8, pp.1942-1954, 1982.

22. R.E.Crochiere, S.A. Webber and J.L.Flanagan, "Digital Coding of Speech in Sub-bands", BSTJ, Vol.55, No.8, pp.1069-1085, 1975.

23. W.R.Daumer, "Subjective Evaluation of Several Efficient Speech Coders", IEEE Trans. Commun., pp.567-573, 1982.

24. M.Honda, N.Kitawaki and F.Itakura, "Adaptive Bit Allocation Scheme in Predictive Coding of Speech", Proc. Int. Conf. Acoustics, Speech and Signal Processing, pp.1672-1675, 1982.

25. R.E.Crochiere, D.J.Goodman, L.R.Rabiner and M.R.Sambur, "Tandem Connections of Wideband and Narrowband Speech Communication Systems: Part 1-Narrow band-to-wideband Link", BSTJ., Vol.56, No.9, pp.1701-1722, 1977.

26. L.R. Rabiner, M.R. Sambur, R.E. Crochiere and D.J.Goodman, "Tandem connections of Wideband and Narrow band speech communication systems: Part 2-Wideband-to-Narrow band Link", BSTJ, Vol.56, No.9, pp.1723-1741, 1977.

27. R.E.Crochiere, "On the Design of Sub-band Coders for Low Bit Rate Speech Communications," BSTJ., Vol.56, No.5, pp.747-770,

1977.

28. R.E.Crochiere and M.R.Sambur, "A Variable-Band Coding Scheme for Speech Encoding at 4.8 kb/s", BSTJ, Vol.56, No.5, pp.771-779, 1977.

29. Fumio Amano, Kohei Iseda, Koji Qkazaki and Shigeyuki Unagami, "An 8 KBPS TC-MQ (Time Domain Compression ADPCM-MQ) Speech Codec", Proc. IEEE Int. Conf. Acoustics, Speech and Signal processing, Vol.1, pp.259-262, 1988.

30. R. Zelinski and P.Noll, "Adaptive Transform Coding of Speech Signals", IEEE Trans. Acoustics, Speech and Signal Processing, pp.299-309, 1977.

31. A.N. Netravali and J.O. Limb, "Picture Coding: A Review", Proc. IEEE, pp.366-406, 1980.

32. R.Zelinski and P.Noll, "Approaches to Adaptive Transform Speech Coding at Low Bit Rates", IEEE Trans. Acoustics, Speech and Signal Processing, Vol.27, No.1, pp.89-95, 1979.

33. J.B.Anderson and J.B.Bodie, "Tree Encoding of Speech", IEEE Trans. Information Theory, Vol.IT-21, pp.379-387, 1975.

34. L.C. Stewart, R.M. Gray and Y.Linde, "The Design of Trellis Waveform Coders", IEEE Trans. Commun., Vol.COM-30, pp.702-710, 1982.

35. H.G. Fehn and P.Noll, "Multipath Search Coding of Stationary Signals with Application to Speech", IEEE Trans. Commun., Vol. COM-30, pp.687-701, 1982.

36. Michael W.Marcellin, Thomas R.Fischer and Jerry D.Gibson, "Predictive Trellis Coded Quantisation of Speech", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Vol.1, pp.247--250, 1988.

37. Jerry D. Gibbson and Greg B.Haschke, "Backward Adaptive Tree Coding of Speech at 16 KBPS", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Vol.1, pp.251-254, 1988.

38. V.Cuperman and A.Gersho, "Adaptive Differential Vector Coding of Speech", Proc. IEEE Int. Conf. Global Telecommunications, Vol.3, pp.1092-1096, 1982.

39. H.Abut, R.M. Gray and G.Rebolledo, "Vector Quantization of Speech and Speech like Waveforms", IEEE Trans. Acoustics, Speech and Signal Processing, Vol.30, pp.423-436, 1982.

40. R.M. Gray and H.Abut, "Full Search and Tree Search Vector Quantization of Speech Waveforms", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.593-596, 1982.

41. A.Gersho, T. Ramstad and I.Versvik, "Fully Vector-Quantized Sub-band Coding with Adaptive Codebook Allocation", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.10.7.1-10.7.4, 1984.

42. A.Buzo, A.H. Gray Jr., R.M.Gray and J.D. Markel, "Speech Coding Based Upon Vector Quantization", IEEE Trans. Acoustics, Speech and Signal Processing, pp.562-574, 1980.

43. M.R. Schroeder, "Linear Predictive Coding of Speech: Review and Current Directions", IEEE Communications Magazine, Vol. 23, No.8, pp.54-61, 1985.

44. B.S. Atal and M.R.Schroeder, "Adaptive Predictive Coding of Speech Signals", BSTJ, Vol.49, pp.1973-1986, 1970.

45. B.S.Atal, "Predictive Coding of Speech at Low Bit Rates", IEEE Trans. Commun., pp.600-614, 1982.

46. B.S.Atal and M.R.Schroeder, "Predictive Coding of Speech Signals and Subjective Error Criteria", IEEE Trans. Acoustics, Speech and

Signal Processing, pp.247-254, 1979.

47. J.L.Flanagan, "Speech Analysis, Synthesis and Perception", Heidelberg, Springer-Verlag, New York, 1972.

48. P.Noll, "On Predictive Quantizing Schemes", BSTJ, pp.1499-1532, 1978.

49. Maurizio Copperi, "A Variable Rate Embedded-Code Speech Waveform Coder", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Vol.1, pp.212-215, 1982.

50. Maurizio Copperi, "A Robust 4800 BPS Full-band Speech Coder", Proc. IEEE Int. Conf. Global Telecommunications, Vol.1, pp.185-189, 1982.

51. Luciano Bertorello and Maurizio Copperi, "Design of a 4.8/9.6 KBPS Base - band LPC Coder Using Split-Band and Vector Quantization", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.1312-1315, 1983.

52. C.S.Southcott, I.Boyd, A.E.Coleman and P.G.Hammett, "Low Bit Rate Speech Coding for Practical Applications", in 'Speech and Language Processing', Ed.C.Wheddon and R.Linggard, Chapman & Hall, London, 1990.

53. M.D.Dankberg and D.Y.Wong, "Development of a 4.8–9.6 kbit/s RELP Vocoder", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.554–557, 1979.

54. Bruce Fette, Wilburn Clark, Cynthia Jaskie, Michelle Tugenberg and William Yip, "Experiments with a High Quality, Low Complexity 4800 bps Residual Excited LPC (RELP) Vocoder", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Vol.1, pp.263–266, 1988.

55. B.S. Atal and J.R.Remde, "A New Model of LPC Excitation for Producing Natural Sounding Speech at Low Bit Rates", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.614–617, 1982.

56. K.Ozawa, S.Ono and T.Araseki, "A Study on Pulse Search Algorithms for Multipulse Excited Speech Coder Realization", IEEE Journal on Selected Areas in Communications, Vol.SAC–4, No.1, pp.133–141, 1986.

57. M.Berouti, H.Garten, P.Kabal and P.Mermelstein, "Efficient Computation and Encoding of the Multipulse Excitation for LPC", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.10.1.1–10.1.4, 1984.

58. S.Singhal and B.S.Atal, "Improving the Performance of Multipulse LPC Coders at Low Bit Rates", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.1.3.1-1.3.4, 1984.

59. K.Ozawa and T.Araseki, "High Quality Multi-pulse Speech Coder with Pitch Prediction", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.1689-1692, 1986.

60. Shigeru Ono and Kazunori Ozawa, "2.4 KBPS Pitch Prediction Multi-pulse speech coding", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.175-178, 1988.

61. F.Deprette Ed and P.Kroon, "Regular Excitation Reduction for Effective and Efficient LP-Coding of Speech", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.25.8.1-25.8.4, 1985.

62. M.R.Schroeder and B.S.Atal, "Code-Excited Linear Prediction (CELP): High Quality Speech at Very Low Bit Rates", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp. 937-940, 1985.

63. Peter Kroon and Bishnu S.Atal, "Strategies for Improving the Performance of CELP Coders at Low Bit Rates", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.151-154, 1988.

64. W.B.Kleijn, D.J. Krasinski and R.H.Ketchum, "Improved Speech Quality and Efficient Vector quantisation in SELP", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.155-158, 1988.

65. Grant Davidson and Allen Gersho, "Multiple-Stage Vector Excitation Coding of Speech Waveforms", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.163-166, 1988.

66. L.A.Hernandez-Gomez, F.J.Casajus-Quiros and R.Garcia-Gomez, "High Quality Vector Adaptive Transform Coding at 4.8 kb/s", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.167-170, 1988.

67. A.M. Kondoz and B.G.Evans, "CELP Base-Band Coder for High Quality Speech Coding at 9.6 to 2.4 KBPS", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.159-162, 1988.

68. M.R.Schroeder and B.S.Atal, "Speech Coding Using Efficient Block Codes", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.1668-1671, 1982.

69. M.Nakatsui and P.Mermelstein, "Subjective Speech-to-Noise Ratio as a Measure of Speech Quality for Digital Waveform Coders", J. Acoust. Soc. Amer., 72, pp.1136-1144, 1982.

70. W.D.Voiers, “Diagnostic Evaluation of Speech Intelligibility” in ’Speech Intelligibility and Speaker Recognition’, Eds. M.Hawley, Dowden Hutchinson Ross, Stroudsburg, Pa., 1977.

71. B.S.Atal and M.R.Schroeder, “Predictive Coding of Speech Signals”, Proc. Int. Conf. on Communication and Processing, pp.360-361, 1967.

72. J.L.Flanagan, “Voices of Men and Machines”, J. Acoust. Soc. Amer. Vol.51, pp.1375-1387, 1972.

73. W.A.Ainsworth, “Mechanisms of Speech Recognition”, Pergamon Press, 1976.

74. J.D.Markel, “The SIFT Algorithm for Fundamental Frequency Estimation”, IEEE Trans. Audio and Electroacoustics, Vol.AU-20, pp.367-377, 1972.

75. L.R.Rabiner, M.J.Cheng, A.E.Rosenberg and C.A.McGonegal, “A Comparative Performance Study of Several Pitch Detection Algorithm”, IEEE Trans. Acoustics, Speech and Signal Processing, Vol.24, pp.399-417, 1976.

76. M.M.Sondhi, “New Methods of Pitch Extraction”, IEEE Trans. Audio

and Elec. Acoust., Vol.AU-16, pp.262-266, 1968.

77. Antoine Chouly and Hikmet Sari, "Six Dimensional Trellis-Coding with QAM Signal Sets", IEEE Trans. Commn., Vol.40, No.1, pp.24--33, 1992.

78. V.Cuperman, A.Gersho, R.Pettigrew and J.Yao, "Low Delay Vector Excitation Coding of Speech at 16 kb/s", IEEE Trans. Commun., Vol.40, No.1, pp.129-139, 1992.

79. Babu P.Anto, "Speaker Identification using Models for Phonemes", Ph.D. Thesis, Dept. of Electronics, Cochin University of Science and Technology, Cochin, 1991.

80. B.S.Atal and L.R.Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition", IEEE Trans. Acoustics, Speech and Signal Processing, Vol.24, pp.201-212, 1976.

81. L.R.Rabiner and M.R.Sambur, "Application of a LPC Distance Measure to the Voiced Unvoiced Silence Detection Problem", IEEE Trans. Acoust., Speech and Signal Processing, Vol.27, pp.338-343, 1979.

82. S.Knoor, "Reliable Voiced/Unvoiced Decision", IEEE Trans.

Acoustics, Speech and Signal Process., Vol.27, pp.263-267, 1979.

83. L.R. Rabiner and M.R.Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances", BSTJ., Vol.54, No.2, pp.297-315, 1975.

84. V.Ramamoorthy, "A Simple Time Domain Algorithm for Voiced/Unvoiced Detection", Signal Processing, Theories and Applications, EURASIP, pp.737-742, 1980.

85. N.K.Narayanan, "Speech Sample Estimation from Composite Zerocrossings and Encoding via Adaptive Switching of Transforms", Ph.D. Thesis, Dept. of Electronics, Cochin University of Science and Technology, Cochin, 1990.

86. Harris, F.J., "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform", Proc. IEEE, Vol.66, No.1, pp.51-83, 1978.

87. K.K.Paliwal, "Effect of Spectral Flattening on the Pitch Estimation Performance of the Autocorrelation Method for Noisy Speech", Acoustics Letters, Vol.7, No.5, 1983.

88. B.S.Atal, M.R.Schroeder and V.Stover, "Voice Excited Predictive Coding System for Low Bit Rate Transmission of Speech", Proc.

ICC, pp.30.37 to 30.40, 1975.

89. M.R.Sambur, "Selection of Acoustic Features for Speaker Identification", IEEE Trans. Acoustics Speech and Signal Processing, Vol.23, No.2, pp.176-182, 1975.

90. L.S.Su, K.P.Li and K.S.Fu, "Identification of Speakers by use of Nasal Coarticulation", J. Acoust. Soc. Amer., Vol.55, pp.1876-1882, 1974.

91. B.S.Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification", J. Acoust. Soc. Am., Vol.55, pp.1304-1312, 1974.

92. Frank Fallside and William A.Woods, Eds., "Computer Speech Processing", Prentice Hall Inc., New Jersey, 1985.

93. Andrej Ljolje and Stephen E.Levison, "Development of an Acoustic Phonetic Hidden Markov Model for Continuous Speech Recognition", IEEE Trans. Signal Processing, Vol.39, No.1, pp.29-39, 1991.

LIST OF PUBLICATIONS OF THE AUTHOR

1. "A New Method of Adaptive Linear Prediction for Voiced Speech Signals", J. Acoust. Soc. of India, Vol.XVIII (3&4), pp.160-163, 1990.

2. "A Knowledge-based Speaker Recognition System", Advances in Modelling and Analysis, B, AMSE Press, France, Vol.27, No.2, pp.53–63, 1993.

3. "A Block Adaptive Model for Speech Signals", Accepted for Publication in the Journal of the Institution of Electronics and Telecommunication Engineers.

4. "New Methods for the Detection of Voiced, Unvoiced, Silent and Transition Regions in Speech Signals", Accepted for presentation at the AMSE International Conference on 'Information Processing', to be held at Orlando, in October 1993.

5. "A Phoneme Identification System", Accepted for presentation at the AMSE International Conference on 'Signals, Data, Systems', to be held at Bangalore, in December 1993.