SPEECH SAMPLE ESTIMATION FROM COMPOSITE ZEROCROSSINGS AND ENCODING VIA ADAPTIVE SWITCHING OF TRANSFORMS

A THESIS SUBMITTED BY

N. K. NARAYANAN

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
COCHIN - 682 022
INDIA

FEBRUARY 1990

Dedicated

To my loving mother
and to the
memory of my beloved father

CERTIFICATE

This is to certify that the thesis entitled "SPEECH SAMPLE ESTIMATION FROM COMPOSITE ZEROCROSSINGS AND ENCODING VIA ADAPTIVE SWITCHING OF TRANSFORMS" is a report of the original work carried out by Mr.N.K.Narayanan under my supervision and guidance in the Department of Electronics, Cochin University of Science and Technology and that no part thereof has been presented for the award of any other degree.

Dr.C.S.Sridhar
Professor
Department of Electronics
Cochin University of
Science and Technology

Cochin 682 022
February 17, 1990

DECLARATION

I hereby declare that the work presented in this thesis is based on 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 that no part thereof has been presented for the award of any other degree.

N.K.Narayanan.

Cochin 682 022
February 17, 1990

ACKNOWLEDGEMENTS

This thesis is the outcome of the education, guidance and inspiration I received from my guide Dr.C.S.Sridhar, Professor of Electronics and Dean, Faculty of Technology, Cochin University of Science and Technology. I wish to express my deep gratitude to him for his invaluable help in steering the course of this study.

I am extremely grateful to Dr.K.G.Nair, Professor and Head of the Department of Electronics for his encouragement and the genuine interest he has shown in this work from time to time.

My special thanks are due to my colleague Mr.Babu P.Anto for his wholehearted co-operation, encouragement and help throughout this endeavour. I am thankful to all the members of the Faculty, Laboratory and non-teaching staff and the research scholars of the Department of Electronics for their kind hearted co-operation during the course of this work.

I thank Dr.N.S.Jayant, AT&T Bell Laboratories, New Jersey, Dr.Allen Gersho, Professor of Elect. & Computer Engineering, University of California, Dr.B.Yegnanarayana, Professor of Computer Science & Engineering, IIT, Madras and Dr.K.K.Paliwal, Reader, Computer Systems and Communications Group, TIFR, Bombay for extending their help and co-operation during the tenure of this work.

I am indebted to the University Grants Commission for generously awarding me a Teacher Fellowship under the Faculty Improvement Programme and the Principal, Government Brennen College, Tellicherry and the Government of Kerala for granting deputation.

Sincere help and co-operation from Mr.K.P.Sasidharan and Mr.V.M.Peter for making this thesis in its present form are gratefully acknowledged.

ABSTRACT

"SPEECH SAMPLE ESTIMATION FROM COMPOSITE ZEROCROSSINGS AND ENCODING VIA ADAPTIVE SWITCHING OF TRANSFORMS"

By

N.K.NARAYANAN

This thesis investigates the potential use of zerocrossing information for speech sample estimation. It provides a new method to estimate speech samples using composite zerocrossings. A simple linear interpolation technique is developed for this purpose. By using this method the A/D converter can be avoided in a speech coder. The newly proposed zerocrossing sampling theory is supported with results of computer simulations using real speech data.

The thesis also presents two methods for voiced/ unvoiced classification. One of these methods is based on a distance measure which is a function of short time zerocrossing rate and short time energy of the signal. The other one is based on the attractor dimension and entropy of the signal. Among these two methods the first one is

simple and requires only very few computations compared to the other. This method is used in a later chapter to design an enhanced Adaptive Transform Coder.

The later part of the thesis addresses a few problems in Adaptive Transform Coding and presents an improved ATC. Transform coefficient with maximum amplitude is considered as 'side information'. This enables more accurate bit assignment and step-size computation. A new bit reassignment scheme is also introduced in this work. Finally, an ATC which applies switching between Discrete Cosine Transform and Discrete Walsh-Hadamard Transform for voiced and unvoiced speech segments respectively is presented. Simulation results are provided to show the improved performance of the coder.

CONTENTS

Page

Chapter 1

INTRODUCTION

1.0 BACKGROUND

Digitization and coding of speech for economical transmission and storage have long been of major engineering concern. The fast evolution of digital hardware technology in the last decade has had a great impact on speech research. Digital speech is less sensitive to noise and can be encripted and stored in computer compatible media. Current trends for world-wide communications in the 1990s and beyond, point to a proliferation of digital transmission as a dominant means of communication for voice and data. Digital speech is also used in the newly emerging man to machine and machine to man voice communication. This involves intensive digital signal processing for the purpose of voice identification, recognition and synthesis [N.S.Jayant and Peter Noll, 1984], [L.R.Rabiner and R.W.Schaffer, 1978]. Another form of digital speech communication is voice-store-and forward (voice mail), which is expected to be an integral part of the future automated office.

Efficient representation of the speech signal in terms of a compact sequence of binary digits (bits) is the basic requirement common to all the above applications of digital speech. The process of converting the analog speech signal into a bit stream is referred to as speech coding. A system that performs this job is called a speech coder, and it is usually accompanied by a speech decoder, which reconstructs the speech signal from its digital representation. The combined speech coder and decoder is sometimes called 'speech codec'.

Speech coders are basically divided into two categories: waveform coders and voice coders (vocoders). The aim of waveform coders is to preserve the shape of the original waveform, that is, to minimize the distortion between the original speech waveform and the reconstructed one. Good quality waveform coders exist in the bit-rate ranging from 16 kbits/s to 64 kbits/s.

Unlike waveform coders, vocoders do not preserve the shape of the waveform. Rather, they are based on mimicking the speech production model. For this purpose, speech is modelled as an output of a linear filter driven by a highly constrained excitation signal. The coder analyzes the input speech, estimates and transmits the filter and excitation parameters. The decoder retrieves these parameters and synthesizes a replica of the speech waveform. Even though the reconstructed waveform may not look like the original, it sounds similar to the original, with a distinct synthetic quality. However, the intelligibility is preserved in this method. Vocoding quality is generally regarded as inadequate for general purpose voice communication. Since vocoders preserve only a very small amount of information they are able to compress speech into very low bit rates. Commercial vocoders operate at 2400 bits/s. A fundamental problem in speech coding is to achieve the minimum possible distortion for a given transmission rate. An important parameter in solving this problem is the cost of encoding or the coder complexity. Due to the advances in digital technology and digital signal processing many coding techniques evolved. These range from the oldest and simplest one--Pulse Code Modulation (PCM) which is a low complexity coder operating at 56-64 kbits/s, to Adaptive Predictive Coding (APC) and Adaptive Transform Coding (ATC)--which are medium and high complexity coders operating at 16 to 32 kbits/s to give good speech quality [J.L.Flanagan et al, 1979].

1.1 MOTIVATION

In all the coders mentioned above, a continuous analog speech signal is to be converted to a discrete form. This involves sampling in the time domain, quantization in the amplitude domain and coding the resultant information into digital form. In the conventional coding systems this process is done using an A/D converter. To find a simple digitizing method is a main motivation of this thesis. To this end we investigate the potential use of zerocrossing informations for speech sample estimation. The zerocrossing based approach requires no sampling of the signal while A/D converter method relies on multilevel quantization of samples taken at prescribed instants of time. With simple digital circuits, it is easier to measure the timings of zerocrossings. Some potential advantages of this approach are that the inherent problems of an A/D converter such as alignment, limited dynamic range and speed can be removed and hardware associated with the sampling circuitry can be simplified.

Another main purpose of this study is to explore the potential capability of Adaptive Transform Coding (ATC) and to find efficient ATC systems for waveform coding.

In this work we propose to use a modified ATC coder to achieve good quality of speech coding at 8 to 16 kbits/s. The basic approach is to transmit the maximum amplitude of the transform coefficient as side information so as to obtain a better stepsize computation and to use a modified bit reassignment scheme. In a later chapter we propose and study an Adaptive Switching Transform Coder (ASTC) based on a simple voiced/unvoiced classification algorithm. This system gives a notable improvement in performance at the expense of a moderate increase in coder complexity.

1.2 OUTLINE OF THE WORK AND MAIN RESULTS

The intent of chapter 2 is to establish a necessary background for the following chapters. The former part of the chapter contains a brief review of the state of art of speech waveform coding. The later part gives a brief survey of different methods for measuring the speech quality. The objective measures like signal-to-noise ratio (SNR), articulation index, log spectral distance, Itakura's likelihood ratio, and Euclidean distance and subjective tests like Diagnostic Rhyme Test (DRT), the Modified Rhyme Test (MRT), the Diagnostic Acceptability Measure (DAM), and Mean Opinion Score (MOS) are reviewed in this section.

Among these the SNR measure is used throughout the thesis because of its simplicity in implementation.

Chapter 3 presents the use of zerocrossing information for estimation of speech samples. In the first part of the chapter we summarise the theory associated with zerocrossing sampling method used to estimate the speech signal. A simple linear interpolation formula for signal estimation from composite zerocrossings is developed. The results of computer simulation experiment, that verifies this approach is presented in the later part.

In chapter 4 we study two methods for voiced/ unvoiced classification. The first one is based on short time zerocrossing rate (STZCR) and short time energy (STE) of speech signal and the second method is based on the second order attractor dimension D_2 and second order Kolmogorov entropy K_2 of speech signal. In the first method, a distance measure is defined as the ratio of STZCR to STE. If this distance is greater than a threshold value, then the segment is classified as unvoiced and otherwise as voiced. Verification of this approach is conducted by manual classification.

We obtain the computed value of the D_2 and K_2 for voiced and unvoiced speech segments and observe that these values are larger for unvoiced speech. The method developed for voiced/unvoiced classification based on D_2 and K_2 requires larger computation time compared to the previous one.

Based on the results obtained in the preceding chapters, in chapter 5, we analyze the applicability of zerocrossing information and the second order attractor dimension and entropy for low bit rate coding. The estimation method developed in chapter 3 can be used with existing waveform coders to reduce the system cost, by replacing the A/D converter with simple digital circuits. Among the two methods presented in chapter 4, for voiced/unvoiced detection the zerocrossing based approach is simple and requires very few computations compared to that based on the attractor dimension and entropy. This technique is used in chapter 7 to design an Adaptive Switching Transform Coder.

In chapter 6, we study a modified Adaptive Transform Coder. ATC proposed by Zelinski and Noll is modified by transmitting the maximum value of the transform coefficient as side information for better computation of the step size. Also a modified bit reassignment is applied. If only one bit is available for a coefficient, that coefficient is not coded and transmitted since it will encode only the sign information. Such bits are reassigned for the coefficients in the lower frequency band. The ATC proposed by Zelinski and Noll and the modified one are implemented on a 3AT6 computer. Both DCT and DWHT are used in this study. Performance of the coders are studied in detail for data blocks of different lengths. The SNR performance of DWHT ATC is very much lower than that of DCT ATC for voiced speech. But for unvoiced speech the SNR of DCT ATC is lesser than that of DWHT ATC, especially when the coder is designed for bit rates below 16 kbits/s. This causes tonal distortion in DCT ATC when designed at bit rates below 16 kbits/s. Based on this result we propose an adaptive switching of transform for better speech quality.

The algorithm for the implementation of Adaptive Switching Transform Coder is presented in chapter 7. The scheme is implemented on a 3AT6 computer. This scheme proves useful in improving the speech quality at a bit rate of 8 to 16 kbits/s.

Finally, chapter 8 concludes this work and suggests a few directions for future research.

Chapter 2

REVIEW OF PREVIOUS WORK

2.1 REVIEW OF THE STATE OF THE ART OF WAVEFORM SPEECH CODING

The purpose of this section is to give a summary of the best results obtained in speech waveform encoding. This is a difficult task as different researchers use different fidelity criteria and different speech data. One of the frequently used quality characterizations in speech coding is based on commentary, toll, communication and synthetic categories [J.L.Flanagan et al., 1979]. All these terms are loosely defined in the literature. Toll quality is typically defined as the quality comparable to that of analog speech having bandwidth 200-3200 Hz, a signal-to-noise ratio of 30 dB and less than 2.3% harmonic distortion. For the definition of other qualities see [J.L.Flanagan et al., 1979]. In literature the term "good speech quality" is often considered equivalent to toll quality.

Jayant in his review paper on speech waveform coding compares ADPCM, log-PCM, and Adaptive Delta Modulation

(ADM) using the signal-to-noise ratio (SNR) as criterion [N.S.Jayant, 1974]. The best performance is achieved by ADPCM at a bit rate of 16 kbits/s (i.e., 2 bits/sample). The SNR obtained is around 11 dB. The performance of Pitch-Adaptive DPCM is compared with that of regular ADPCM with a three tap fixed predictor in [N.S.Jayant, 1977]. Both these coders use adaptive quantization. The comparison was done for four utterances, two male and two female, at 16 kbits/s for ADPCM and 17 kbits/s for Pitch Adaptive DPCM. The average results on the four utterances were 11.5 dB for ADPCM and 15.25 dB for Pitch Adaptive DPCM (the performance of the latter system varied from 13.5 dB to 18 dB, the higher performances were obtained on female utterances and the average on male utterances was 14 dB).

Crochiere, Webber and Flanagan studied the relative performance of sub-band speech coding with standard ADPCM at 16 kbits/s and ADM at 9.6 kbits/s [R.E.Crochiere et al., 1976]. At 16 kbits/s, although the SNR was practically the same (11.1 dB for sub-band, 10.9 dB for ADPCM), 94% of listeners preferred sub-band quality. At 9.6 kbits/s the SNR for sub-band encoder was 9.9 dB as compared with 8.2 dB achieved by 10.3 kbits/s ADM.

Flanagan et al. in a review paper on speech coding, present the Adaptive Predictive Coding (APC), Adaptive Transform Coding (ATC), the Phase Vocoder and the Voice Excited Vocoder (VEV) as 'best' systems. All of these systems are claimed to attain "toll quality" at 16 kbits/s and "communication quality" at 7.2 kbits/s [J.L.Flanagan et al., 1979]. All these systems are considered high complexity coders (relative complexity 50 on a scale where ADPCM has relative complexity 1; [J.L.Flanagan et al., 1979]). There are only a few objective signal-to-noise measurements reported for such systems. There are some discrepancies between the available objective measurements and the claim of subjectively achieving 'toll' quality. For example, in [R.V.Cox and R.E.Crochiere, 1981] a segmental SNR of 18.9 dB is reported for an ATC system simulated in real time at 16 kbits/s when the expected value of SNR for a 'toll' quality system would be around 30 dB. Generally, these discrepancies are explained by the fact that the segmental SNR (SEGSNR) does not exactly reflect the subjective quality. We may also note that some of these systems may hardly be considered as waveform coders (for example, VEV uses parameter extraction together with waveform coding).

Another category presented in the above review paper contains Pitch Predictive ADPCM, sub-band coding, and ADPCM. Systems from this category achieve toll quality at 24-32 kbits/s (3-4 bits/sample) and communication quality at 9.6-16 kbits/s (1.2-2 bit/sample). The relative complexity of these systems is in the range 1-5 on the scale where the complexity of ADPCM is 1.

Xydeas, Evci and Steele present the performance of the ADPCM systems in which the predictor coefficients are changed adaptively at each sample instant using gradient techniques [C.S.Xydeas et al., 1982]. They report that these systems perform better than standard ADPCM, but at the expense of complexity. The number of multiplications per sample is of the order of 3n-5n, where n is the number of predictor coefficients. The best achieved performance under the SEGSNR criterion at a rate of 2 bits/sample (16 kbits/s) was 13.5 dB. The SEGSNR is considered to relate better to the subjective quality of speech than the standard SNR. Depending on the system, a difference of around 1-2 dB may be found between SNR and SEGSNR measured on the same system with same speech data.

Anderson and Bodie investigated Tree and Trellis encoding of speech. They used a search algorithm which pursues a fixed number of paths at each level throughout the code tree and a transversal code generator [J.B.Anderson and J.B.Bodie, 1975]. The performance obtained was as high as 21 dB at 2 bits/sample and 12 dB at 1 bit/sample. However, these performances were obtained on a short utterance (2 sec. of speech) and generally were not confirmed by later researchers. Stewart et al. studied the design of tree and trellis speech coders [L.C.Stewart, R.M.Gray and Y.Linde, 1982]. They classified the results as "inside the training sequence" whenever the results were obtained on the speech data used for encoder design and "outside the training sequence" otherwise. The results obtained are 16 dB inside and 13.5 dB outside the training sequence for a rate of 2 bits/sample and 12.2 dB inside and 8.7 dB outside for a rate of 1 bit/sample. Fehn and Noll studied the performances of different tree and trellis encoding schemes at a rate of 1 bit/sample [H.G.Fehn and P.Noll, 1982]. The performances were compared with those of Adaptive Transform Coding (ATC) as reported by R.Zelinski and P.Noll [1977]. The SEGSNR obtained for tree and trellis encoding was 12 dB as compared with 12.3-14.9 dB achieved by ATC (different values for different speakers).

Waveform vector quantization of speech has been studied by several researchers [H.Abut, R.M.Gray and G.Rebollede, 1982], [R.M.Gray and H.Abut, 1982], [V.Cuperman and A.Gersho, 1982], [A.Gersho, T.Ramstad and I.Versvik, 1984]. For a long training sequence (640,000 samples) the signal-to-noise ratio inside the training sequence was found to be 13.5 dB for 2 bits/sample and 9.7 dB for 1 bit/sample. Outside the training sequence the corresponding results were 12.7 dB and 8.8 dB. Using a shorter training sequence (128,000 samples) the performances inside the training sequence increase by about 1 dB, but the performances outside the training sequence decrease by about 0.5 dB.

A comparison of the performance achieved by the known systems is very difficult since different authors use different distortion criteria and different speech data. Also, some researchers consider the 'subjective quality' of speech as the only adequate criterion which is loosely defined and difficult to compare (since the results in a subjective test depend on the training of the team performing the test, on the speech data used etc.).

To summarize, the performance of standard scalar waveform coding systems are limited to a signal-to-noise ratio of about 15 dB at a rate of 2 bits/sample and 10 dB at 1 bit/sample. High complexity encoding schemes such as APC with tree encoding of the residual or ATC may achieve higher performance (18-20 dB at 2 bits/sample and 12 dB at 1 bit/sample). Vector quantization (dimension upto 4 for 2 bits/sample and 8 for 1 bit/sample) achieves a performance of 12-14 dB at 2 bits/sample and 8-9 dB at 1 bit/sample which is lower than the best scalar systems.

2.2 REVIEW OF THE STATE OF THE ART OF THE USE OF ZERO-CROSSINGS IN SPEECH CODING

Licklider in 1946 measured the intelligibility of speech after peak clipping at different levels [J.C.R. Licklider, 1946]. For 0 to 20 dB peak-clipped speech he obtained 96 per cent intelligibility. Between 20 to 50 dB peak-clipped speech, he obtained an intelligibility of about 70 per cent. But the most fascinating result Dr.Licklider obtained was that the intelligibility remained constant at about 70 per cent even with infinite clipping of speech (Infinite clipping is the condition where the signal is rectangular in shape, but crosses the axis at the same portions as the original speech wave).

Licklider and Pollack later demonstrated the results of their investigation upon the perception of speech which had been distorted in various ways [J.C.R.Licklider and I.Pollack, 1948]. They distorted the speech signal by employing differentiation, integration, and infinite clipping. About 100 per cent intelligibility was obtained for speech signal which is distorted by differentiation only, and integration only. Differentiation and clipping, and differentiation, clipping and integration produced about 97 per cent intelligibility, whereas clipping, clipping and integrating, and clipping and differentiating produced about 70 per cent intelligibility. Integrating and clipping, and integrating, clipping and differentiating produced very poor results.

The results demonstrated by Licklider and Pollack have guided the attention of the investigators: 1) to explain this perceptual phenomenon and 2) to utilize this result in practical speech processing systems. These two issues have been studied by many researchers. Morris provides a review of much of this past work and gives new insights into the role of zerocrossings in speech recognition and processing [L.R.Morris, 1970, 1972].

Voelcker demonstrated that a waveform can be completely represented by real and complex zeros [H.B. Voelcker, 1966]. Haavik showed that repeated differentiation converts complex zeros into real zeros (zerocrossings)

[S.Haavik, 1966]. Later on Morris demonstrated that Voelcker's and Haavik's theories can be used to explain the high intelligibility of clipped speech [L.R.Morris, 1970, 1972]. This work provides a mathematical and theoretical explanation of the high intelligibility of clipped and clipped differentiated speech.

Licklider also performed a study which provided data about the number of bits per second necessary to represent a clipped differentiated speech signal [J.C.R. Licklider, 1950]. In this study, clipped differentiated speech was time quantized over a range from about 1000 to 40,000 quanta/s. In viewing this experiment in modern technological terms, it can be considered as sampling the clipped differentiated waveform at continuous rates from 1000 to 40,000 samples/s. Since each A/D sample must be either a "0" or "1", this can be considered in representing the clipped speech with 1000 to 40,000 bits/s. A conclusion that can be arrived at from Licklider's results is that clipped differentiated speech sampled with a 1 bit A/D converter is approximately 90 per cent intelligible at bit rates as low as 9 kbits/s.

In a recent paper Niederjohn, Krutz and Brown present the results of an experimental investigation that provides an interesting perspective on the relative importance of zerocrossing locations and zerocrossing intervals for speech perception [R.J.Niederjohn et al., 1987]. In the experiments reported, the intelligibility degradation of clipped filtered speech subjected to averaging and reordering of the zerocrossing interval sequence was studied. The main conclusion drawn from this study is that the set of zerocrossings of a speech waveform represents a near minimal set of informational attributes in the sense that any reordering or averaging of the zerocrossing intervals has a significant detrimental effect upon speech intelligibility.

Kay and Sudhakar studied a zerocrossing based spectrum analyzer in [S.M.Kay and R.Sudhakar, 1986]. A reconstruction method suitable for the recovery of a low frequency noisy sinusoid from its zerocrossings is presented in this paper. It is also suitable for periodic band-limited signal, which can be recovered within a scale factor from its zerocrossing sequence, using a sine wave product expansion formula. In the case of a periodic signal which is represented by an infinite product, the signal can only be approximately recovered.

To summarize, the infinitely clipped speech signals still retain all the important aspects of intelligibility and even to some extent recognizability. This is possible only if the information in the zerocrossing can be extrapolated to yield the original speech signal. Therefore further investigations are necessary to reconstruct the speech signal from their zerocrossings.

2.3 MEASUREMENT OF SPEECH QUALITY

Assessment of the relative performance of speech coders is one of the most difficult tasks in speech coding. It is not completely understood how the human ear and the brain process the speech signal. Because of this it has not been possible to quantify in a mathematical expression what is meant by the words "speech quality". However, based on the present day knowledge of speech understanding many performance measurement methods have evolved. These are mainly classified into two: objective measurement using mathematical expressions and subjective measures by listening tests. We will describe the objective measures like signal-to-noise ratio (SNR), articulation index, log spectral distance, Itakura's likelihood ratio, and Euclidean distance and subjective tests like Diagnostic Rhyme Test (DRT), the

Modified Rhyme Test (MRT), the Diagnostic Acceptability Measure (DAM), and Mean Opinion Score (MOS) in this section.

2.3.1 Objective measures of speech quality

Speech coding systems are mainly classified into two. The first one, called waveform coders, tries to preserve the shape of the speech waveform. The second one, usually called vocoders, is not concerned with the exact shape of the waveform but the resulting speech sounds like the original with a synthetic quality. In order to suit these distinct types of coders we use two types of objective measures. The signal-to-noise ratio (SNR) - related measures are better suited for waveform coders, while spectral distance measures are better suited for vocoders.

2.3.1.1 Signal-to-noise ratio (SNR)

Signal-to-noise ratio (SNR) is the most commonly used objective measure in waveform coders. Let X(n) be the original speech signal and Y(n) the corresponding coded signal at the sampling instant n. Then the coding error signal is given by

$$e(n) = X(n) - Y(n) \qquad (2.1)$$

To find the SNR for a record of N samples, we compute the original signal variance as,

$$E_s^2 = \frac{1}{N} \sum_{n=1}^{N} [X(n) - \frac{1}{N} \sum_{n=1}^{N} X(n)]^2 \qquad (2.2)$$

and the error signal variance as

$$E_e^2 = \frac{1}{N} \sum_{n=1}^{N} [e(n) - \frac{1}{N} \sum_{n=1}^{N} e(n)]^2 \qquad (2.3)$$

The SNR is defined as the ratio of the original signal variance to error signal variance,

i.e., $$SNR = \frac{E_s^2}{E_e^2} \qquad (2.4)$$

It is expressed in dB as

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

The speech signal is characterized by its time varying nature. This results in some speech segments with high energy and other segments with low energy. If the error variance E_e^2 is more or less constant, the resulting SNR will be high. But the perceptual effects of the noise in the regions of lower E_s^2 will be more severe. To take into account this fact, the performance of a coding system is measured in terms of segmental SNR which is denoted as SEGSNR. To compute the SEGSNR, we divide the speech signal into segments of 64-256 sample length, and compute the SNR(m) dB where m = 1,2,...,M represents the block number. Then the segmental SNR is defined by

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

By averaging the SNRs of different segments as in (2.6) the strong portions of the signal do not overwhelm the SNR.

2.3.1.2 Articulation index

The articulation index (AI), originally used with analog signals, is a method of assessing the speech quality. To compute the AI, the signal is bandpass filtered into 20 bands as in [N.S.Jayant and P.Noll, 1984]. For each band m, we compute the signal-to-noise ratio SNR(m) dB, and from that we can obtain the articulation index

$$\text{AI} = \frac{1}{20} \sum_{m=1}^{20} \frac{\min(\text{SNR}(m)\text{dB},\ 30)}{30} \qquad (2.7)$$

The SNR value for each band is limited to a maximum of 30 dB. The articulation index is analogous to SEGSNR, except that the segmentation takes place in the frequency domain instead of the time domain.

2.3.1.3 Log spectral distance

In the case of vocoders, only the magnitude of the spectrum of speech is usually preserved. This is according to the hypothesis that the human ear is not very sensitive to the short-term phase. As a result, the vocoder output waveform can be quite different from the original speech, and still sound the same. Therefore it is no longer meaningful to use the signal-to-noise ratio as a measure of reproduction fidelity of vocoder outputs. Here, we have to use distance metrics that are sensitive to spectral differences. Log-spectral distance, Itakura's likelihood ratio and the Euclidean distances are such metrics.

The log spectral distance measures for vocoders are often used in the context of the LPC method. In LPC, the spectral envelope representing the vocal tract is given by the expression

$$H(e^{j\omega}) = \frac{G}{A(e^{j\omega})} \tag{2.8}$$

where the inverse filter A(z) is given by

$$A(z) = 1 + a_1 z^{-1} + \ldots + a_p z^{-p} \tag{2.9}$$

Usually, p = 10. The log-spectral distance between two LPC models $H_1(e^{j\omega})$ and $H_2(e^{j\omega})$ is defined by

$$d = [\int_{-\pi}^{\pi} |\ln|H_1(e^{j\omega})|^2 - \ln|H_2(e^{j\omega})|^2|^2 \frac{d\theta}{2\pi}]^{\frac{1}{2}} \tag{2.10}$$

The log-spectral distance is a reasonable measure to use for the determination of quality if we assume that one of the spectra H_1 is the true representation of the speech signal, while the other is an approximation whose goodness we are testing.

2.3.1.4 Itakura's likelihood ratio

Let a_1 be the coefficient vector and R_1 be the autocorrelation matrix for the LPC vocal tract model H_1. Also, let a_2, R_2 be the corresponding quantities for H_2. Then, the likelihood ratio can be defined either as

$$d_{LR_1} = \frac{a_2^T R_1 a_2}{a_1^T R_1 a_1} \tag{2.11}$$

or

$$d_{LR_2} = \frac{a_1^T R_2 a_1}{a_2^T R_2 a_2} \qquad (2.12)$$

The log likelihood ratios are the logarithms of these expressions:

$$d_{LLR} = 10 \log_{10}(d_{LR}) \qquad (2.13)$$

and are expressed in dB.

2.3.1.5 Euclidean distance

The Euclidean distance metric is defined as

$$d_{LAR} = [\sum_{i=1}^{p} (LAR_{1i} - LAR_{2i})^2]^{\frac{1}{2}} \qquad (2.14)$$

Here LAR_{1i} and LAR_{2i} are the sets of log-area ratios corresponding to H_1 and H_2. The log-area ratios, LAR_i are derived from the reflection coefficients K_i according to the relation

$$LAR_i = \ln \frac{1+K_i}{1-K_i} \qquad (2.15)$$

2.3.2 Subjective measures

Subjective tests are performed by listening to the coded speech. They are divided into two basic categories:

one testing the intelligibility and the other testing the quality of the coded speech (good intelligibility does not necessarily mean good quality but the converse is true). In the following we will discuss both categories.

2.3.2.1 Intelligibility tests

Diagnostic Rhyme Test (DRT) is the most widely used intelligibility test in speech coding. In this scheme, the listener is presented with one word from a pair of encoded words differing only in one phoneme and asked to determine what word was spoken. A correct response from the listener indicates that the coded speech is intelligible. The DRT score in per cent is given by

$$P = \frac{R-W}{T}100 \tag{2.16}$$

where

R = number of right answers

W = number of wrong answers

T = total number of items involved.

Typical values of DRT range between 75 and 95. A "good" system must have a DRT score of about 90.

In DRT the 'word pair' are so chosen that they differ only in one attribute of the first consonant. That is, DRT is based on differences of initial consonants only, and the listener is asked to select among pairs of words.

The Modified Rhyme Test (MRT) is another intelligibility test. In this, the listener is presented with one encoded word, and is asked to select his answer from a list of six words rather than two words in DRT. Also both groups of words that differ in beginning consonant and ending consonant are used in MRT.

2.3.2.2 Quality tests

Subjective judging the quality of encoded speech is an extremely difficult task. This is because different speech encoder systems introduce different types of distortion, and different people have different preferences. It is also probable that these preferences change over time. These limitations should be kept in mind when considering the different kinds of quality tests.

Diagnostic Acceptability Measure (DAM) is a highly systematic approach to determining the speech quality. This requires well trained listener crews who are able to determine any drift in the individual performance.

In DAM, encoded sentences are taken from the Harvard list of phonetically balanced sentences (e.g., "Add the sum to the product of these three", "An icy wind racked the beach" etc.). The listener is presented with these sentences and asked to rate the speech quality both in terms of overall acceptability and in terms of the individual characteristics (parametrically). The listener is asked to evaluate the Hissing, Buzzing, Babbling, Rumbling etc. characteristics of the encoded speech, by giving a grade between 0 and 100 to each characteristics. Finally the overall quality is judged by evaluating the grade given to each characteristic.

Mean Opinion Score (MOS) is another subjective quality test. In MOS, the listener is asked to rate a system on an absolute scale, usually ranging between 1 and 5. The meaning of these grades are:

5. excellent
4. good
3. fair
2. poor
1. bad.

The mean value of the grades rated by a number of listeners is taken as the MOS.

To summarize, this section reviewed the objective and subjective measures that are used to evaluate a speech coding system. Among these the objective measure based on the signal-to-noise ratio (SNR) is used in this thesis because of its simplicity in implementation.

Chapter 3

SPEECH SAMPLE ESTIMATION FROM ITS COMPOSITE ZEROCROSSINGS

3.1 INTRODUCTION

In digital speech processing, a continuous analog speech signal is to be converted to a discrete form suitable for processing in a digital computer. This involves sampling in the time domain, quantization in the amplitude domain and coding the result into digital form. All these processes impose limitations on the speech data obtained, and can give rise to various errors. In a conventional speech processing system the above mentioned processing is done using an A/D converter. In this chapter we present a new method to estimate the speech samples from zerocrossings.

A band limited signal can be represented by the real and complex zeros of the signal [F.E.Bond and C.R.Cahn, 1958], [H.B.Voelcker, 1966], [A.A.G.Requicha, 1980]. The zerocrossing based approach requires no sampling of the signal, while the A/D converter method relies on multilevel quantization of samples taken at prescribed instants of time. With simple digital circuits, it is easier to measure the timings of zerocrossings. Some potential advantages of this approach are that the inherent problems of an A/D converter such as alignment, limited dynamic range and speed can be alleviated, and the hardware associated with the sampling circuitry can also be simplified.

Kay and Sudhakar [ASSP, 1986] proposed a method suitable for the recovery of a low frequency noisy sinusoid from its zerocrossings. This is also suitable for periodic band limited signal, which can be recovered within a scale factor from its zerocrossing sequence, using a sine wave product expansion formula [A.Seeky, 1970]. For aperiodic signals which are represented by an infinite product, the signal can only be approximately recovered.

The zerocrossings of a segment of speech are the result of a nonlinear operation, and hence analysis is extremely complicated. Further, the statistical aspects have not been investigated completely though C.S.Sridhar attempted to study this problem from statistical angle [C.S.Sridhar, 1975]. Recently speech segments are modelled using the information regarding zerocrossings and it is shown that this helps in extracting features like voiced/unvoiced boundary [Babu P.Anto, N.K.Narayanan and C.S.Sridhar, 1987].

In this chapter a simple new method is presented for the reconstruction of speech signals from their composite zerocrossings. To begin with, the theory associated with zerocrossing sampling method used to estimate the speech signal is presented. A simple linear interpolation formula for signal estimation from composite zerocrossings is derived. This is followed by the results of a computer simulation experiment on the recovery of speech signals from zerocrossings.

3.2 ZEROCROSSING SAMPLING THEORY

Consider a bandlimited signal S(t) with the highest frequency value W_n. A cosine wave of frequency equal to the highest frequency W_n, present in the signal S(t) and of a larger amplitude than the maximum amplitude of S(t), is added to S(t) so that a composite signal X(t) is obtained as

$$X(t) = S(t) + A\cos(2\pi W_n t) \qquad (3.1)$$

where $A > |S(t)|_{max}$.

It can be shown that $X(t)$ has exactly one zerocrossing in each of the sub-intervals of duration $(1/2W_n) = T_S$, i.e., if the signal is sampled at the Nyquist rate, then between adjacent samples there will be one zerocrossing.

From equation (3.1) we can write,

$$X(0) = S(0) + A > 0,$$

i.e., $X(0) > 0$

and $X(T_s) = S(T_s) + A \cos (2\pi W_n T_s)$

$$= S(T_s) - A < 0. \quad \text{i.e., } X(T_s) < 0.$$

When the composite signal $X(t)$ is sampled at its Nyquist rate, $X(0)$ and $X(T_s)$ are two adjacent sample values. Since $X(t)$ changes sign, there must be an odd number of zerocrossings of $X(t)$ in the interval $(0, T_s)$. Similarly, it can be shown that between any two consecutive samples there will be an odd number of zerocrossings. Since the signal $X(t)$ is bandlimited to W_n, the maximum number of zerocrossings in a duration of unit second is $2W_n$. Remember the composite signal $X(t)$ is sampled at its Nyquist rate, i.e., sampling frequency $f_s = 2W_n$. Therefore the above conditions can be satisfied only if there is only one zerocrossing between adjacent samples of X(t). Hence in a sequence of N samples $\{X(kT_s)\}$ there will be (N-1) zerocrossings. Let t_o, t_1, $t_2, \ldots, t_{N-2}$ be these zerocrossings. Now we can derive a linear interpolation formula for estimating the signal samples using these zerocrossings.

According to the sampling theorem, for any band-limited signal which is sampled at the Nyquist rate, the reconstruction formula that provides perfect reconstruction of the signal is defined by the infinite sum of weighted sample values as

$$X(t) = \sum_{k=-\infty}^{\infty} X(kT_s) \frac{\text{Sin}[\pi(t-kT_s)/T_s]}{\pi(t-kT_s)/T_s} \tag{3.2}$$

For an N periodic sequence the above formula reduces to

$$X(t) = \sum_{k=0}^{N-1} X(kT_s) \frac{\text{Sin}[\pi(t-kT_s)/T_s]}{\pi(t-kT_s)/T_s} \tag{3.3}$$

Now the zerocrossing time, $t_o, t_1, t_2 \ldots$ can be obtained by substituting X(t) = 0, in the above formula. Therefore, the interpolation formula (3.3) will give (N-1) equations representing the N-1 zerocrossings as

$$\sum_{k=0}^{N-1} X(kT_s) \frac{\mathrm{Sin}[\pi(t_n-kT_s)/T_s]}{\pi(t_n-kT_s)/T_s} = 0 \qquad (3.4)$$

where n = 0,1,2,...N-2.

To obtain the values of the zerocrossing times $t_o, t_1, t_2, \ldots$, we have to solve N-1 equations. For simplicity we can assume that the interpolation between two adjacent samples is mainly contributed by these two samples only, so that the terms with other sample values in (3.4) can be neglected. Now the interpolation formula for the signal between $X(nT_s)$ and $X((n+1)T_s)$ becomes

$$\sum_{k=n}^{n+1} X(kT_s) \frac{\mathrm{Sin}[\pi(t_n-kT_s)/T_s]}{\pi(t_n-kT_s)/T_s} = 0 \qquad (3.5)$$

Simplifying this equation we can write

$$\frac{t_n-nT_s}{T_s} = \frac{X(n)}{X(n)-X(n+1)}$$

Considering the alternate sign changes in signal amplitude, we can write

$$t_n = \frac{T_s|X(n)|}{|X(n)|+|X(n+1)|} + nT_s \tag{3.6}$$

Equation (3.6) gives a relationship between the n^{th} zero-crossing time and the n^{th} and $(n+1)^{th}$ sample magnitudes. Putting $t'_n = t_n - nT_s$ in equation (3.6) and rearranging we get

$$|X(n+1)| = |X(n)|\frac{T_s-t'_n}{t'_n} \tag{3.7}$$

Using (3.7), assuming a suitable scale factor for X(0), say unity, the consecutive signal magnitudes |X(1)|, |X(2)|,..., can be calculated using the information regarding zerocrossing time. Formula (3.7) can be obtained by a simple triangular interpolation method as discussed below.

Consider two consecutive sample values X(n) and X(n+1) of the composite signal. Let t'_n be the zerocrossing distance of the n^{th} zerocrossing from the n^{th} sample, so that

$$t_n = t'_n + nT_s.$$

In the Fig.3.1, |X(n)| and |X(n+1)| represent the sample amplitudes. The point at which the line joining the edges of the sample amplitude crosses the time axis is taken as the zerocrossing.

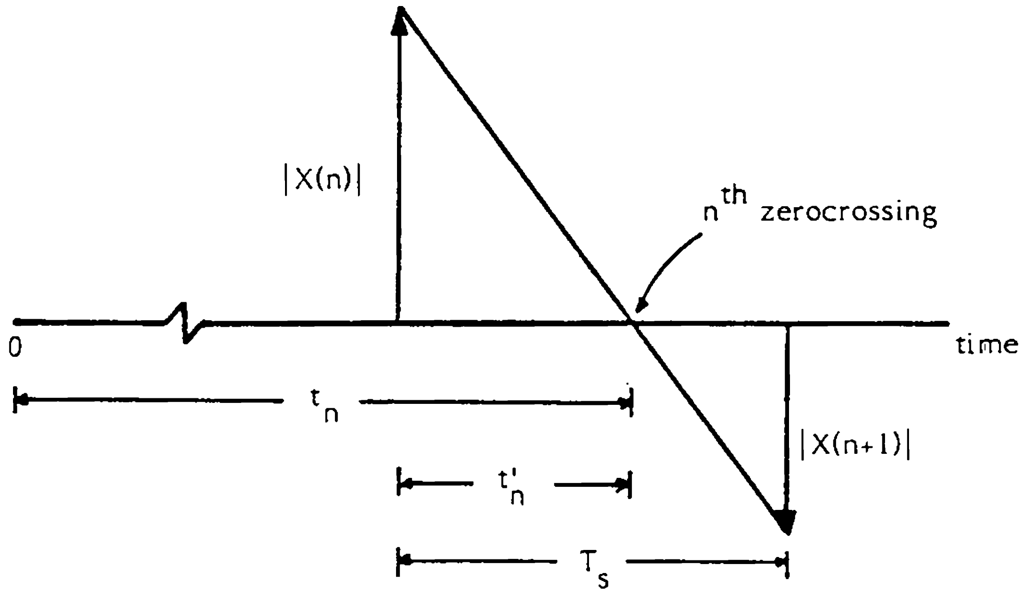


Fig. 3.1 Triangular Interpolation

From the figure 3.1, using the theorem of similar triangles we can write

$$\frac{T_s - t'_n}{t'_n} = \frac{|X(n+1)|}{|X(n)|}$$

$$\therefore \quad |X(n+1)| = |X(n)| \frac{T_s - t'_n}{t'_n} \qquad (3.8)$$

Equations (3.7) and (3.8) are the same. The sequence of sample amplitude obtained by this formula, after multiplying the alternate sample by -1 gives the reconstructed version of the sample values of the composite signal X(t). Therefore we can write a triangular interpolation formula (TIF) for the reconstruction of samples using Nyquist rate zerocrossing as

$$X(n+1) = (-1)^{n+1} |X(n)| \frac{T_s - t'_n}{t'_n} \qquad (3.9)$$

where n = 0,1,2,...N-2.

The true signal can be obtained from this by subtracting the added sinusoid. Let X(n) and S(n), where n = 0,1,2,...N-1 be the Nyquist rate samples of the composite signal and the original signal respectively, then we can write $S(n) = X(n) - A \cos(2\pi W_n T_s)$, where A is the amplitude of the added sinusoid.

3.3 SIMULATION EXPERIMENT AND RESULTS

Fig.3.2 shows the schematic representation of the method of the reconstruction of speech samples from the zerocrossings. Speech signal is lowpass filtered to bandlimit to 4 kHz, and is normalized, so that the maximum amplitude of the speech signal under study is unity. This is summed with a sinusoid whose frequency is half of the required sampling rate f_s (here 4 kHz). The amplitude of the sinusoid is taken as twice that of the maximum amplitude of speech signal. The composite signal X(t) is passed through the zerocrossing location extractor to find the zerocrossing location and is then quantized by the zerocrossing location quantizer. The quantized zerocrossing locations (ZCL) are used to reconstruct the composite signal samples $\{X(n)\}$ using the TIF (3.9). This sequence $\{X(n)\}$ after subtracting the amplitude of the added sinusoid (in the present case 2) gives the reconstructed version of the speech sample sequence $\{S(n)\}$.

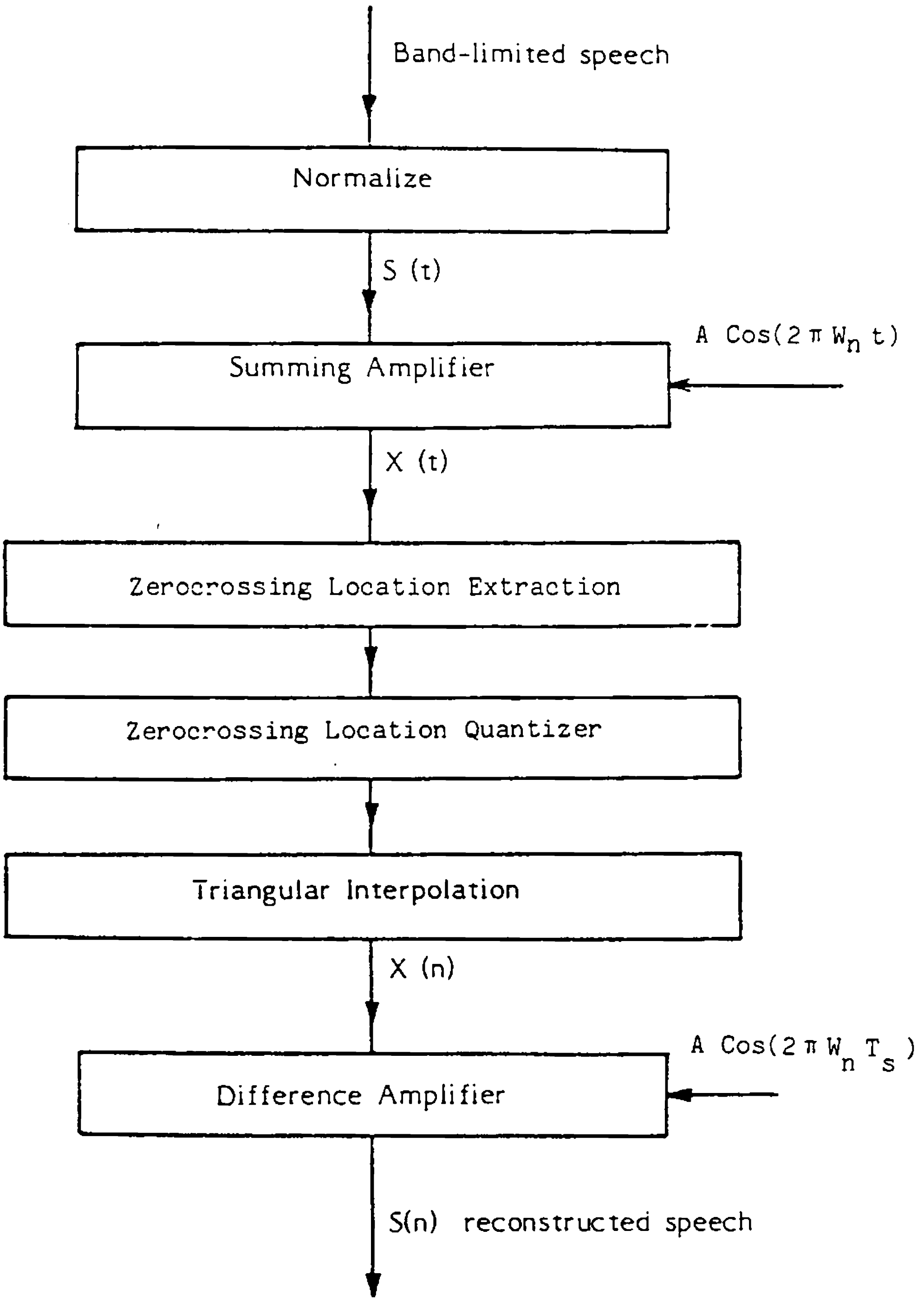


Fig. 3.2 Schematic block diagram for the method of reconstruction of the speech samples from the zerocrossings.

3.3.1 Verification of the Zerocrossing Theory

Simulation experiments are conducted to verify the triangular interpolation method developed in the previous section. The speech data base used in this experiment contains 45 sec. of speech spoken by two speakers; a male and a female. The data base consists the following speech material denoted as S1, S2,..., S9.

S1 : An icy wind racked the beach.

S2 : The pipe began to rust while new.

S3 : Cats and dogs hate each the other.

S4 : Oak is strong and also gives shade.

S5 : Thieves who rob friends deserve jail.

S6 : Open the crate but do not break the glass.

S7 : Add the sum to the product of these three.

S8 : Joe brought a young girl.

S9 : A lathe is a big tool.

These utterances were chosen since they are phonetically well balanced, including voiced speech, plosives, fricatives etc. The actual zerocrossing location are determined by sampling the speech waveform at a very high rate than the Nyquist rate as in [J.C.R.Licklider,1950] and [R.J.Niederjohn et al., 1987]. The speech waveform, band limited to 4 kHz and digitized using a 12 bit A/D converter at a sampling rate of 64 kHz is stored in the data base of the computer.

Now the samples of the sinusoid of frequency 4 kHz at the same sampling rate of speech, i.e., 64 kHz are generated in software using the formula

$$A \cos \left(\frac{2 \pi \times 4000 \times k}{64000} \right)$$

where k = 0,1,2,...etc. The amplitude of the sinusoid, 'A' is chosen to be equal to 2. By adding the normalized speech sample and the sinusoid sample we obtained the samples of the composite signal at a sampling rate of 64 kHz. Since the amplitude of the sinusoid is chosen to be equal to 2, the maximum amplitude of the composite signal is equal to 3 times that of original speech signal. Fig.3.3(a) represents a segment of the speech signal. Fig.3.3(b) represents the 4 kHz sinusoid that is to be added with the speech segment in Fig.3.3(a). Fig.3.3(c) gives the resultant composite signal. Fig.3.4(a), (b) and (c) represents the 64 kHz sampled form of the signal in 3.3(a), (b) and (c) respectivel

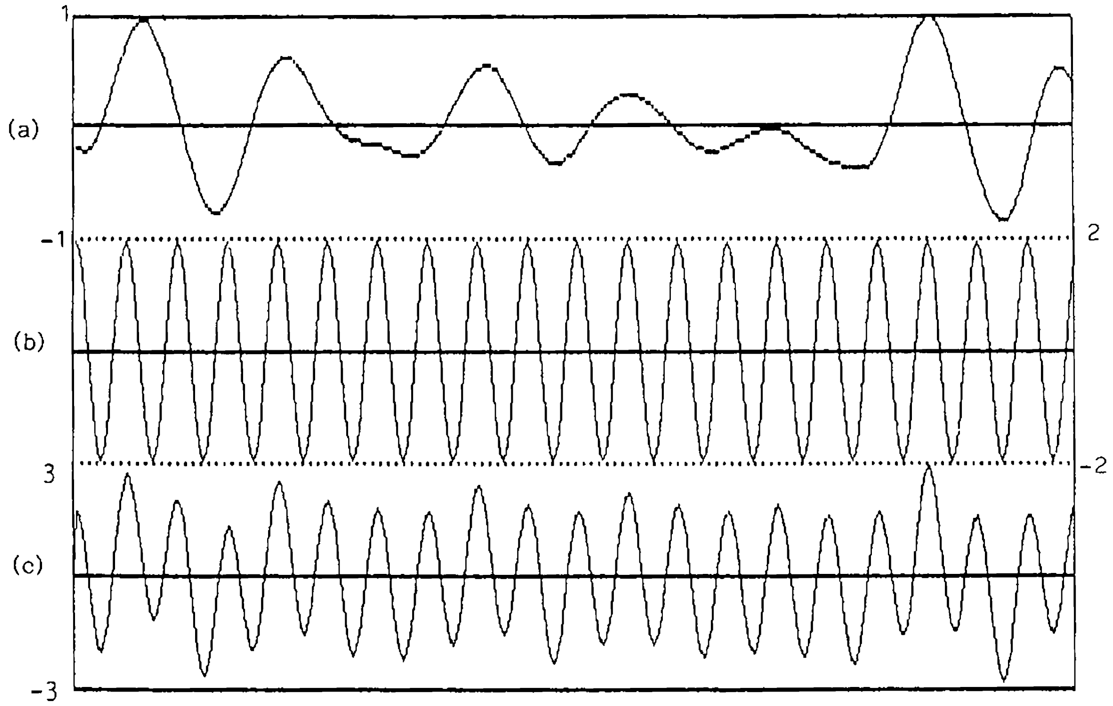


Fig.3.3(a) Original speech segment
(b) 4 kHz sinusoid
(c) Composite Signal (speech + sinusoid)

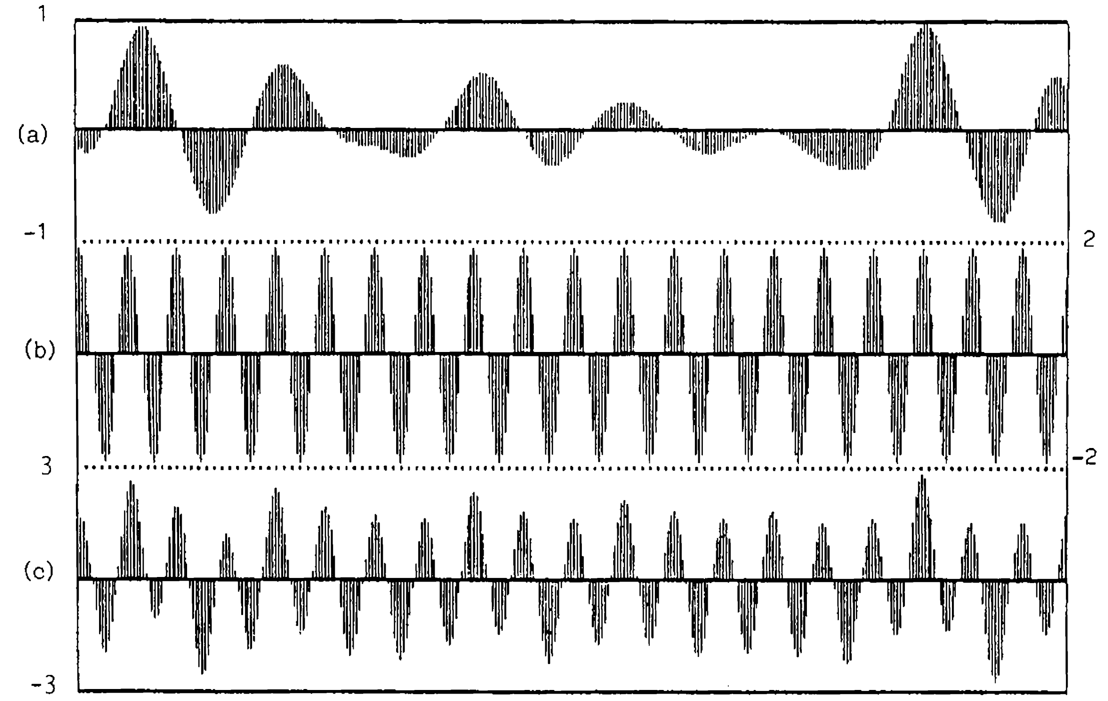


Fig.3.4(a) 64 kHz sampled form of original speech
(b) 64 kHz sampled form of 4 kHz sinusoid
(c) 64 kHz sampled form of composite signal

Fig.3.5 illustrates the zerocrossing locations in the composite signal. The zerocrossing locations of the band limited composite signal are determined by linear interpolation between each successive pair of samples which differ in sign (remember these sample pairs are not the Nyquist rate samples of the composite signal that differ in sign, which are to be estimated using the triangular interpolation formula. The Nyquist rate samples of the original composite signal are shown in Fig.3.5 with arrow heads). The linear interpolation is carried out to achieve a greater accuracy in zerocrossing interval measurement than that would be accomplished by simply choosing either the sample location nearest to each zerocrossing, or the middle point between two samples. Thus the original zerocrossing locations (OZCL) are determined within an error limit of 7.8125 micro second in the present experiment.

The zerocrossing locations shown in Fig.3.1 (hereafter referred as pseudo zerocrossing locations (PZCL)) are obtained by linear interpolation between adjacent Nyquist rate samples of the composite signal. This is done for the sake of comparison between OZCL and PZCL. The OZCL, PZCL and their difference, for a speech segment

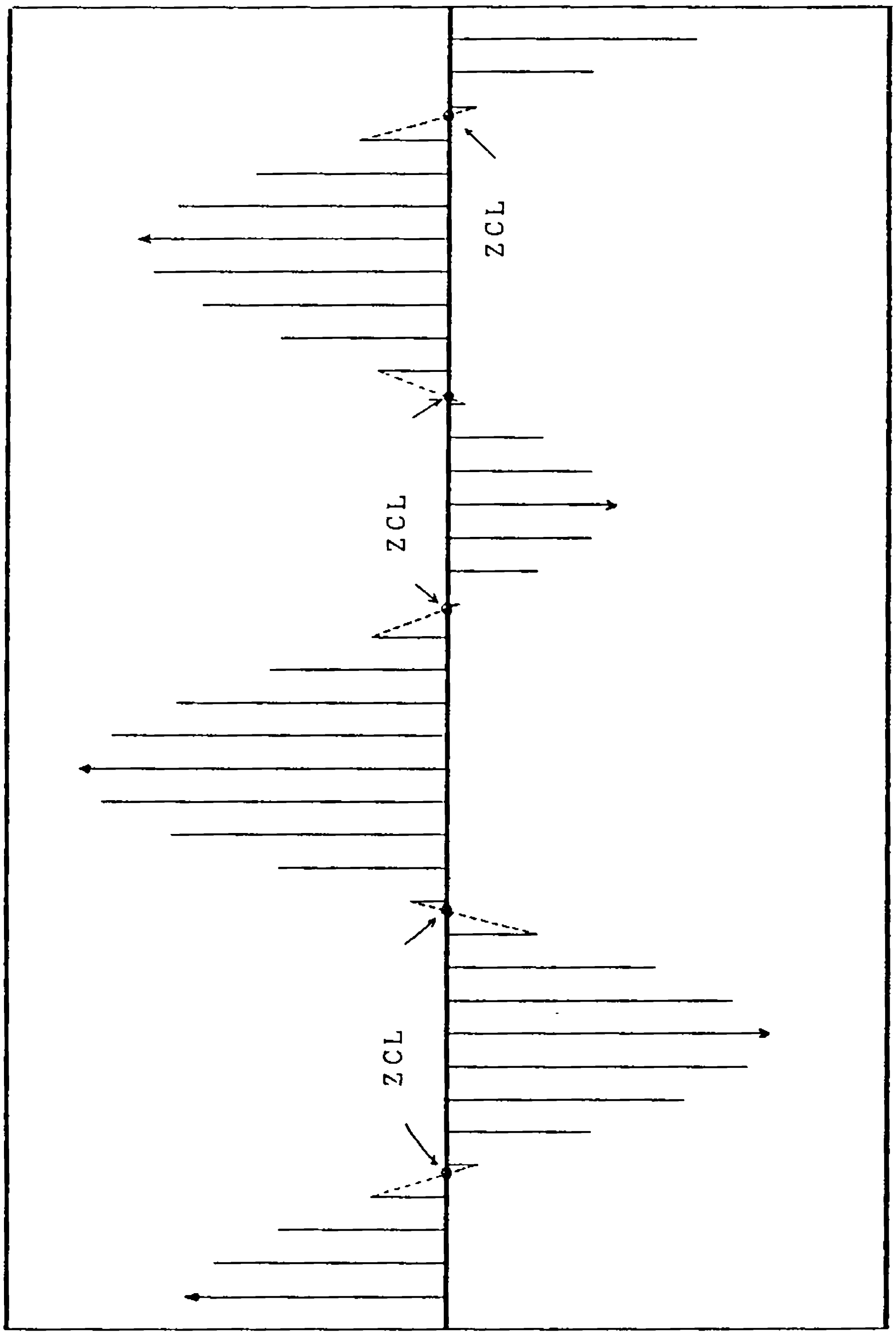


Fig.3.5 Zerocrossing Locations (ZCL) in the composite signal

are listed in Table 3.1. The maximum difference is about 6 micro seconds. These are illustrated in the Fig.3.6.

Since the OZCL differ from the PZCL the estimated sample values will suffer two types of errors, viz., error due to interpolation and error due to ZCL quantization.

The estimated OZCL are quantized to different number of bits. The Nyquist rate composite signal sequence is estimated using the TIF with the quantized OZCL, by assuming the amplitude of the first sample value as unity. This sequence is then scaled to make the maximum amplitude value as 3 which is the maximum amplitude value of the original composite signal. By subtracting the amplitude of the added sinusoid from the reconstructed composite signal, the reconstructed speech signal sequence $\{S(n)\}$ is obtained. Figs.3.7(al,a2,...a7) represents the segments of original speech signal and Figs.3.7(bl,b2,...,b7) represents the reconstructed speech signal using OZCL quantized to 8 bits. It may be noted that the reconstructed speech signal is distorted with high frequency noise. This is studied by computing the power spectral densities (PSD) of the original speech segment and the reconstructed one

Table 3.1

Occurrence of zerocrossings from the last NYQUIST rate sample in micro seco

ZCL	original	psuedo	difference	ZCL	original	psuedo	difference
No	ZCL time	ZCL time	between	No	ZCL time	ZCL time	between
	(OZCL)	(PZCL)	OZCL&PZCL		(OZCL)	(PZCL)	OZCL&PZCL
1	67.53	70.13	-2.60	21	60.15	59.14	1.01
2	61.06	59.27	1.79	22	65.84	67.34	-1.50
3	62.36	62.16	0.20	23	59.16	57.20	1.96
4	63.94	64.27	-0.33	24	65.61	67.55	-1.94
5	59.89	58.08	1.81	25	58.72	58.16	0.56
6	66.25	67.24	-1.00	26	61.53	59.37	2.16
7	62.08	62.91	-0.82	27	71.44	77.70	-6.25
8	57.12	55.66	1.46	28	49.35	48.25	1.10
9	68.60	71.04	-2.45	29	62.73	63.26	-0.52
10	60.56	59.37	1.19	30	74.16	75.57	-1.42
11	59.90	59.21	0.69	31	53.98	49.66	4.32
12	65.58	66.46	-0.88	32	61.64	60.72	0.92
13	62.30	61.85	0.45	33	70.60	72.95	-2.34
14	60.36	59.72	0.64	34	56.21	54.44	1.77
15	65.02	66.13	-1.12	35	62.63	64.44	-1.81
16	61.15	60.82	0.32	36	63.27	64.03	-0.76
17	61.51	61.21	0.29	37	60.67	59.57	1.11
18	64.82	65.63	-0.80	38	66.79	68.09	-1.30
19	60.71	59.56	1.15	39	59.01	57.26	1.75
20	63.87	64.96	-1.09	40	61.06	59.95	1.11

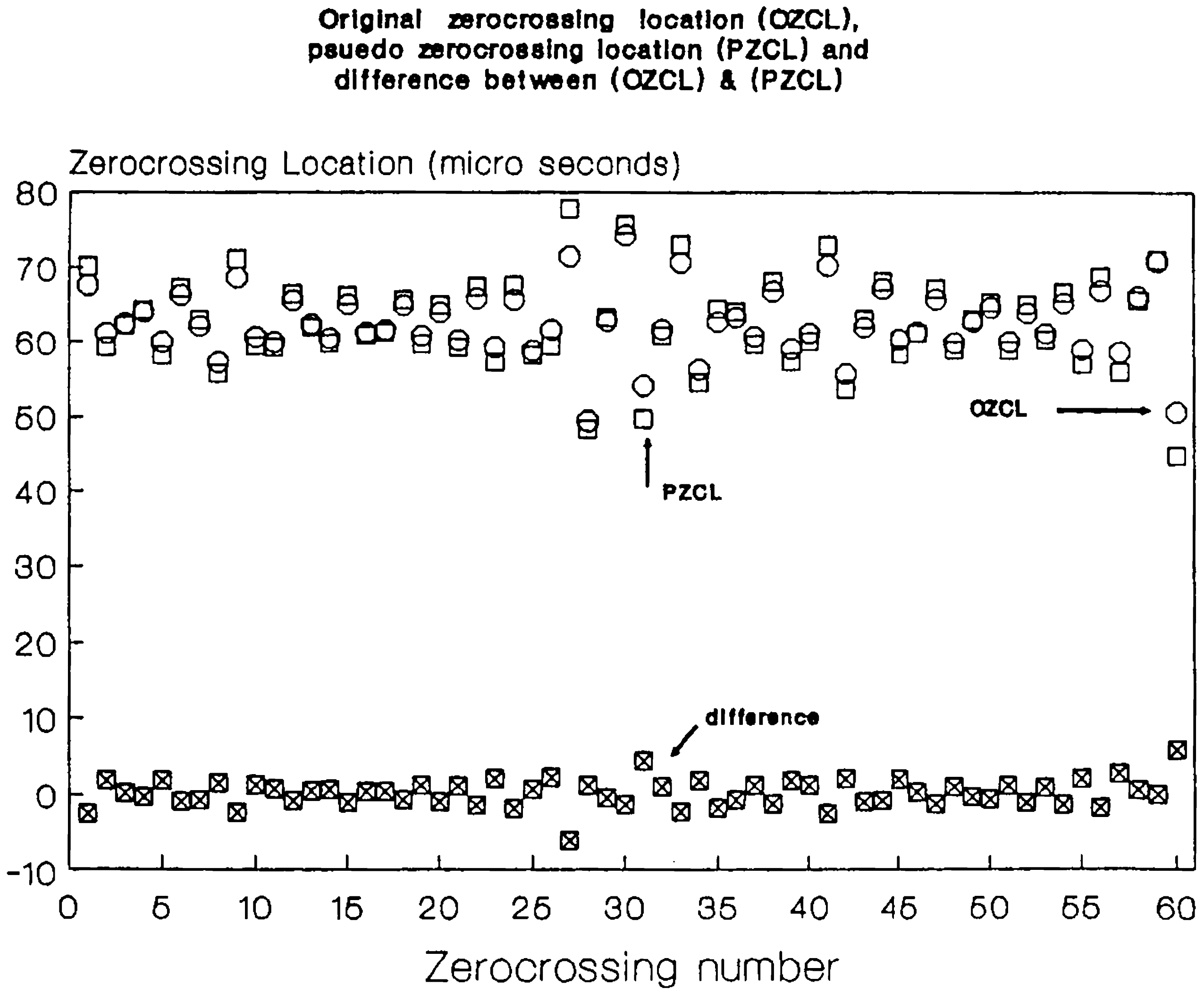


Fig 3.6 Comparison between OZCL & PZCL

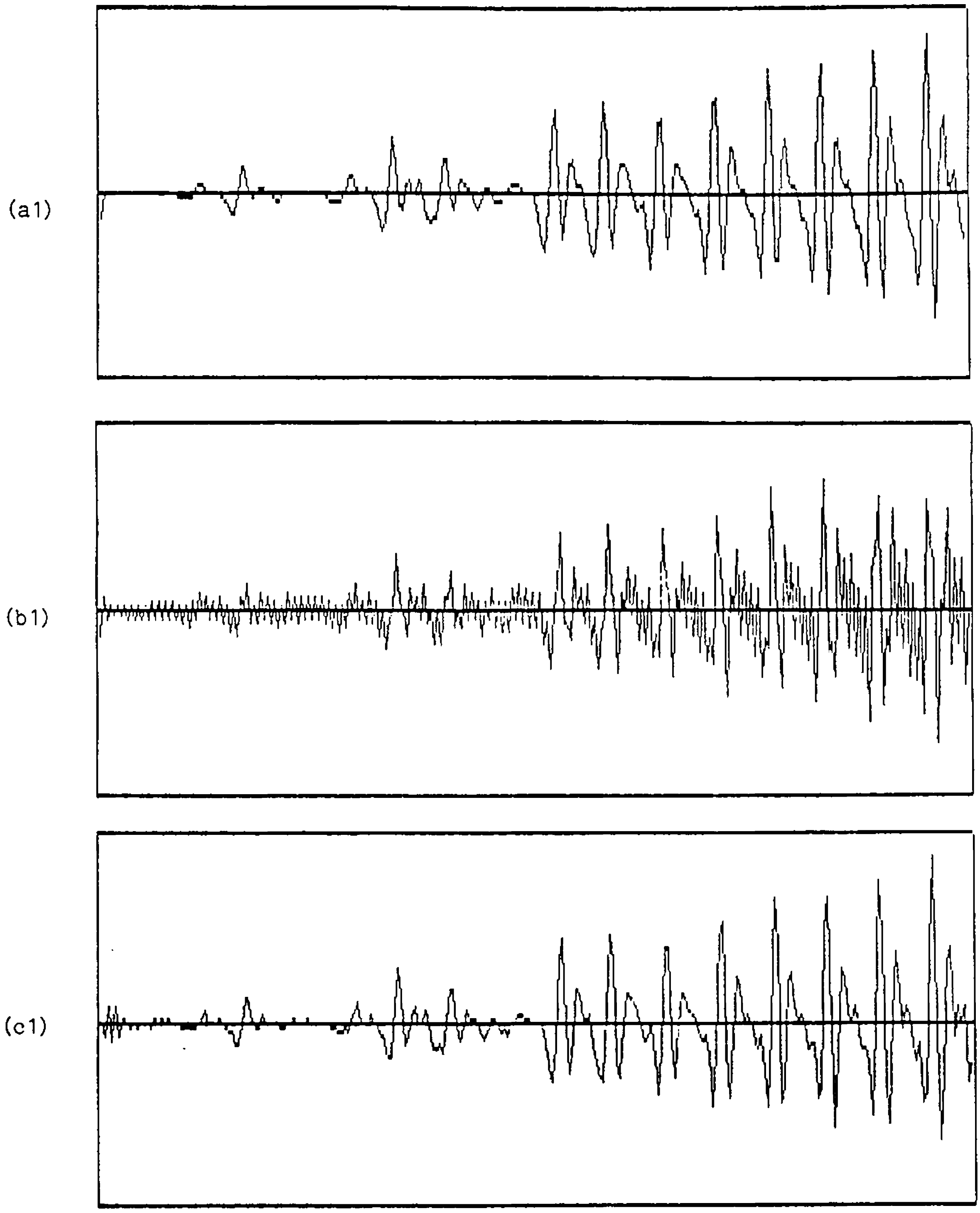


Fig. 3.7 (a1) original speech
(b1) reconstructed speech (unfiltered)
(c1) reconstructed speech (filtered)

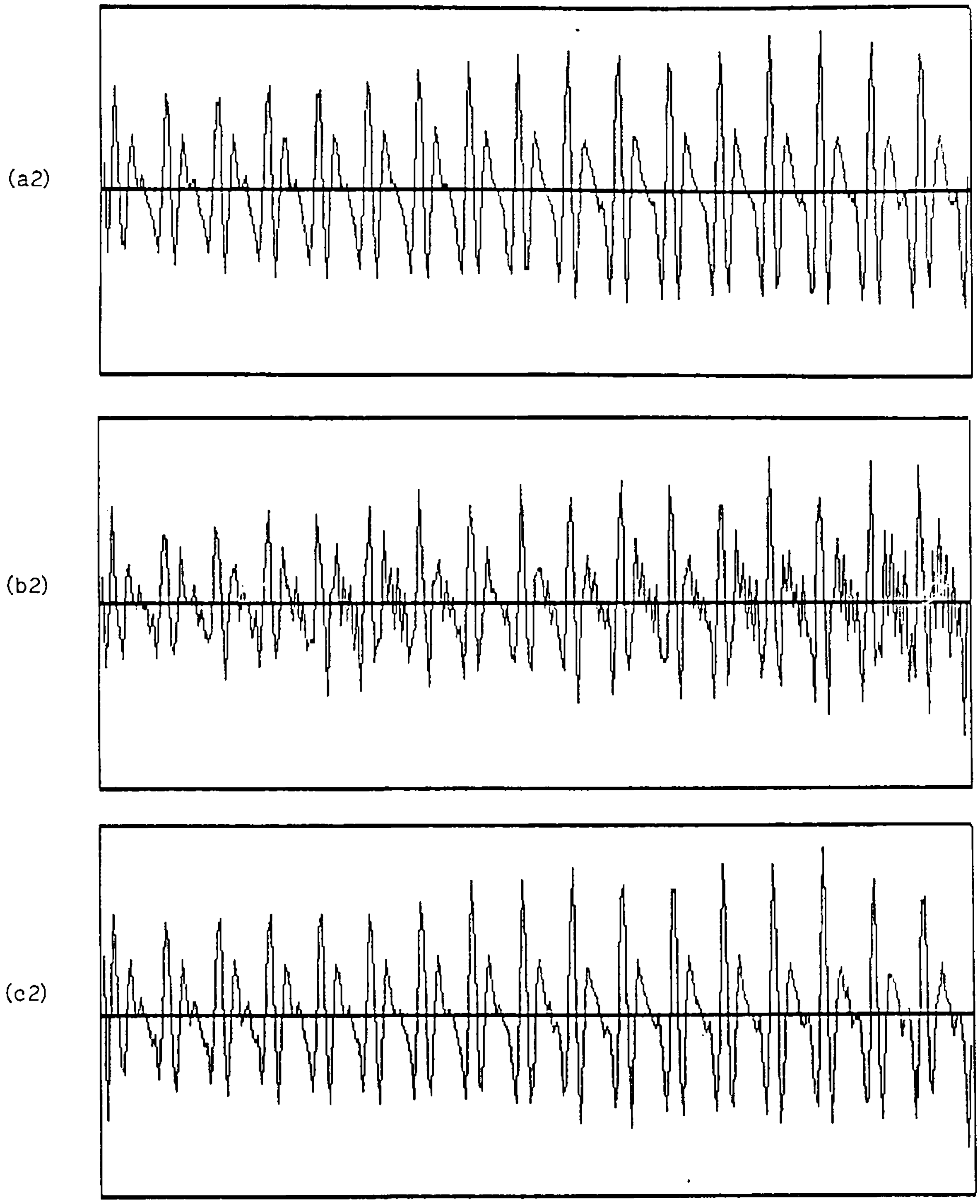


Fig. 3.7 (a2) original speech
(b2) reconstructed speech (unfiltered)
(c3) reconstructed speech (filtered

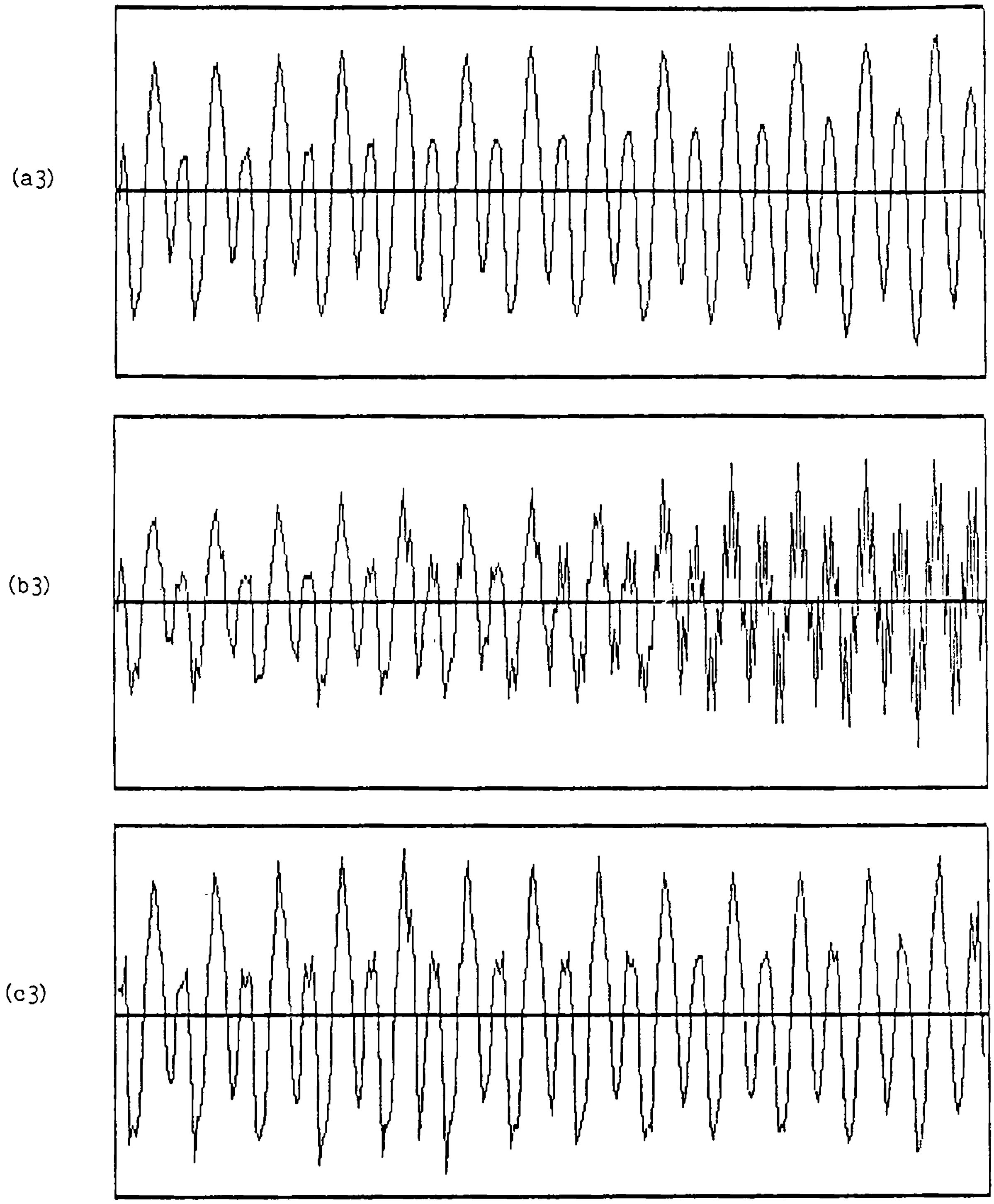


Fig. 3.7 (a3) original speech

(b3) reconstructed speech (unfiltered)

(c3) reconstructed speech (filtered)

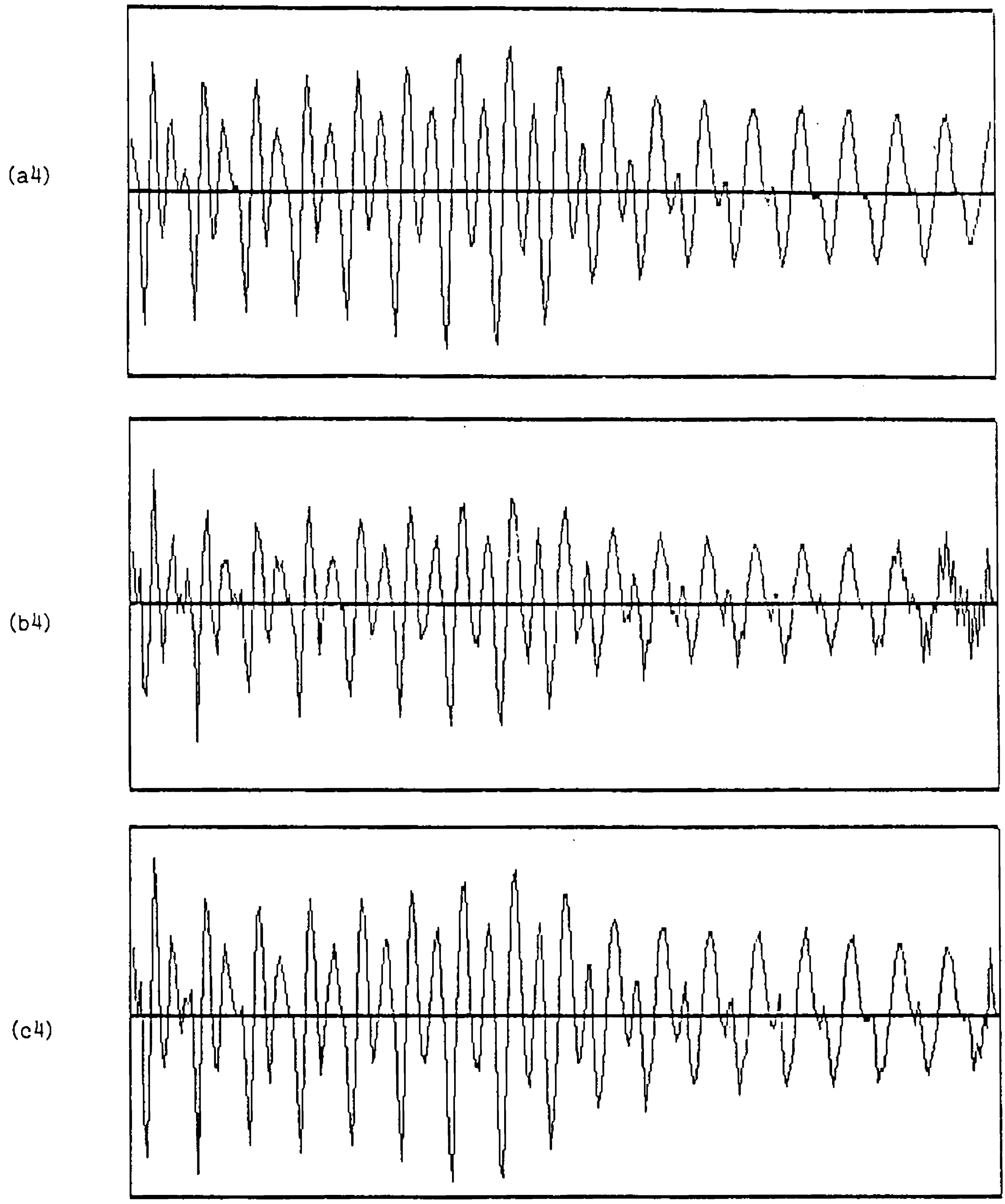


Fig. 3.7 (a4) original speech
(b4) reconstructed speech (unfiltered)
(c4) reconstructed speech (filtered)

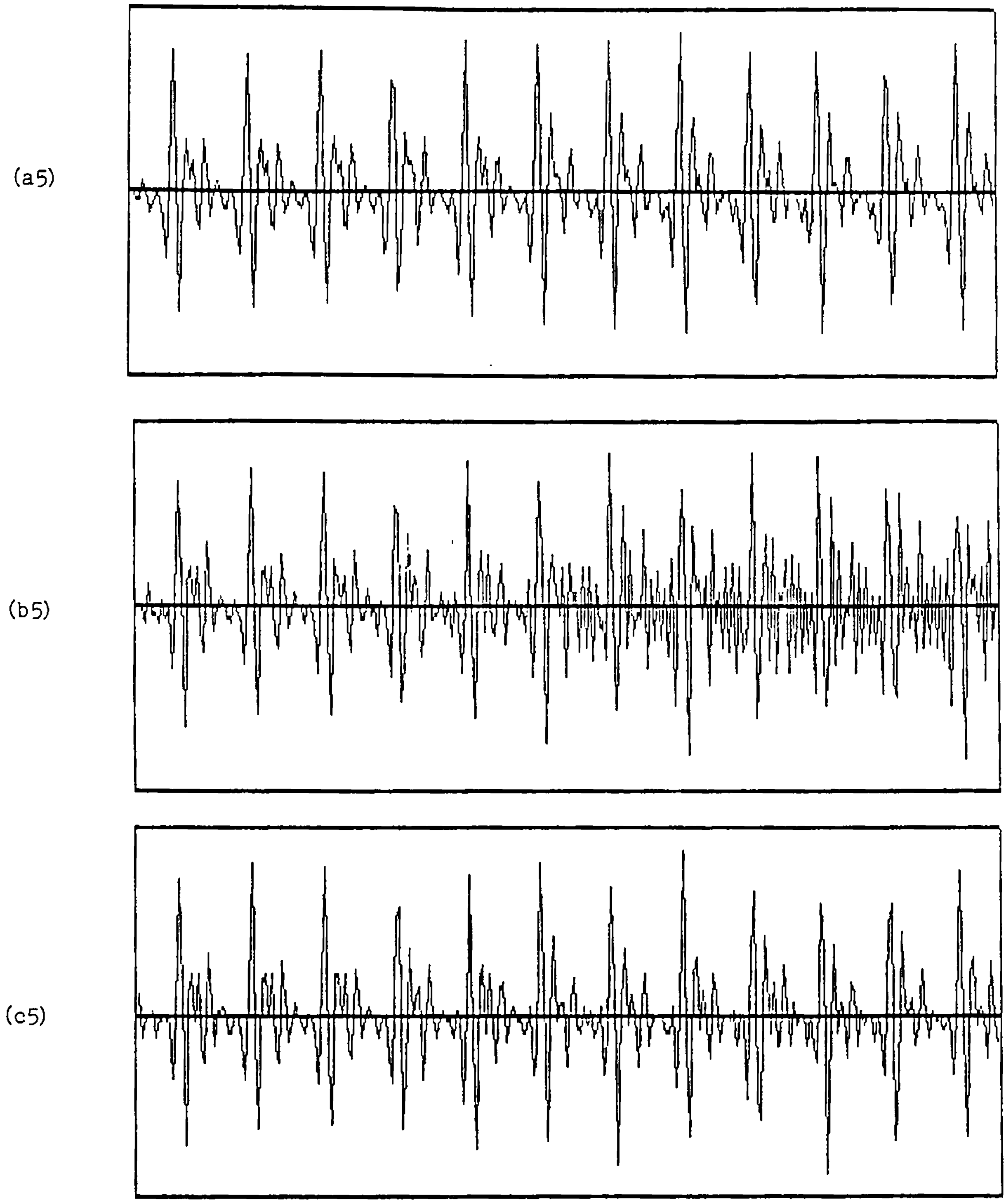


Fig. 3.7 (a5) original speech
(b5) reconstructed speech (unfiltered)
(c5) reconstructed speech (filtered)

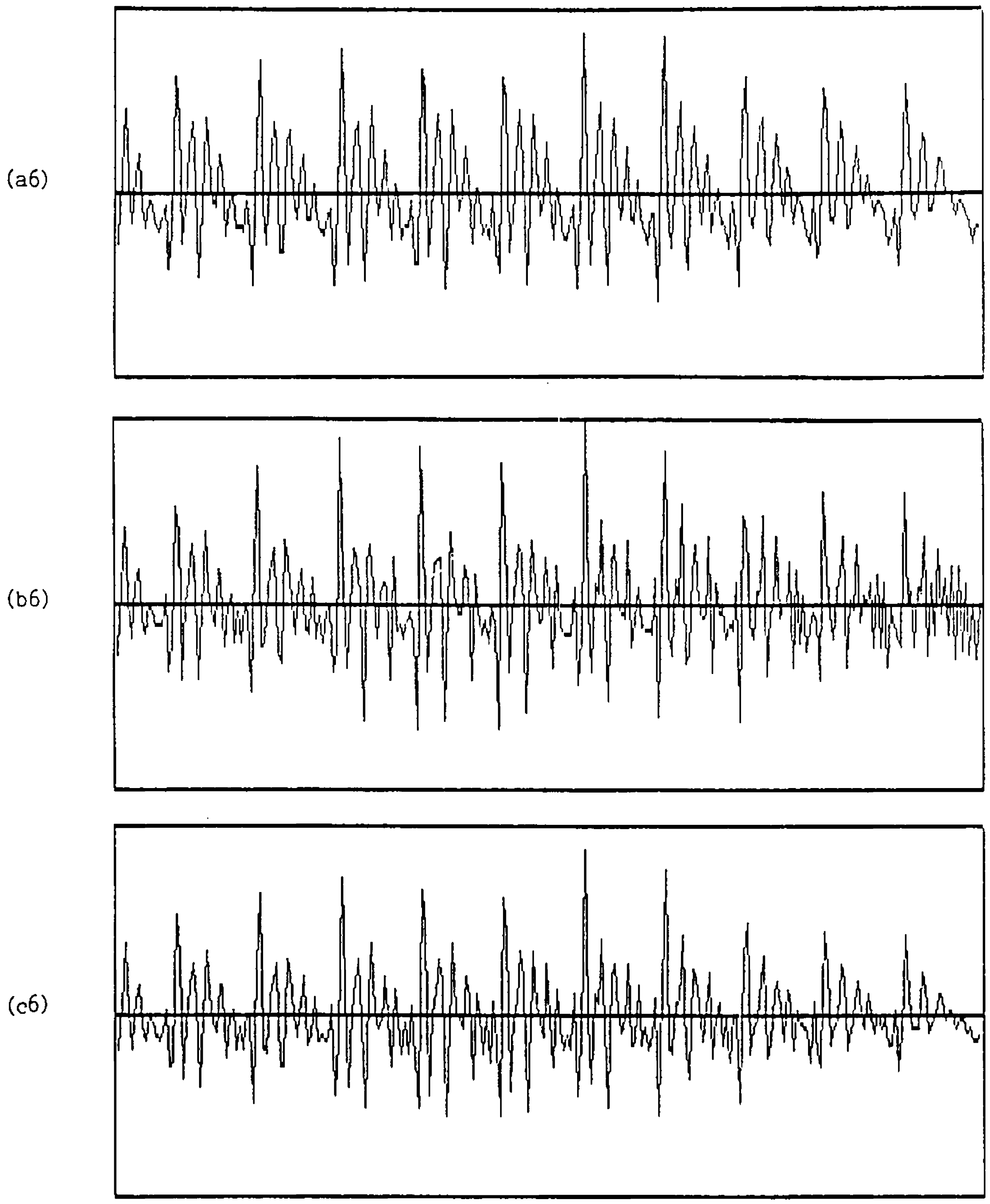


Fig. 3.7 (a6) original speech
(b6) reconstructed speech (unfiltered)
(c6) reconstructed speech (filtered)

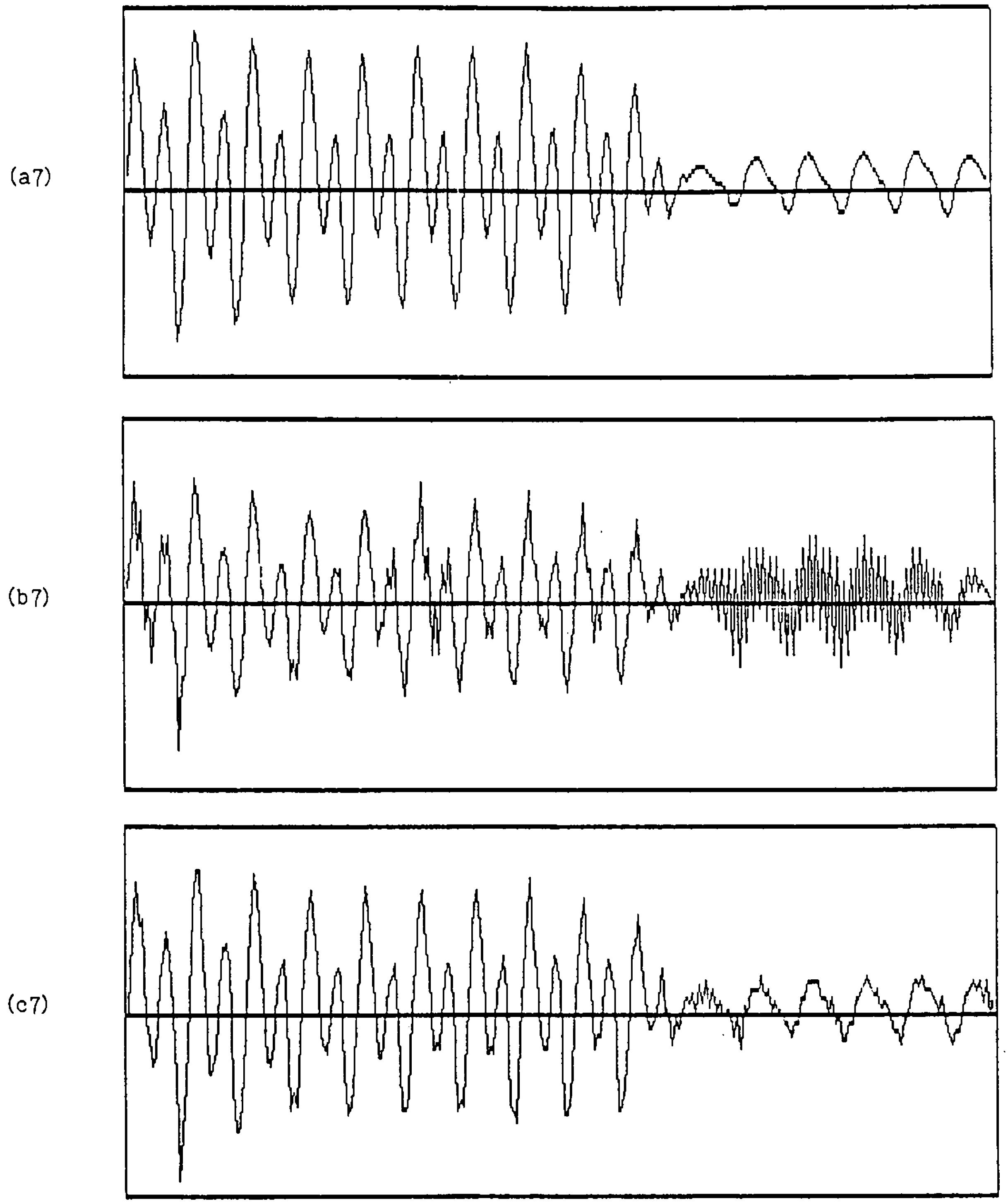


Fig. 3.7 (a7) original speech
(b7) reconstructed speech (unfiltered)
(c7) reconstructed speech (filtered)

using FFT. Figs.3.8(a) and (b) are the PSDs of the original and the reconstructed speech segment. The PSD plots indicate that the TIF introduce high frequency noise in the estimated speech. It may be noted that the major noise components are near the added sinusoidal frequency. On examining a large number of reconstructed speech segments it is understood that if the reconstructed speech is filtered above 3750 Hz, major components of the noise introduced by the TIF can be suppressed. Fig.3.8(c) represents the PSD of the reconstructed speech lowpass filtered to 3750 Hz. The reconstructed speech shown in Fig.3.7(bl,b2,...,b7) are lowpass filtered to 3750 Hz and are shown in Figs. 3.7(cl,c2,...,c7) respectively. Figs.3.7(al,a2,...,a7) and Figs.3.7(cl,c2,...,c7) illustrate that there is little apparent loss of information by reconstructing the signal using the TIF from the 8-bit quantized values of the zerocrossings of the original signal plus the large-level sinusoid.

The perceptual quality of the estimated speech using 8 bit quantized zerocrossing is examined by subjective listening tests. This is carried out with sixteen listeners.

The original speech, zerocrossing estimated (unfiltered) and zerocrossing estimated (filtered) are presented

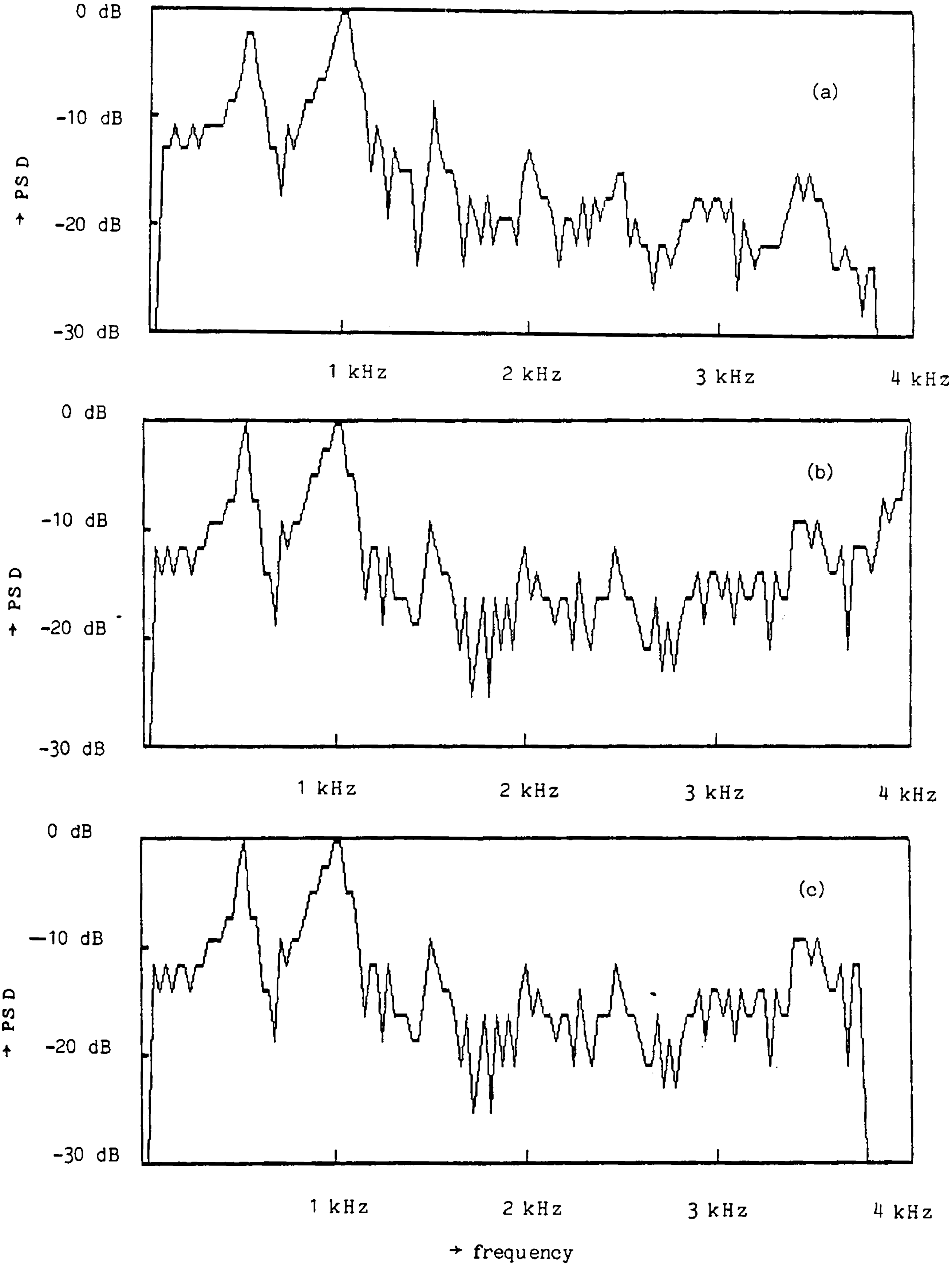


Fig. 3.8 (a) PS D of original speech

(b) PS D of reconstructed speech (unfiltered)

(c) PS D of reconstructed speech (filtered) in that order to the listeners. They are asked to judge the quality of the estimated speech (both unfiltered and filtered) by comparing with the original on the basis of its intelligibility, clarity, crispness, hoarseness and warbling effect. The following description is a summary of the subjective evaluation of the listeners.

The reconstructed speech (both unfiltered and filtered) is fully intelligible. The tonal quality is maintained in the filtered speech. The unfiltered estimated speech has a warbling effect indicating the presence of high frequency distortion. However, the intelligibility is still perfect and the distortions are not very objectionable. When the estimated speech is lowpass filtered to 3750 Hz, the quality is improved in the sense that the warbling noise level is reduced considerably. The filtered reconstructed speech has almost same level of quality as that of the original speech.

In order to obtain the Mean Opinion Score (MOS), the listeners are asked to rate the filtered reconstructed speech on an absolute scale, ranging between 1 and 5 by comparing with the original. The meaning of these grades are:

5 - excellent

4 - good

3 - fair

2 - poor

1 - bad

The mean value of the grades rated by the listeners, the MOS is equal to 4.56, which is a sufficient score for a good communication quality speech encoder. Therefore, the 8 bit quantized zerocrossing based speech sample estimation can be rated as having good communication quality.

3.3.2 Effect of Zerocrossing Quantization

Performance of the zerocrossing based sample estimation method using zerocrossings quantized to different number of bits are compared in this section. A figure of merit used very often to compare waveform coding system is the signal-to-noise ratio (SNR). The SNR(dB) of the reconstructed speech using zerocrossings quantized to different number of bits is calculated using the formula (2.4) defined in Chapter 2. Table 3.2 lists the SNR values for some segments of reconstructed speech using zerocrossings quantized to different number of bits. The SNR values of the same reconstructed speech segments lowpass filtered

Table 3.2

SNR(dB) OF RECONSTRUCTED SPEECH USING ZCL QUANTIZED TO DIFFERENT BITS

Seg. NO.	4 bit	5 bit	6 bit	7 bit	8 bit	9 bit	10bit	11bit	12bit	13bit	14bit
1	0.58	0.63	2.33	2.16	6.71	13.63	13.41	12.69	12.64	13.04	12.91
2	0.28	0.35	0.97	6.53	3.94	5.34	5.33	5.39	5.98	5.06	5.18
3	-0.18	0.55	1.53	4.01	10.84	12.13	12.11	11.96	12.44	12.80	12.85
4	-3.04	1.34	1.98	5.55	6.47	7.08	7.04	7.26	7.62	7.62	7.59
5	-2.65	0.01	0.50	1.87	2.28	2.68	2.51	2.64	2.43	2.45	2.42
6	-2.64	1.20	2.08	3.17	7.29	6.63	7.59	7.75	8.08	8.14	8.01
7	-3.74	-0.16	0.98	5.13	10.51	10.58	10.54	11.25	11.44	11.32	11.45
8	-0.58	0.44	0.80	8.23	9.04	9.56	9.65	9.09	9.00	9.79	9.52
9	-0.53	0.44	1.19	1.77	2.67	3.78	6.60	6.29	7.10	7.66	7.53
10	0.01	0.13	0.50	4.24	4.25	4.37	4.11	4.57	4.64	4.67	4.62
11	-0.37	0.35	1.73	3.96	4.46	7.07	6.74	6.78	6.86	6.96	7.14
12	0.75	1.18	1.60	4.41	6.12	6.10	6.44	7.14	7.45	7.15	7.14
13	-0.17	0.06	0.39	0.91	2.05	2.92	2.96	3.92	3.94	3.77	3.92
14	-1.63	-0.39	-0.39	5.46	4.43	7.67	7.35	7.54	7.02	7.32	7.22
15	-0.77	-0.78	0.28	0.34	6.29	6.45	6.63	6.78	7.19	6.94	6.87
16	-0.22	-0.12	0.21	0.68	3.20	4.43	4.15	4.79	4.81	4.83	4.89
17	-0.27	1.69	2.13	8.70	10.53	9.06	9.84	9.70	9.21	9.54	9.41
18	-1.75	-0.73	4.42	4.57	5.44	6.98	7.13	6.99	7.01	7.06	7.24
19	0.05	0.38	0.95	2.59	4.61	5.47	5.61	5.76	5.78	5.77	5.77
20	0.61	1.48	2.04	6.53	7.89	7.62	7.37	7.59	7.04	7.14	7.11
21	-0.37	0.01	1.69	2.62	4.49	8.01	8.41	8.33	8.12	7.90	8.21
22	-1.12	-0.80	1.16	1.14	1.92	2.98	3.83	3.91	3.63	3.68	3.65
23	-0.15	-0.06	0.18	4.70	6.52	8.52	8.66	8.08	8.11	8.19	8.16
24	-0.67	0.05	2.84	7.08	7.66	7.85	8.65	9.00	9.38	9.87	9.78
25	-2.25	1.16	4.46	8.52	8.25	8.11	8.44	8.80	8.26	8.15	8.18

Table 3.3

SNR(dB) OF FILTERED RECONSTRUCTED SPEECH USING ZCL QUANTIZED TO DIFFERENT BITS

Seg. NO.	4 bit	5 bit	6 bit	7 bit	8 bit	9 bit	10bit	11bit	12bit	13bit	14bit
1	1.85	10.06	14.68	17.33	22.17	24.39	24.62	24.81	24.77	24.81	24.86
2	2.41	8.57	9.06	17.91	24.18	23.76	24.45	24.38	24.32	24.61	24.57
3	4.54	5.24	8.66	15.75	19.89	20.04	20.00	19.77	20.14	20.25	20.26
4	2.31	10.07	15.28	17.04	19.81	21.39	21.60	21.49	21.45	21.60	21.59
5	1.21	5.71	9.73	15.21	16.64	17.09	17.07	16.87	16.94	16.95	16.95
6	3.76	8.48	12.80	18.52	17.45	19.96	18.54	18.70	19.01	19.06	19.06
7	6.00	10.92	9.92	19.06	17.81	18.64	18.40	18.47	18.35	18.24	18.23
8	3.35	7.31	9.31	12.66	12.61	12.27	12.34	12.14	12.21	12.19	12.17
9	1.66	2.71	10.20	10.67	13.94	17.46	18.35	18.87	19.17	18.98	18.96
10	4.25	7.32	9.19	14.85	16.78	15.86	15.35	16.24	16.02	16.03	16.01
11	1.28	5.08	9.00	13.78	16.61	16.48	16.80	16.70	16.64	16.65	16.67
12	6.63	7.47	12.24	14.71	16.63	16.48	16.84	17.07	17.12	17.10	17.08
13	3.21	8.37	12.32	15.70	17.81	18.02	18.03	18.30	18.33	18.31	18.33
14	1.90	8.69	12.97	16.85	17.66	18.97	19.54	18.92	18.89	18.93	18.92
15	3.59	5.60	13.25	15.95	18.39	18.92	18.57	18.83	18.85	18.75	18.77
16	2.29	4.60	9.90	11.90	17.20	18.38	18.34	18.84	18.96	18.94	19.00
17	7.41	12.40	14.89	15.34	15.97	15.37	15.86	15.83	15.76	15.82	15.79
18	2.48	4.62	11.68	12.44	13.22	13.45	13.44	13.44	13.44	13.43	13.43
19	3.26	6.31	11.82	12.71	14.76	15.63	15.92	16.20	16.09	16.09	16.08
20	4.28	10.81	13.26	19.08	20.47	20.27	20.66	20.27	20.46	20.52	20.45
21	2.72	7.37	13.60	13.95	15.37	18.24	18.18	18.89	18.57	18.52	18.49
22	3.62	3.86	11.96	13.54	14.04	14.17	14.06	13.94	13.97	13.96	14.00
23	3.38	7.01	9.42	12.87	12.40	13.05	13.19	13.14	13.30	13.31	13.32
24	7.53	12.09	15.64	19.34	18.32	19.96	20.81	22.07	21.72	22.01	21.91
25	2.60	9.12	11.59	12.98	13.33	13.48	13.35	13.26	13.26	13.25	13.25

to 3750 Hz are listed in Table 3.3. It may be·noted that there is a notable increase in SNR when the lowpass filtering is performed. This confirms that the major noise components introduced by the zerocrossing interpolation method are in the frequency region above 3750 Hz. The SNR values versus number of bits for zerocrossing quantization for reconstructed speech segments (unfiltered and filtered) are illustrated in Fig.3.9(1-24).

Let $SNR_{zc}(m)$ (dB), where m = 1,2,...,M represents the signal-to-noise ratio of the M segments, for each particular number of quantization bits. Now we can define the average SNR over a frame of speech containing M segments as

$$\text{SEGSNR} = \frac{1}{M} \sum_{m=1}^{M} SNR_{zc}(m)\,(\text{dB})$$

The SEGSNR obtained using the speech data base for different number of bits for zerocrossing quantization is shown in Table 3.4. The SEGSNR comparison of reconstructed and filtered reconstructed speech versus number of bits for zerocrossing quantization is illustrated in Fig.3.10.

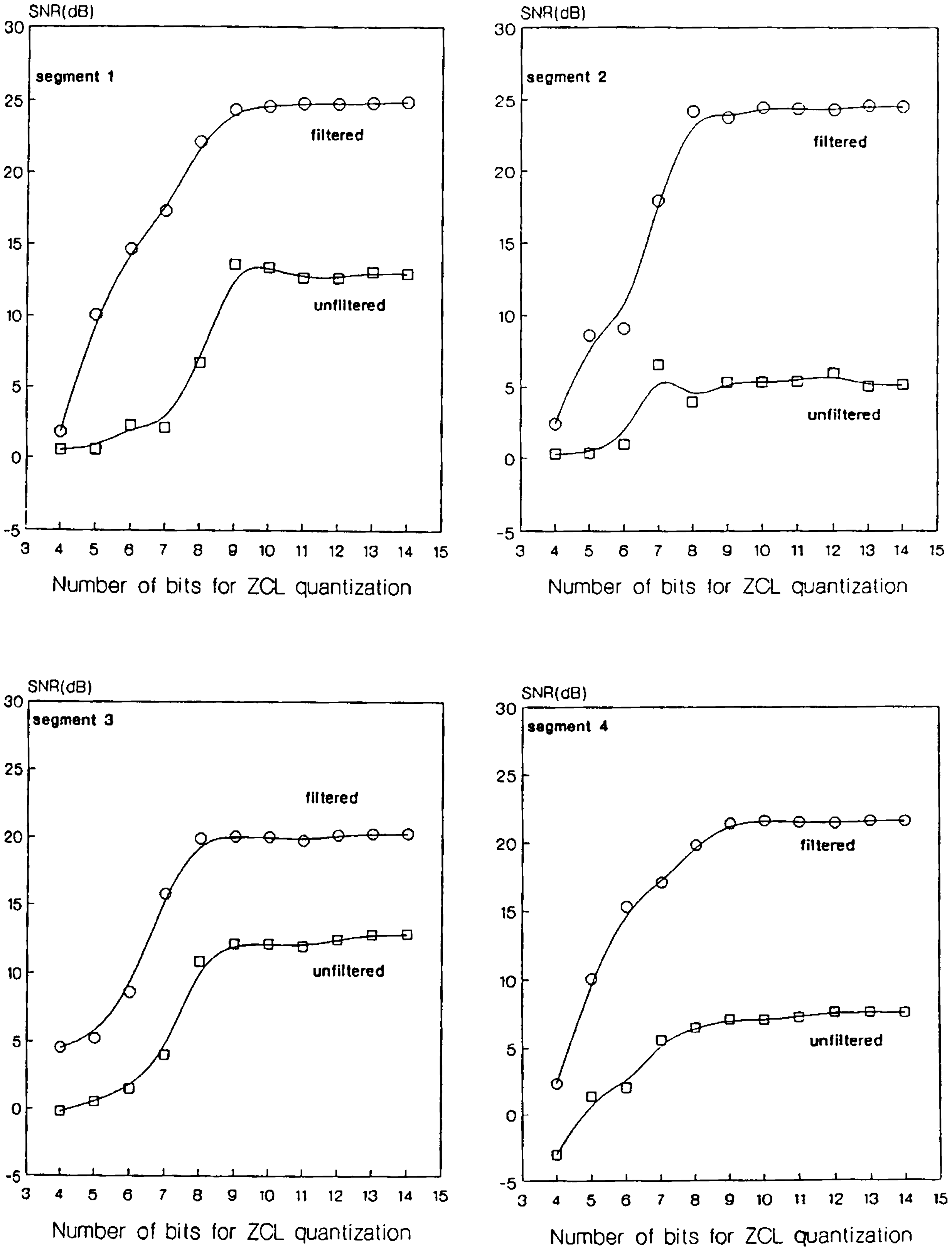


Fig. 3.9(1-4) SNR comparison of reconstructed and filtered reconstructed speech vs number of bits for zerocrossing quantization

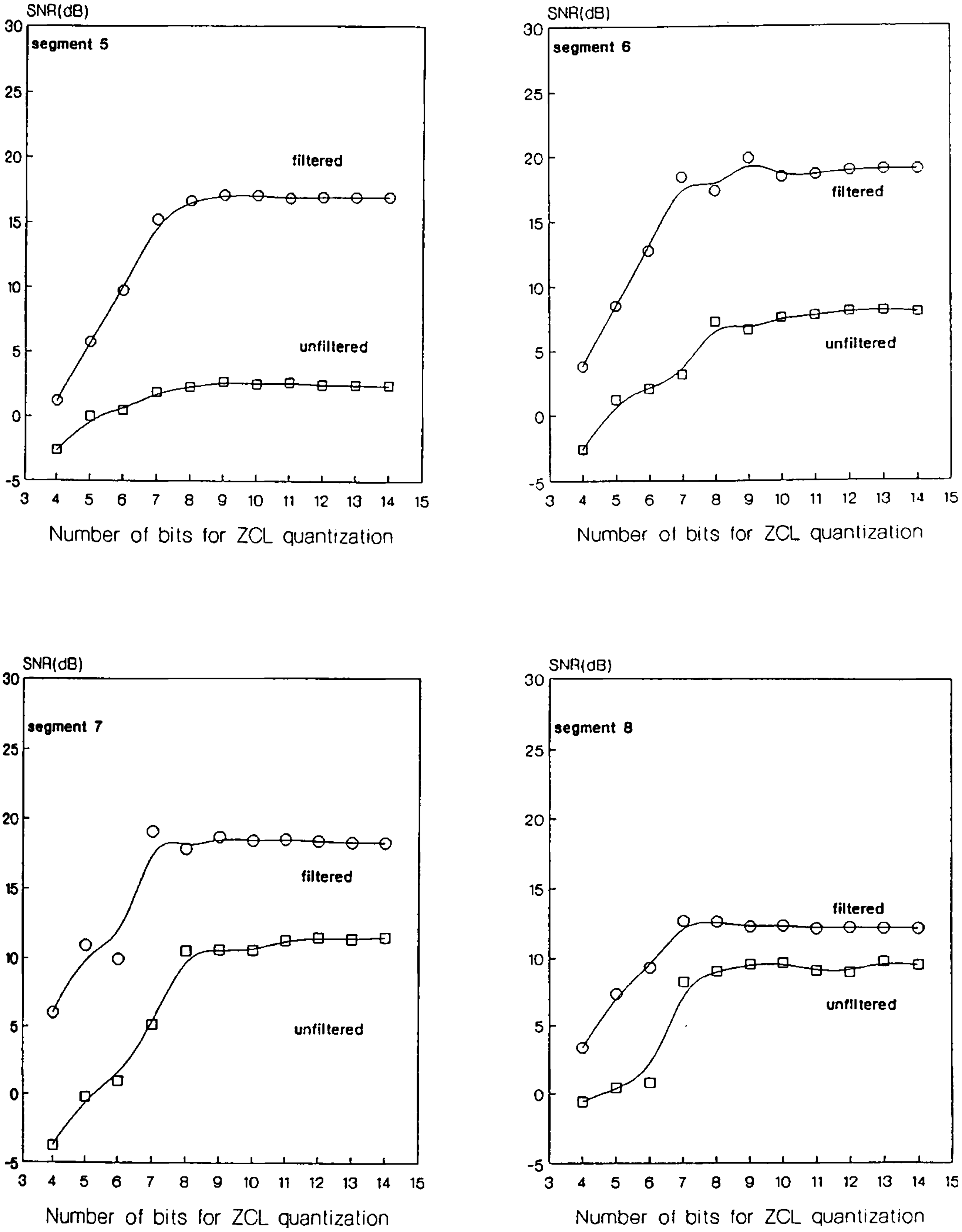


Fig. 3.9(5-8) SNR comparison of reconstructed and filtered reconstructed speech vs number of bits for zerocrossing quantization

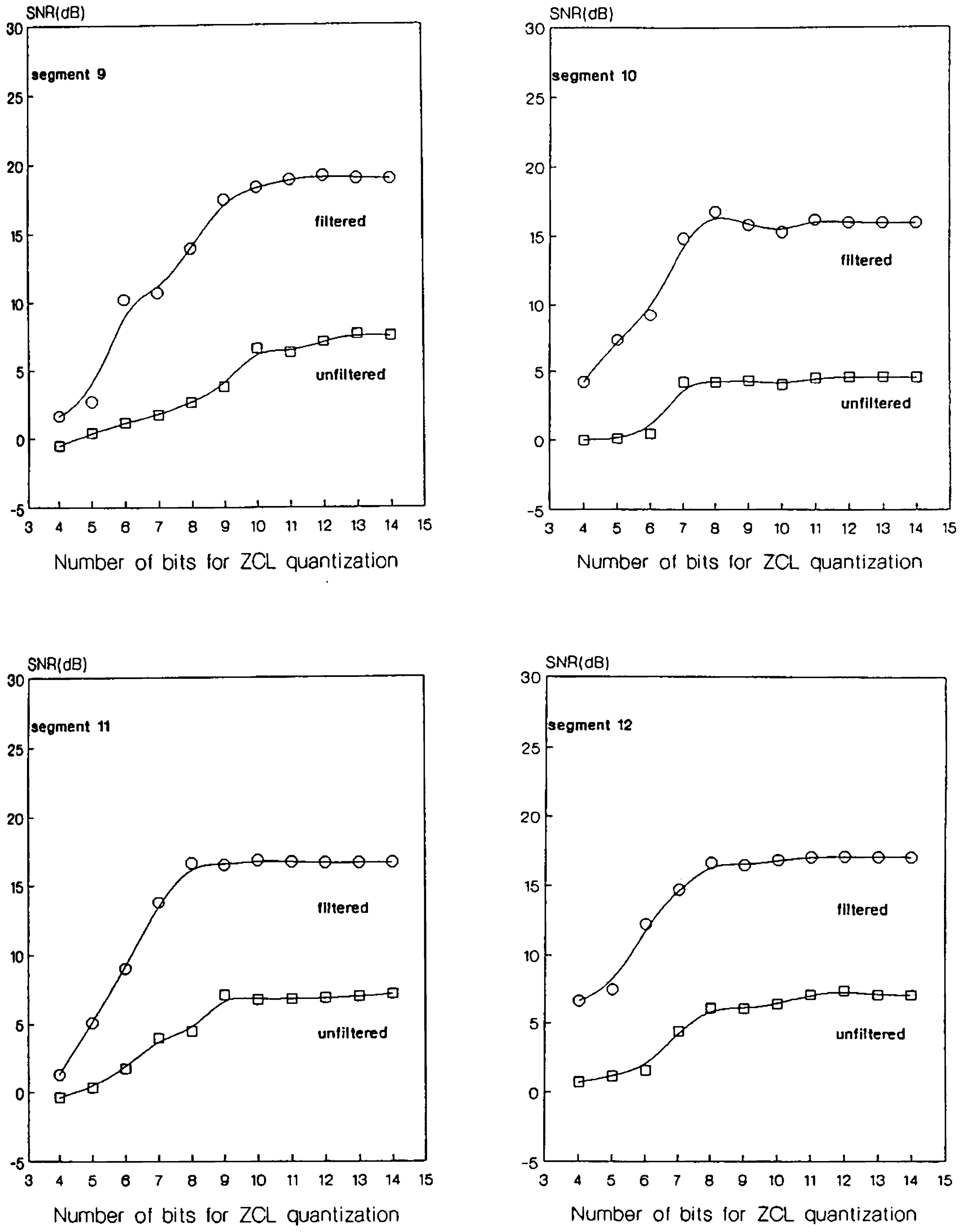


Fig. 3.9(9-12) SNR comparison of reconstructed and filtered reconstructed speech vs number bits for zerocrossing quantization

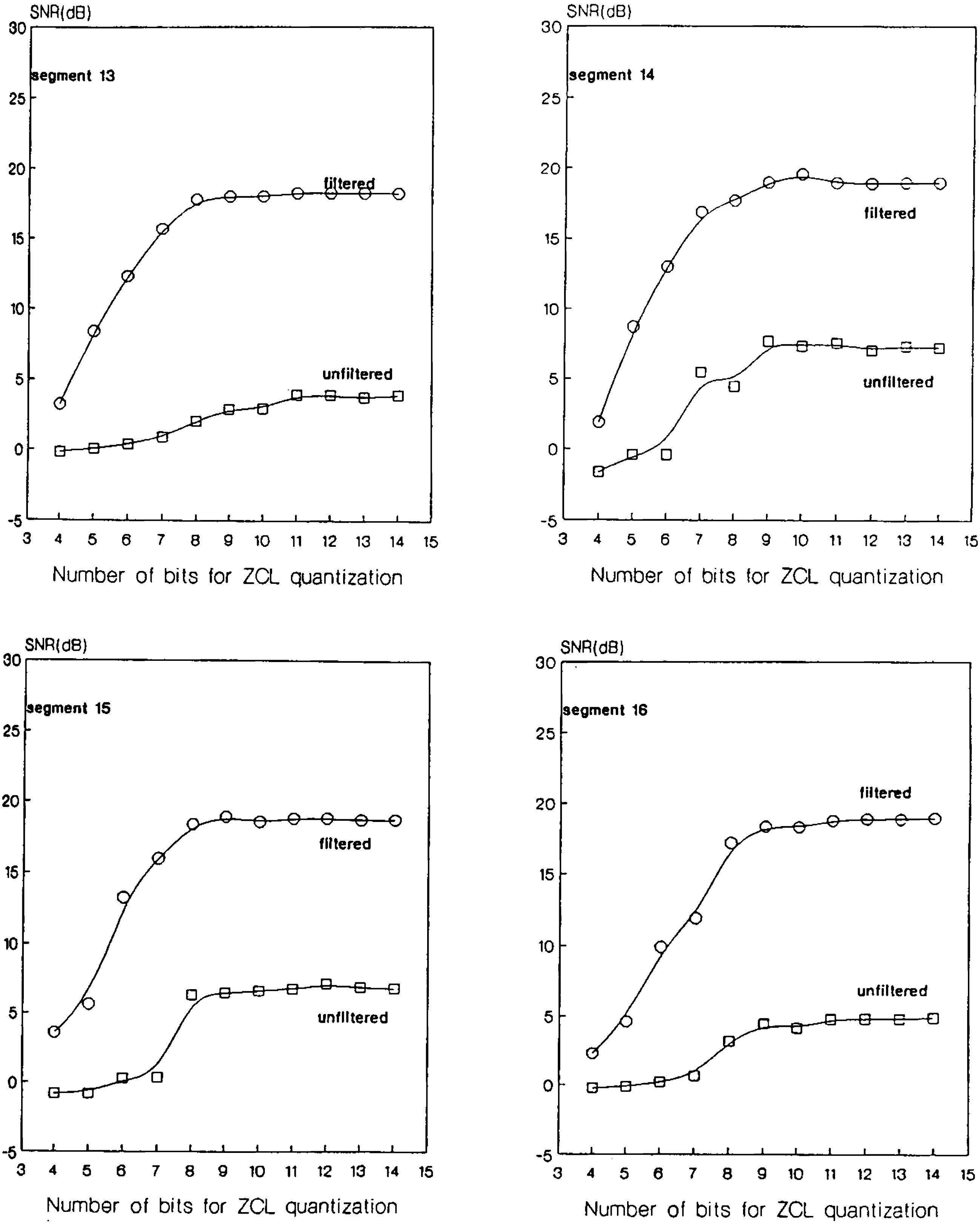


Fig. 3.9(13-16) SNR comparison of reconstructed and filtered reconstructed speech vs number of bits for zerocrossing quantisation

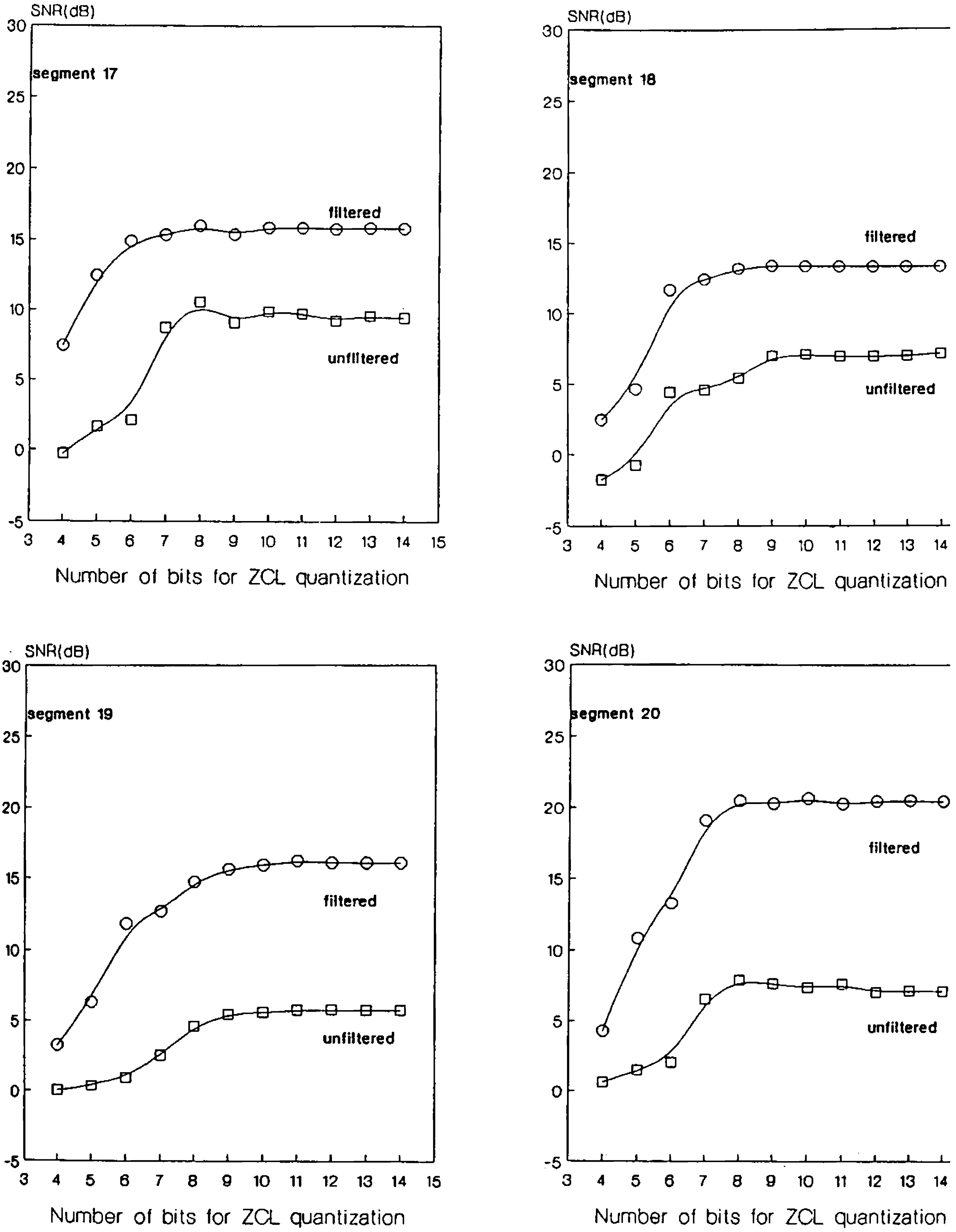


Fig.3.9(17-20) SNR comparison of reconstructed and filtered reconstructed speech vs number of bits for zercrossing quantization

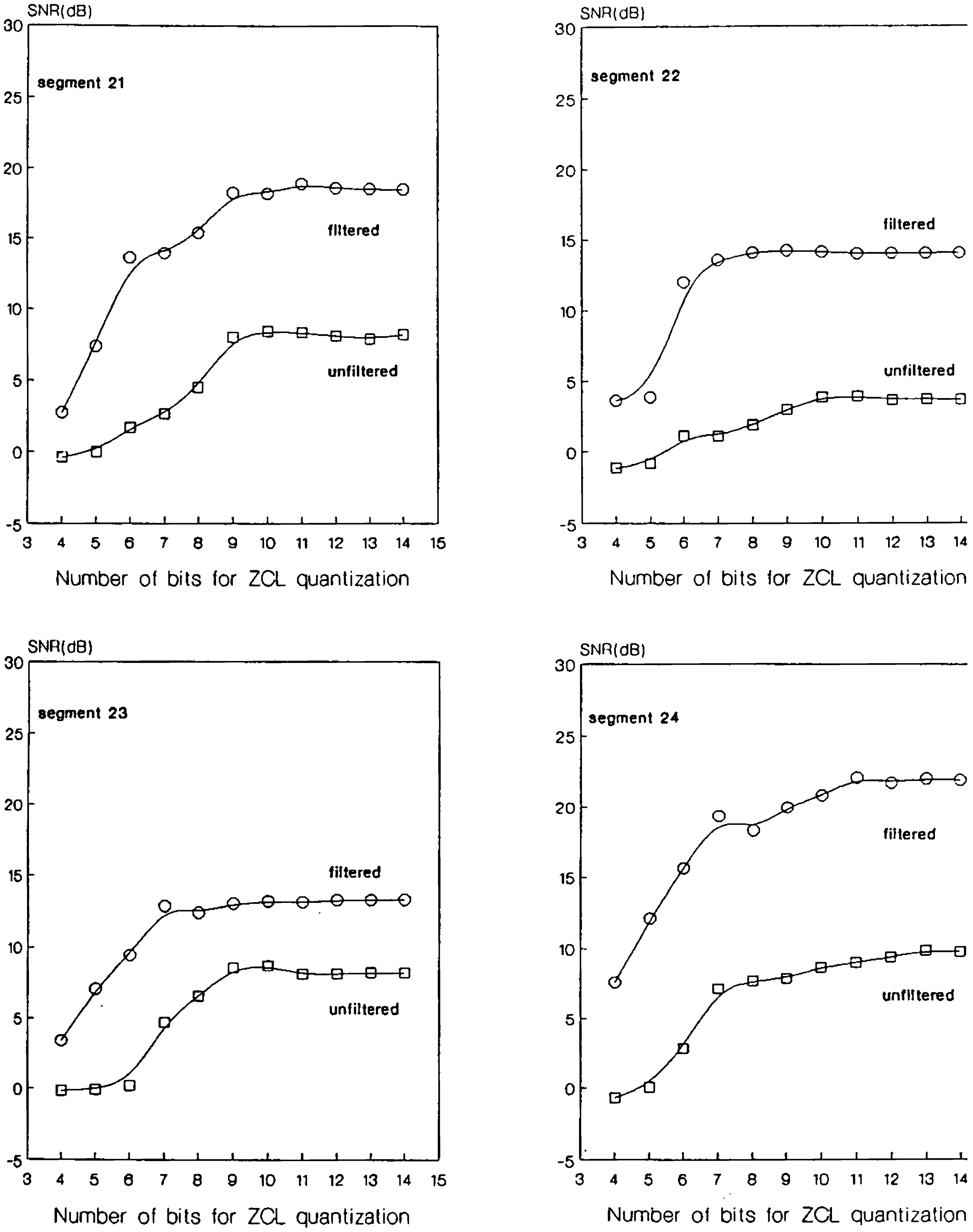


Fig. 3.9(21-24) SNR comparison of reconstructed and filtered reconstructed spe
vs number of bits for zerocrossing quantization

Table 3.4 SEGSNR (dB) of reconstructed speech

Number of bits for zerocrossing quantization	4	5	6	7	8	9	10	11	12	13	14
SEGSNR (dB) of reconstructed speech (Unfiltered)	-0.87	0.31	1.73	4.29	6.03	7.14	7.60	7.59	7.63	7.68	7.68
SEGSNR (dB) of reconstructed speech (Lowpass filtered)	3.31	6.94	11.86	15.31	16.91	17.63	17.75	17.88	17.88	17.87	17.89

SEGSNR comparison of reconstructed and filtered reconstructed speech vs number of bits for zerocrossing quantization

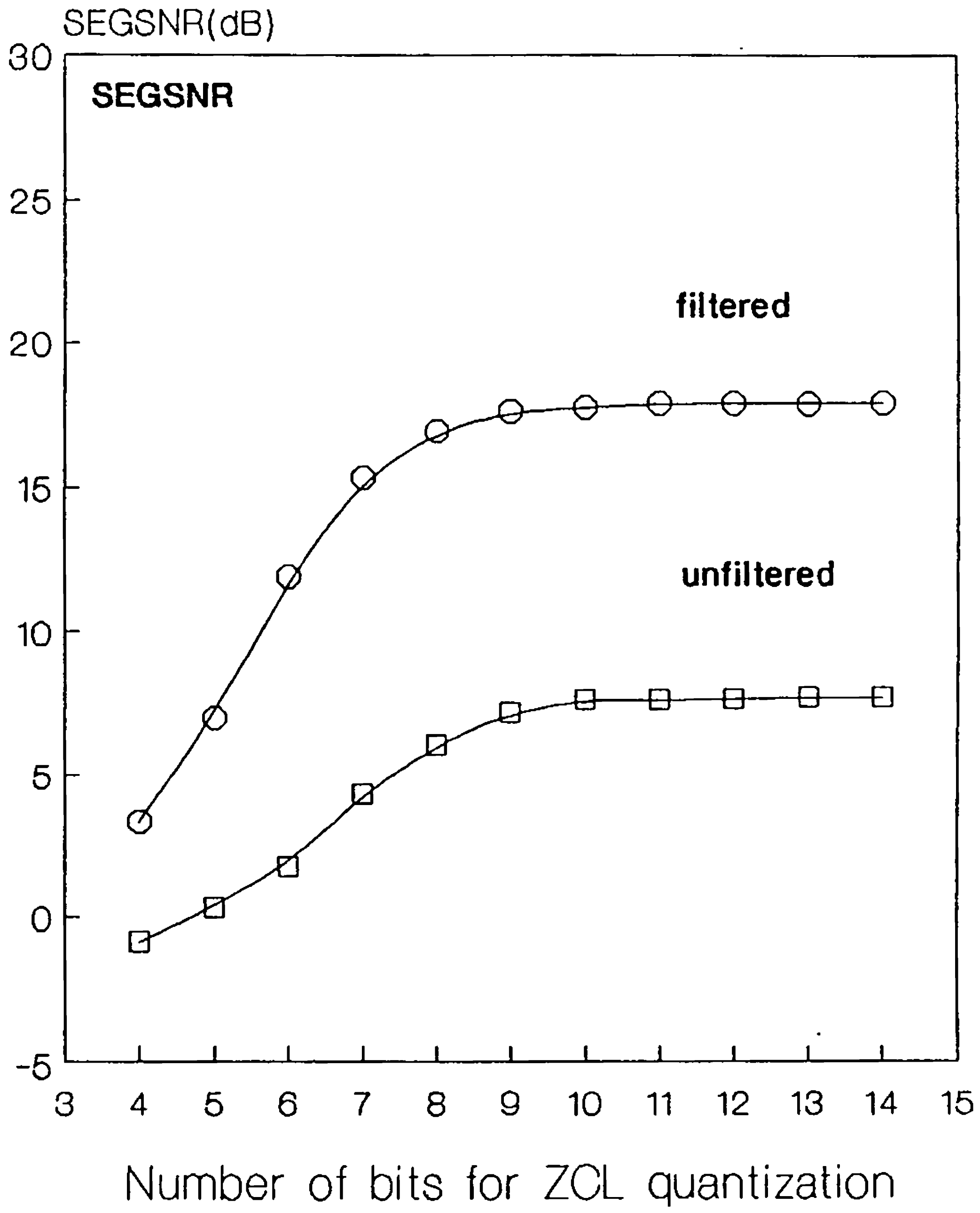


Fig. 3.10

The subjective quality tests performed in previous section confirmed that the filtered reconstructed speech using 8 bit quantized zerocrossings is having good communication quality. The experimental results also indicate that the SNR values are almost same when more than 8 bits are used for zerocrossing quantization. This is because of the reason that the quantization error is very small compared to the interpolation error for these bit ranges. Therefore the zerocrossing based speech sample estimation method may be used as an alternative to the A/D converter method. By using this method, the cost of the digitizer system can be considerably reduced, since simple digital circuits are sufficient for the extraction of zerocrossings.

3.3.3 Effect of Signal Statistics

The applicability of the proposed method for noisy speech signal is studied by adding zero mean Additive White Gaussian Noise (AWGN) of different variance values to the original speech and repeating the above experiment. The SEGSNR obtained for the reconstructed signal for zerocrossing quantized to different number of bits for the signal (speech + AWGN) is listed in Table 3.5. The variance of the AWGN is selected to get 40 dB, 30 dB, 10 dB, 3 dB

Table 3.5(a)-(f) SEGSNR (dB) of reconstructed signal (speech + AWGN) for zerocrossing quantized to different number of bits

Table 3.5(a) 40 dB Speech

Number of bits for zerocrossing quantization	4	5	6	7	8	9	10	11	12	13	14
SEGSNR (dB) (unfiltered)	-3.86	-0.53	1.57	5.79	6.21	7.02	7.12	7.17	7.20	7.23	7.19
SEGSNR (dB) (Lowpass filtered)	1.88	7.78	11.32	16.06	17.04	17.58	17.56	17.33	17.40	17.43	17.45

Table 3.5(b) 30 dB speech

Number of bits for zerocrossing quantization	4	5	6	7	8	9	10	11	12	13	14
SEGSNR (dB) (Unfiltered)	-5.92	-2.75	1.07	3.45	5.08	5.58	6.07	6.35	6.31	6.32	6.33
SEGSNR (dB) (Lowpass filtered)	1.6	7.19	9.64	14.95	16.71	16.77	16.82	16.80	16.81	16.81	16.84

Table 3.5(c) 20 dB speech

Number of bits for zerocrossing quantization	4	5	6	7	8	9	10	11	12	13	14
SEGSNR (dB) (Unfiltered)	-9.24	-5.85	-1.38	1.50	1.78	2.54	2.93	3.07	3.01	3.01	3.03
SEGSNR (dB) (Lowpass filtered)	1.18	6.52	8.44	12.35	14.62	14.57	14.75	14.95	14.97	14.93	14.92

Table 3.5(d) 10 dB speech

Number of bits for zerocrossing quantization	4	5	6	7	8	9	10	11	12	13	14
SEGSNR (dB) (Unfiltered)	-11.97	-7.93	-4.14	-3.73	-3.66	-3.63	-3.62	-3.61	-3.62	- 3.64	-3.63
SEGSNR (dB) (Lowpass filtered)	0.28	4.02	7.02	8.98	9.42	9.40	9.53	9.54	9.56	9.53	9.54

Table 3.5(e) 3dB speech

Number of bits for zerocrossing quantization	4	5	6	7	8	9	10	11	12	13	14
SEGSNR (dB) (Lowpass filtered)	-14.98	-10.50	-8.86	-8.42	-7.68	-7.82	-7.82	-7.82	-7.82	-7.82	-7.82
SEGSNR (dB) (Lowpass filtered)	-0.88	2.71	3.45	4.28	4.64	4.72	4.72	4.72	4.73	4.73	4.73

Table 3.5(f) 0 dB speech

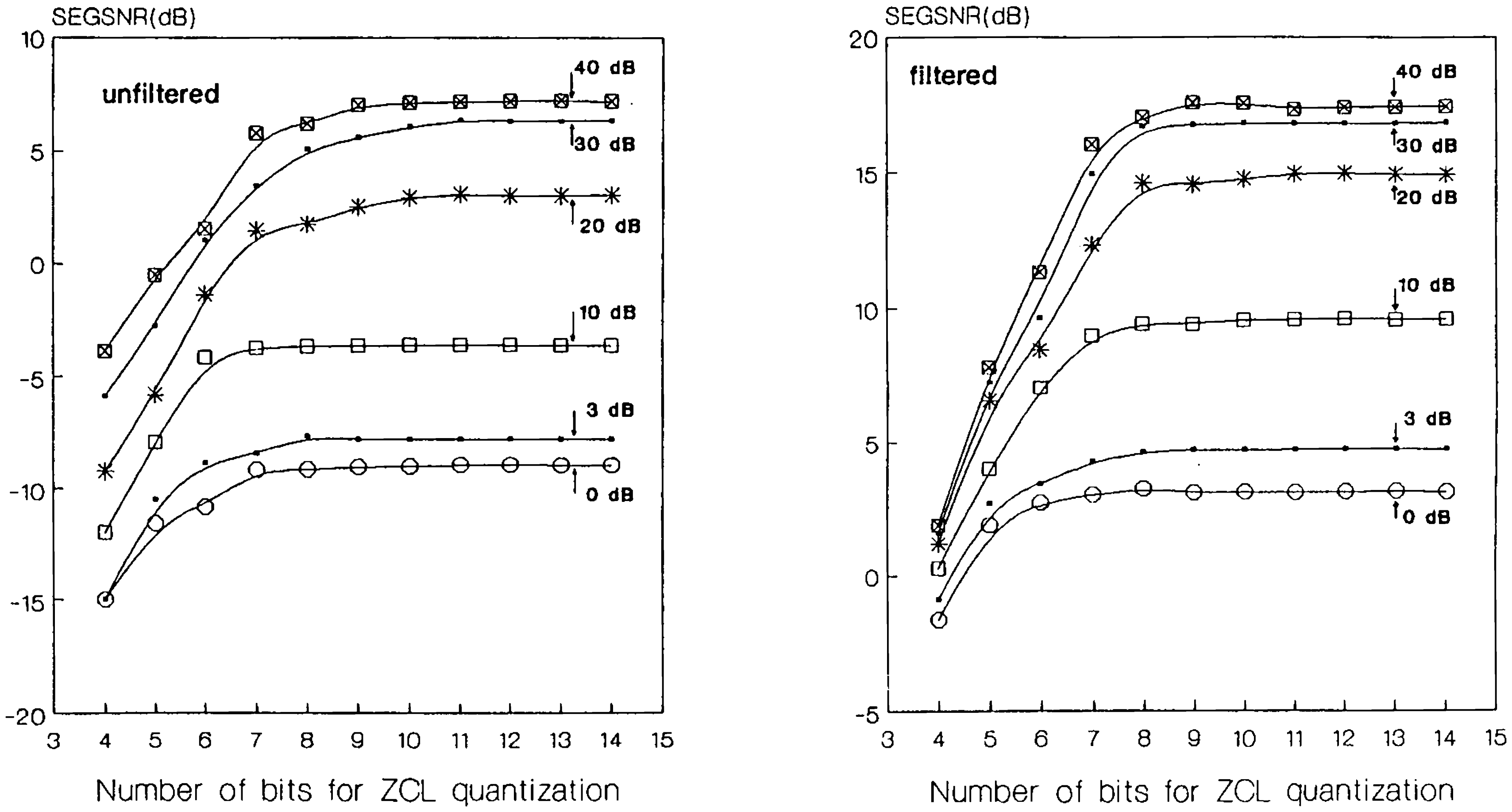
Number of bits for zerocrossing quantization	4	5	6	7	8	9	10	11	12	13	14
SEGSNR (dB) (Unfiltered)	-15.00	-11.56	-10.84	-9.16	-9.13	-9.04	-9.01	-8.95	-8.95	-8.97	-8.97
SEGSNR (dB) (Lowpass filtered)	-1.63	1.9	2.74	3.04	3.26	3.11	3.13	3.11	3.13	3.14	3.13

and 0 dB SNR for speech signal, i.e., the resulting signal statistics is changed over a wide range. The performance of the TIF for these signals is illustrated in Fig.3.11. It is noted that when the noise level is low, the SEGSNR of the reconstructed speech is not much affected. But there is considerable reduction in SEGSNR when the noise level is high.

3.3.4 Effect of the Amplitude of the Added Sinusoid

In the earlier part of the study the amplitude of the added high frequency sinusoid was equal to twice the maximum amplitude of the original signal i.e., $A = 2S(t)_{max}$. And the SNR value of the reconstructed speech signal was calculated changing the number of bits for quantization of the ZCL. In this section the effect of sinusoidal amplitude on the SNR performance is studied.

Computer simulation experiment carried out in this case is similar to that of the earlier case, except that there is a provision to change the sinusoidal amplitude. According to theory, the amplitude of the sinusoid must be greater than the maximum amplitude of the signal. In this study the signal used is normalized so that the


SEGSNR of unfiltered reconstructed noisy speech (speech + awgn) with different statistcs vs bits for ZCL quantization

SEGSNR of filtered reconstructed noisy speech (spech+awgn) with different statistics versus bits for ZCL quantization

Fig. 3.11

maximum amplitude is unity. We set the initial value of the amplitude of the added sinusoid equal to 1.10 in the program. The SNR value for this amplitude is computed and recorded. The amplitude of the sinusoid is then changed by an increment of 0.10 and the reconstruction experiment is repeated upto an amplitude equal to 4.00.

Figs.3.12(i-iv) illustrate the SNR variation of the reconstructed speech signal for different amplitudes of the added sinusoidal signal. It may be noted that the SNR is fluctuating randomly with sinusoidal amplitude about a mean value. The variation is large when lesser number of bits are used for quantizing the ZCL.

We have statistically analysed this random fluctuation of the SNR vs. sinusoidal amplitude. This is studied by testing the statistical distribution of the zerocrossing quantization error as well as the reconstruction error. For this, the simulation program is modified by incorporating provisions to compute the mean, variance, standard deviation and to test the chi-square goodness of fit for following probability distribution functions. The distributions tried are: (1) Normal, (2) Uniform, (3) Laplace, (4) Gamma, and (5) Log-Normal.

SNR(dB) OF RECONSTRUCTED SPEECH FOR DIFFERENT SINUSOIDAL AMPLITUDES (6&8 bits for ZCL quantization)

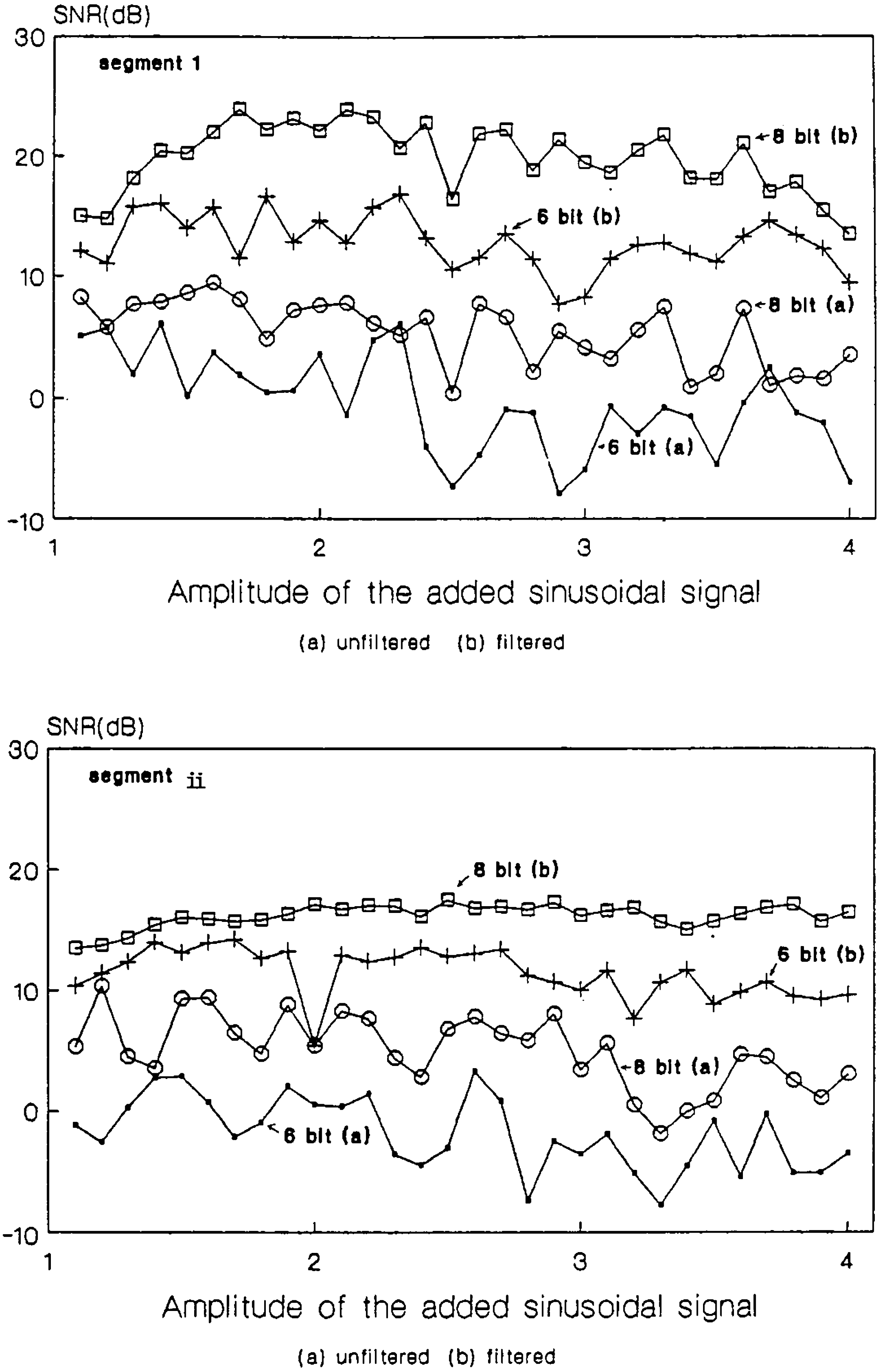


Fig. 3.12(i & ii)

SNR(dB) OF RECONSTRUCTED SPEECH FOR DIFFERENT SINUSOIDAL AMPLITUDES (6&8 bits for ZCL quantization)

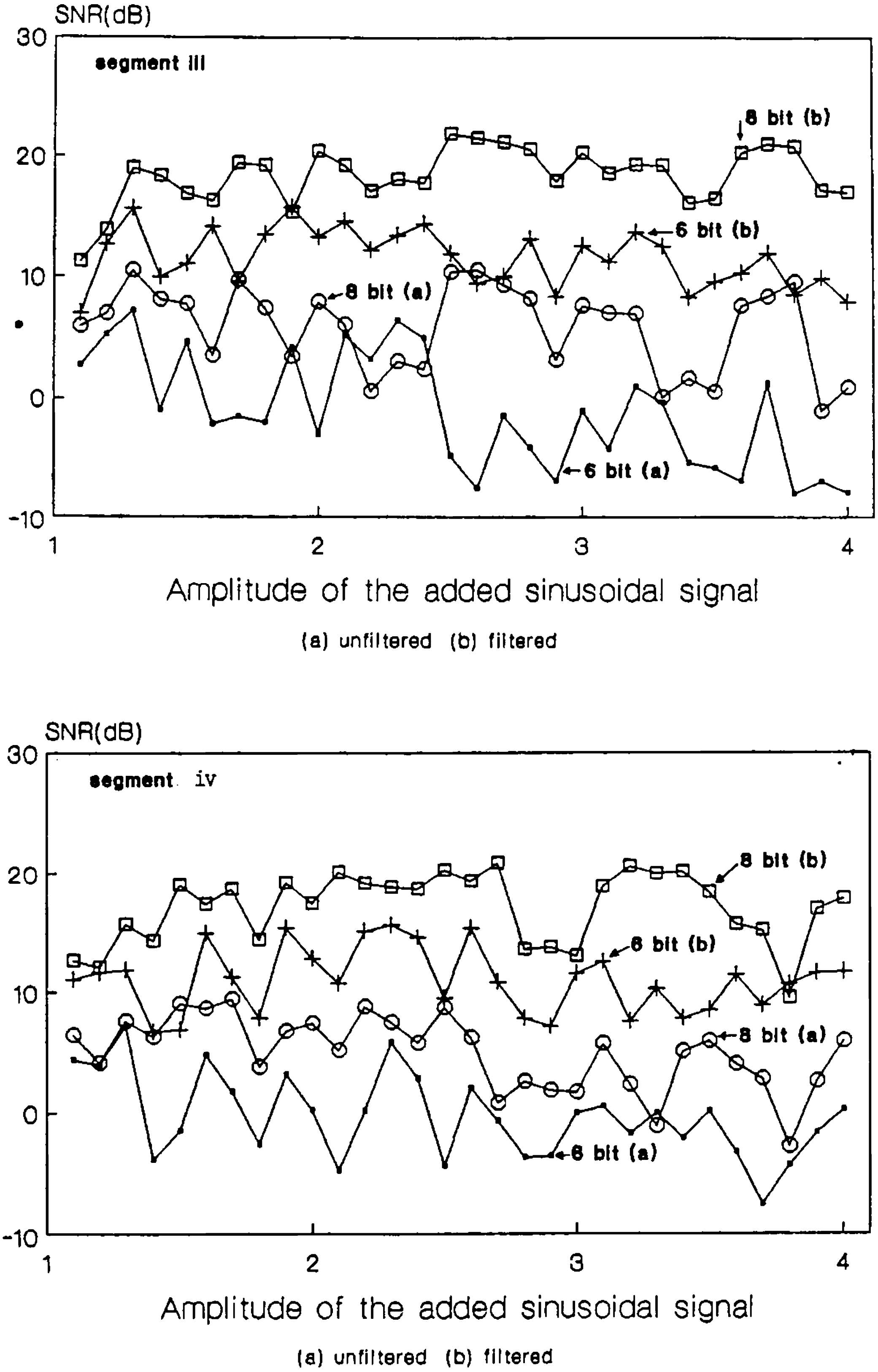


Fig. 3.12(iii & iv)

HISTOGRAM STUDIES

Simulation study was performed to find the statistical behaviour of the zerocrossing quantization error and the reconstruction error using histogram and chi-square methods.

The zerocrossing quantization error (ZCQE) is recorded as ZCQE = $\{e_{q_1}, e_{q_2}, e_{q_3}, \ldots, e_{q_N}\}$

where $e_q = (t_i' - t_{iq}')$.

Here $t_i' = i^{th}$ zerocrossing time before quantization

t_{iq}' = quantized value of i^{th} zerocrossing time.

The maximum and minimum values of the ZCQE is determined and the difference is divided into 'k' class intervals. Then the number of e_{q_i} s falling in each class interval is determined and the histogram is obtained. Fig.3.13 is the histogram obtained for the ZCQE for a speech segment.

In order to test the 'goodness of fit' for different distribution functions, the mean, variance and standard deviations of the ZCQE are computed. The expected value

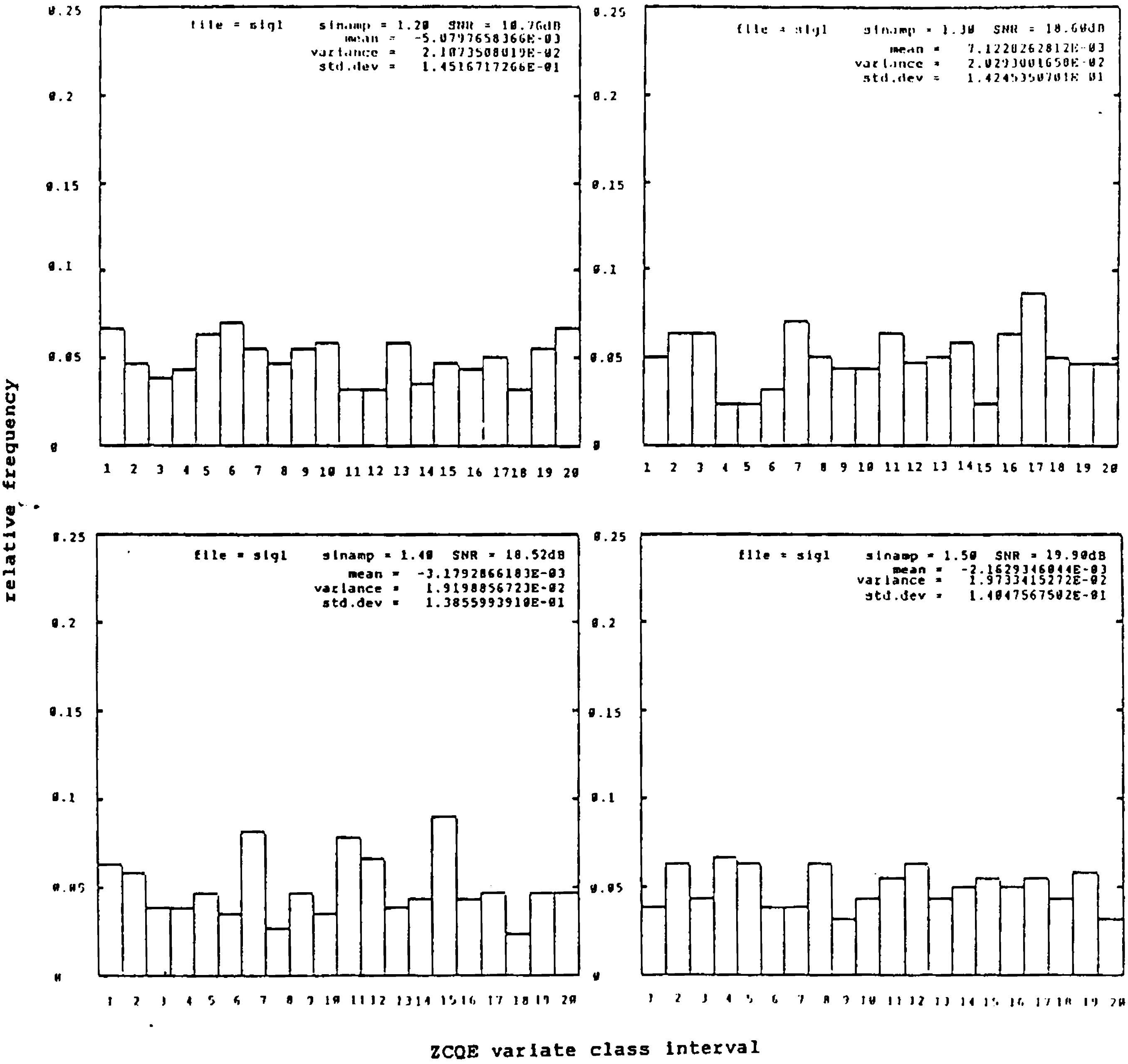


Fig. 3.13 HISTOGRAM for zerocrossing quantization error (ZCQE)

of the relative frequency E_j of the ZCQE in each class j for different distribution functions is computed and the observed value of the relative frequency O_j in each class is obtained directly from the histogram. The chi-square value is computed using equation

$$\chi^2 = \sum_{j=1}^{k} \frac{(E_j - O_j)^2}{E_j}$$

Figs.3.14(i-iv) present the χ^2 values for the five distribution functions for which the 'goodness of fit' test is applied vs. the amplitude of the added high frequency sinusoid. The result shows that the distribution of the ZCQE is uniform (other distributions tried are Gaussian, Laplace, Gamma, Log-normal. These shows poor fit on chi-square test).

A similar 'goodness of fit' test is conducted in the case of the reconstruction error, i.e., the noise introduced by the zerocrossing location based sample estimation method. Figs.3.15(i-iv) present the χ^2 values for the five distribution functions for which the 'goodness of fit' test is applied vs. the amplitude of the added

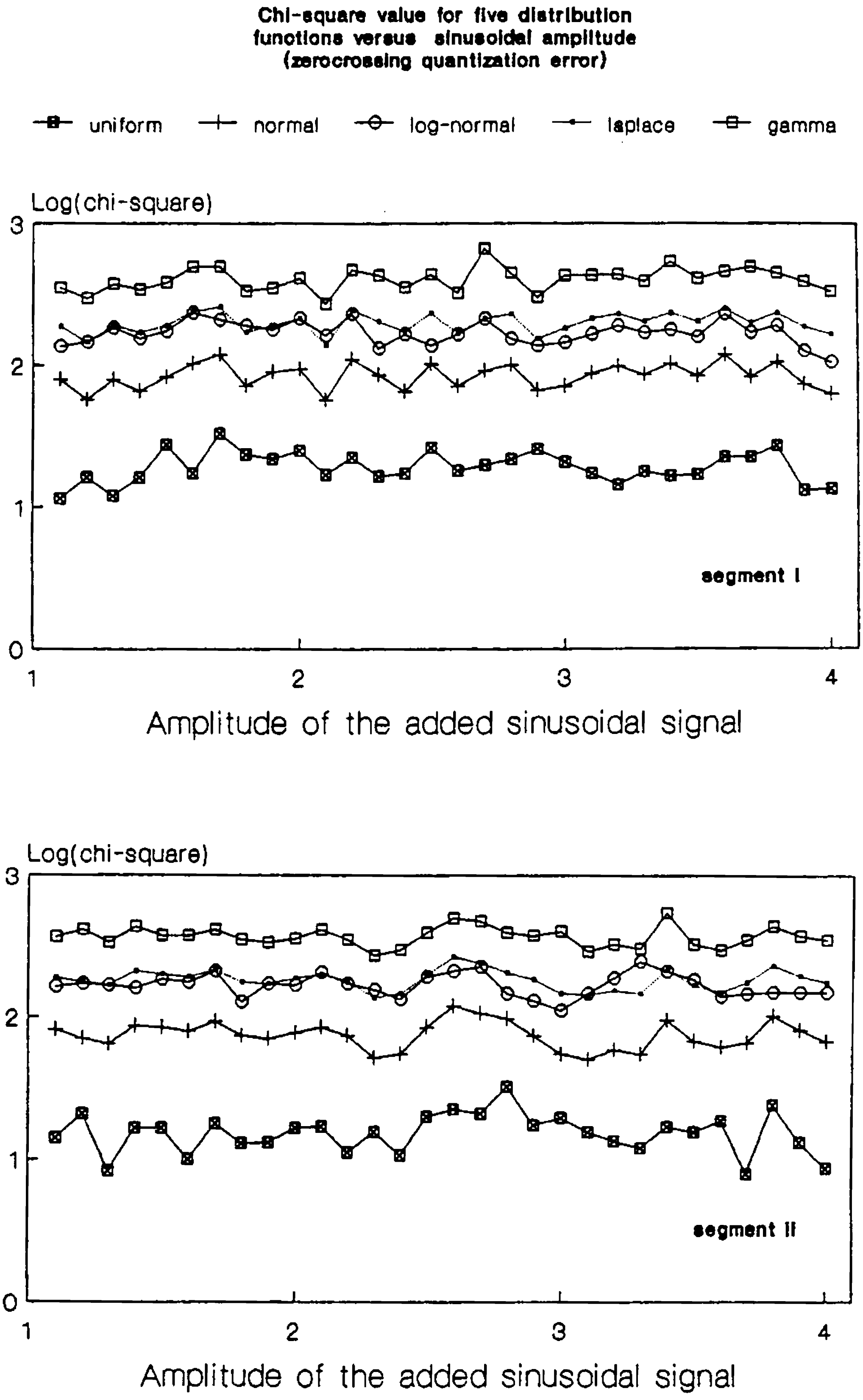

Chi-square value for five distribution
functions versus sinusoidal amplitude
(zerocrossing quantization error)
uniform
normal
log-normal
laplace
gamma
Log(chi-square)
3
2
1
0
1
2
3
4
Amplitude of the added sinusoidal signal
segment I
Log(chi-square)
3
2
1
0
1
2
3
4
Amplitude of the added sinusoidal signal
segment II

Fig. 3.14(i & ii)

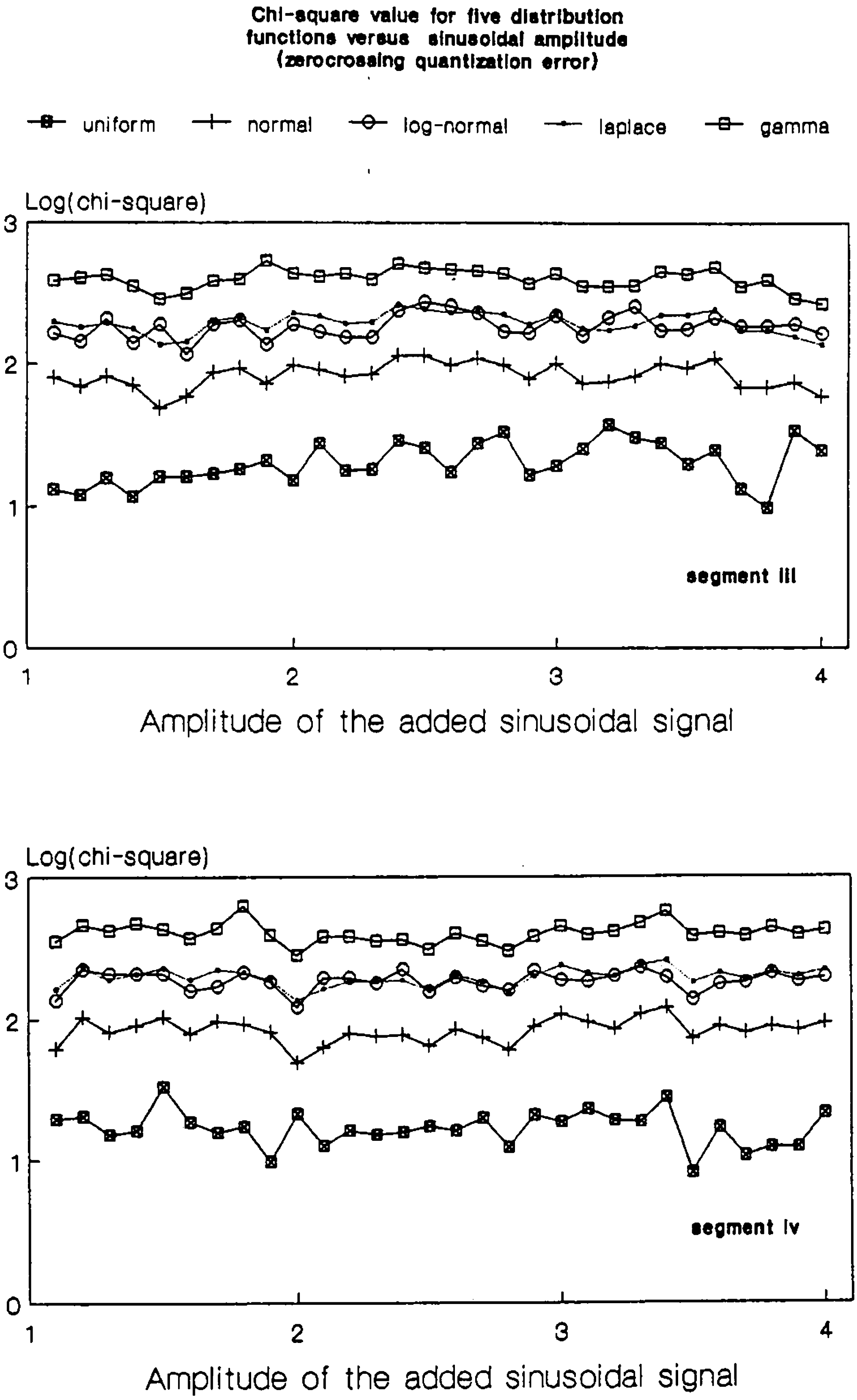


Fig. 3.14(iii & iv)

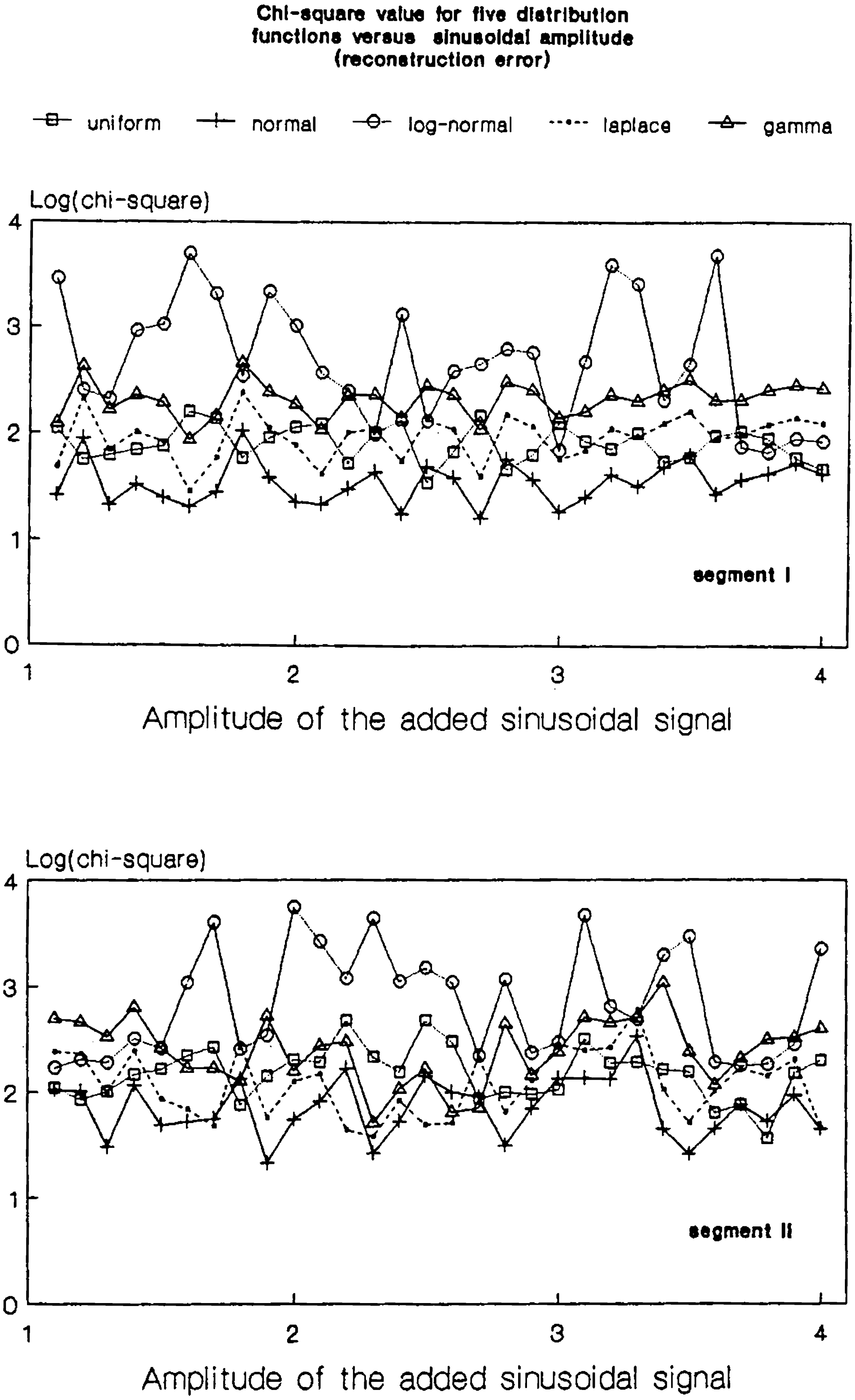

Chi-square value for five distribution functions versus sinusoidal amplitude (reconstruction error)
uniform
normal
log-normal
laplace
gamma
Log(chi-square)
4
3
2
1
0
1
2
3
4
segment I
Amplitude of the added sinusoidal signal
Log(chi-square)
4
3
2
1
0
1
2
3
4
segment II
Amplitude of the added sinusoidal signal

Fig. 3.15(i & ii)

Chi-square value for five distribution functions versus sinusoidal amplitude (reconstruction error)

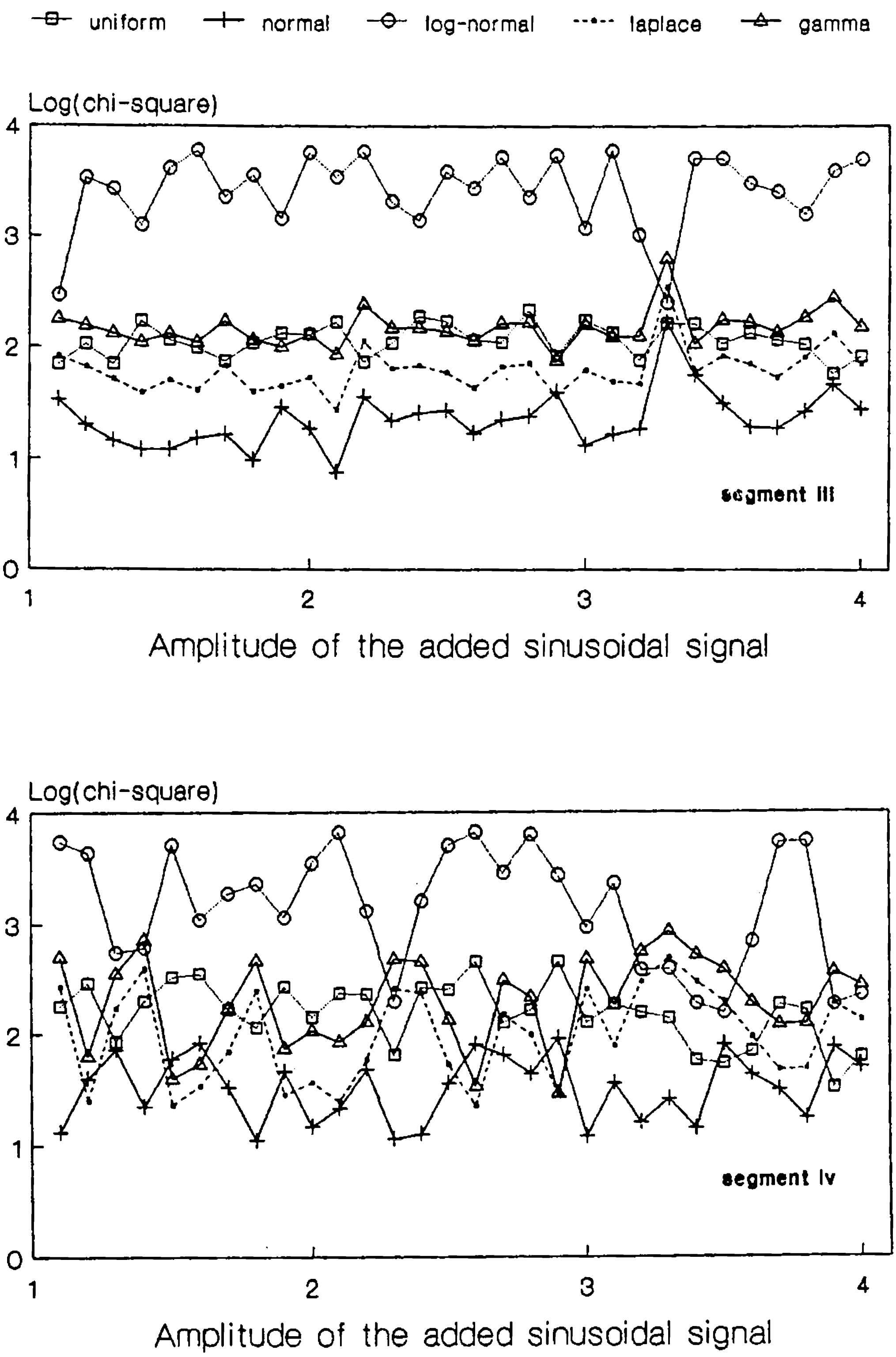


Fig. 3.15(iii & iv)

high frequency sinusoid. Fig.3.16 shows the histogram obtained for the reconstruction error.

It is seen that the Gaussian distribution gives the best fit for the reconstruction error.

In summary, the statistical analysis of the random behaviour of SNR gives the following results.

(a) The distribution of the zerocrossing quantization error is uniform.

(b) Reconstruction error is Gaussian with no influence on distribution of segments or number of bits.

(c) Sinusoidal amplitude variation results in random change in variance of quantization error.

(d) Segment to segment there is slight difference in SNR and quantization error variance.

3.4 CONCLUSIONS

In this chapter we project the use of composite zerocrossings in speech sample estimation. The zerocrossing

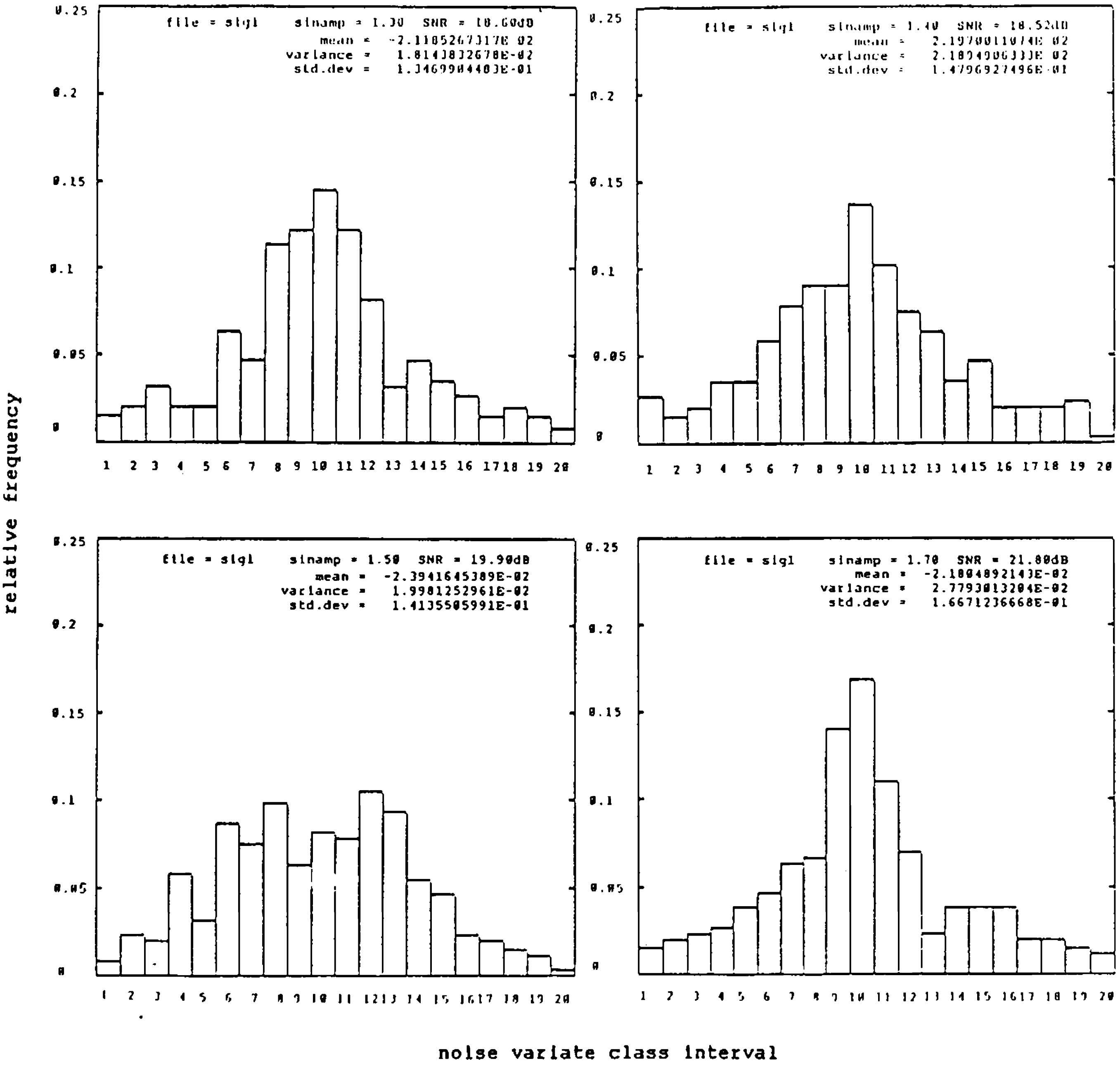


Fig. 3.16 HISTOGRAM for reconstruction error (noise)

based approach requires no sampling of the signal while the conventional A/D converter method relies on multilevel quantization of samples taken at prescribed instants of time. With simple digital circuits it is easier to measure the timings of zerocrossings. The proposed method uses only a simple linear interpolation formula for the estimation of speech samples from zerocrossings. The results of the computer simulation study verifies that the reconstructed speech is of good quality when the zerocrossing location is quantized using 8 bits.

Chapter 4

VOICED/UNVOICED CLASSIFICATION

4.1 INTRODUCTION

At least two processes with significantly different statistics are present in a speech waveform, namely voiced and unvoiced processes. Voiced speech is generated due to the vibration of the vocal cords by forcing air through the glottis. It is a quasi-periodic waveform with highly correlated samples and with high energy (/a/,/i/, /I/,/e/,/ æ/,/ ə/,/u/,/v/,/m/,/n/ etc. are some examples of voiced speech). Unvoiced speech is noise-like and of low energy and low correlation. It is produced by exciting the vocal tract by a steady air flow which becomes turbulent in the region of a constriction in the vocal tract (/f/,/θ /,/s/,/sh/ etc. are some examples of unvoiced speech) [L.R.Rabiner and R.W.Schafer, 1978].

The underlying process of speech production can be modelled using a Markovian generating process with different states for different segments of speech. Fig.4.1 represents a two state Markov model for the speech waveform,

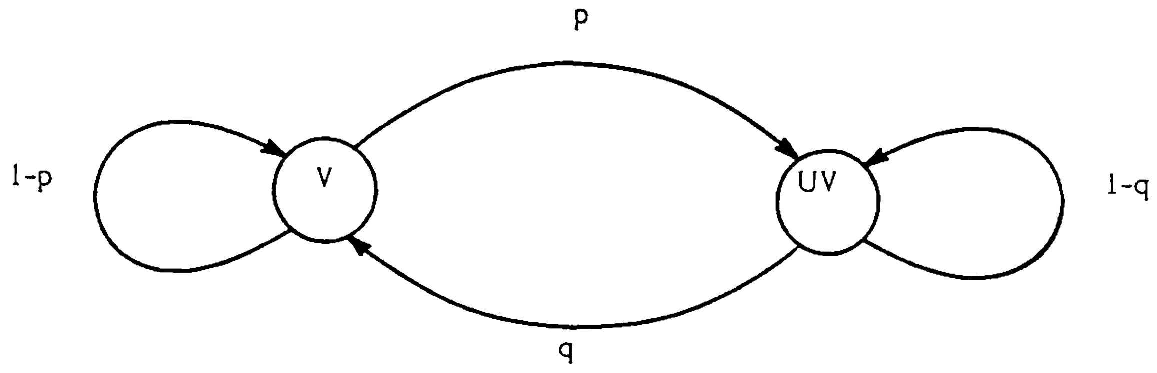


Fig.4.1 A two state Markov model for speech waveform.

with states 'V' for voiced and 'UV' for unvoiced waveform. The transitions from one state to the other state can occur at any instant. Therefore there is a finite probability 'p' of passing from the voiced state into the unvoiced state and a probability 'q' of the opposite transition. The generalization to a multi-state model is straight forward and the experimental results are actually computed for models with three states viz., voiced, unvoiced and silent respectively.

Fig.4.2 gives a binary tree classification scheme for speech waveform proposed by Wiren and Stubbs [J.Wiren and H.L.Stubbs, 1956].

It is evident that the further classification of the speech signal will be perfect only if the voiced/ unvoiced decision is achieved correctly. This is a difficult problem in speech analysis. In this chapter we present two methods for voiced/unvoiced decision. The first one is based on the short time zerocrossing rate and short time energy of the signal and the second one is based on the second order attractor dimension and second order Kolmogorov entropy of speech signal.

4.2 ALGORITHM BASED ON SHORT-TIME ZEROCROSSING RATE AND SHORT-TIME ENERGY

Many algorithms are available in the literature for voiced/unvoiced detection. The main idea of all these algorithms is to find different features of speech signals that can help in voiced/unvoiced decision. Atal and Rabiner considered the voiced/unvoiced classification problem as a pattern recognition problem [B.S.Atal and L.R.Rabiner, 1976]. They have considered five features like energy, zerocrossing rate, correlation coefficient, L.P.C. predictor

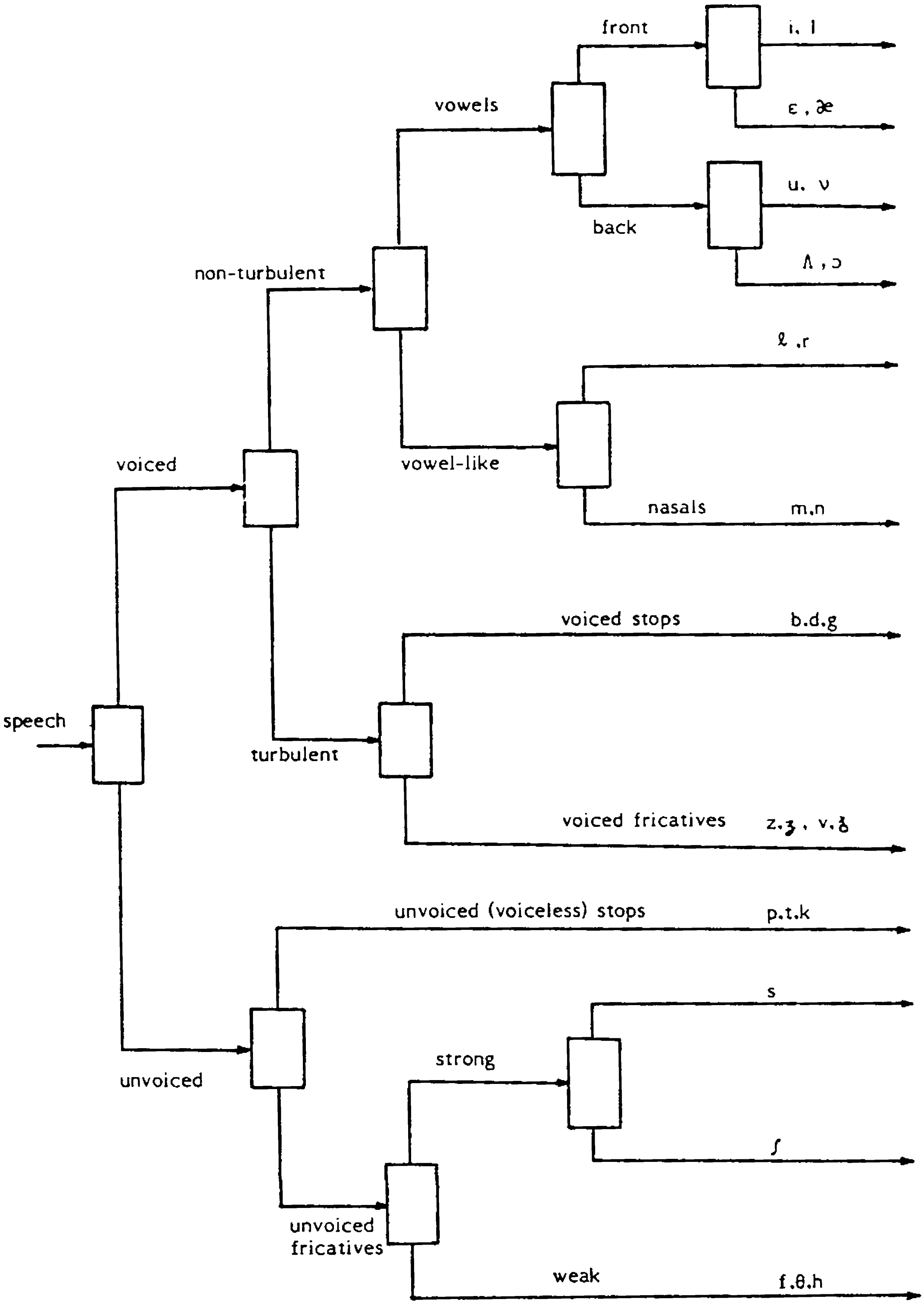


Fig. 4.2 The binary classification scheme of Wiren and Stubbs [1956]

coefficients and predictor error energy for deciding the speech segment as voiced or unvoiced. Rabiner and Sambur proposed an L.P.C. distance measure for voiced-unvoiced-silence detection [L.R.Rabiner and M.R.Sambur, 1979]. Knoor presented another technique for voiced/unvoiced classification by filtering the speech and comparing the rectified filter outputs [S.Knoor, 1979]. In all these papers mentioned above, the selection of features is mainly on the basis of the knowledge acquired from various trials and hence requires involved computation.

The distance measure that we present here for voiced/unvoiced classification is a function of Short-Time Zerocrossing Rate (STZCR) and Short-Time Energy (STE) of speech segments. The facts that the STZCR is larger for unvoiced speech than voiced speech and also, the STE is lesser for unvoiced speech compared to voiced speech are made use of in defining this distance measure [V.Ramamoorthy, 1980].

Let us consider a set A with voiced segments as its elements and another set B with unvoiced segments as elements. This is represented by Fig.4.3. Clearly $A \cap B = \emptyset$, where $\emptyset$ represents the Null set.

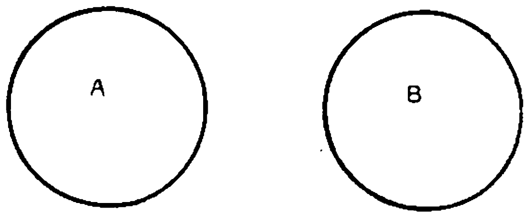


Fig.4.3.

Normally the distribution of both STZCR and STE for voiced speech overlaps to a certain extent with that for unvoiced speech [B.S.Atal and L.R.Rabiner, 1976], [L.R.Rabiner and R.W.Schafer, 1978]. i.e., there can be low-level voiced speech segments with STE comparable to that of unvoiced segments. Also there can be voiced speech segments with STZCR comparable to that of unvoiced segments. Therefore a detection procedure based either only on STZCR or STE leads to large amount of error because of the spread in the distribution of these features.

Let us assume that the STZCR of voiced segments form a set C and that for unvoiced segments form another set D. Because of the finite probability of voiced and unvoiced segments having comparable STZCR values [L.R. Rabiner and R.W.Schafer, 1978] we can represent C and D using Fig.4.4 and clearly $C \cap D \neq \emptyset$.

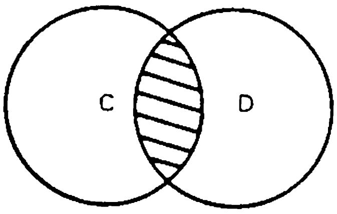


Fig.4.4.

Let x_1, x_2,...,x_n be the speech segments that fall in the shaded area of Fig.4.4. Let these segments form a set X, i.e., X is a set whose elements are voiced and unvoiced segments having comparable STZCR values.

Now let us also assume that the STE of voiced segments form a set E and that for unvoiced segments form another set F. Because of the finite probability of voiced and unvoiced segments having comparable STE values, we can represent E and F by Fig.4.5 and clearly $E \cap F \neq \emptyset$.

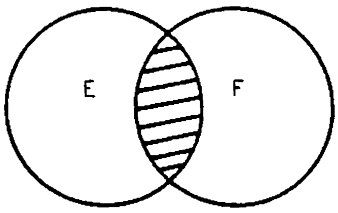


Fig.4.5.

Let $y_1, y_2, \ldots, y_m$ be the speech segments that fall in the shaded area of Fig.4.5 and Y be a set with these segments as its elements, i.e., Y is a set whose elements are voiced and unvoiced segments having comparable STE values.

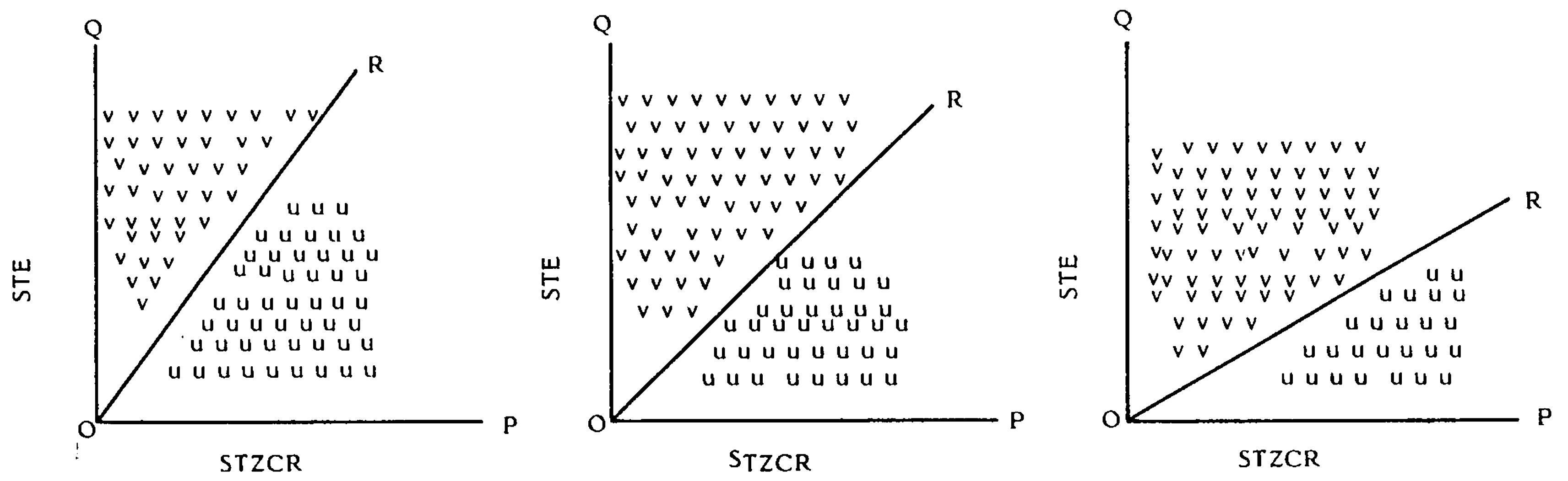
If STE of voiced and unvoiced segment have nearly equal values, their STZCR cannot have comparable values because of their distinct features. Similarly if the STZCR of a voiced and unvoiced segment have nearly equal values, then their STE cannot have comparable values. This characteristics lead to the assumption that, $X \cap Y = \emptyset$. These principles are utilized in defining the new distance measure as a function of STZCR and STE as

$$D = \frac{STZCR}{STE} \tag{4.1}$$

Now let us form two more sets V and U with elements as the values of D for voiced and unvoiced segments respectively. The numerical values of the elements of V will be much less compared to those of the elements of U. The probability of overlapping V with U is very small so that we can assume $V \cap U = \emptyset$. This assumption clearly enables us to define a threshold value for D to classify the voiced/unvoiced segments. i.e., voiced/unvoiced classification can be performed by comparing the value of D for a particular segment with a threshold value. But because of the large dynamic range and the high speaker dependence of speech signal, the value of the elements of the sets V and U varies considerably. Hence a universal constant threshold is not possible and requires an adaptive threshold. This idea is illustrated in Fig.4.6, where OR represents the threshold boundary.

The threshold we use here is relative to the recent most minimum value (RMMV) of D. The RMMV is defined as the lowest value of D observed in the most recent voiced state (The number of segments n, considered in determining the RMMV of D is 12). The effective value of this adaptive threshold 'T_D' can be obtained as

$$T_D = k \times (\text{RMMV}) \text{ of } D,$$



(a) speaker 1

(b) speaker 2

(c) speaker 3

Fig. 4.6 Illustration for the speaker dependence of the threshold boundary (v - voiced, u - unvoiced).

where k is a constant. The value of k is experimentally found to be equal to 150 (The values of n and k are estimated by trial and error method to minimize the error in V/UV decision inside the speech data that is used in the simulation experiment to train the algorithm. The algorithm is validated using speech data outside the training data which will be discussed in the next section). Decisions made from observing the crossing of this adaptive threshold value 'T_D' are essentially independent of the signal dynamic range and type of speakers since the RMMV is a speaker and speech dependent parameter.

In addition to the voiced/unvoiced classification based on the value of D a preliminary test is carried out on each segment of speech to classify the silence based on the value of STE. For this a threshold STE level 'T_S' is obtained from the background noise. If the STE is less than this threshold, the segment is classified as silent. The flow chart of the detection algorithm formed based on STZCR and STE is shown in Fig.4.7. The initial value of T_D is determined by trial and error method so that the first segment is classified correct.

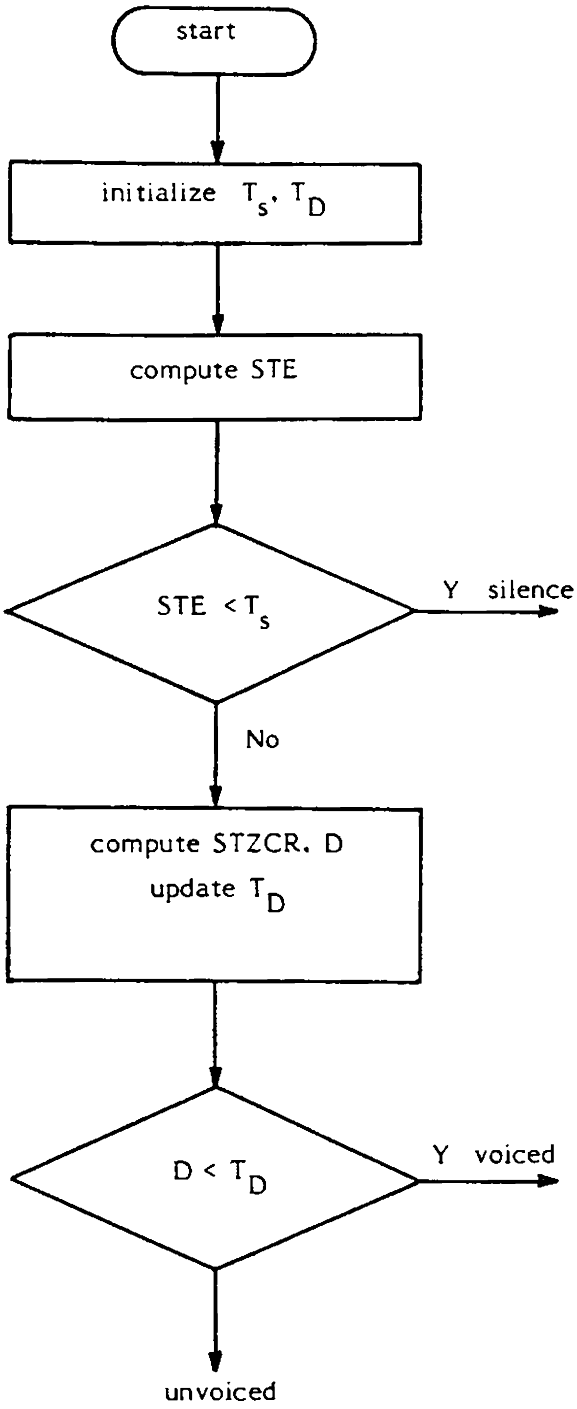


Fig. 4.7 Flow chart of the v/uv detection algorithm based on STZCR and STE. (The initial values of T_s and T_D are experimentally obtained to be equal to 10^{-5} and 9×10^{5} respectively)

4.2.1 Simulation experiment and results

Two sets of speech samples are used in this simulation experiment,--(i) for training the algorithm, and (ii) for validating the algorithm.

The speech data used for training the algorithm is of a sentence "An icy wind racked the beach" [N.S.Jayant, private commn.] spoken by two speakers, a male and a female. This is a 4 kHz band limited noise free signal sampled at 8 kHz rate and quantized to 16 bits. The normalized speech sequence is divided into 256 sample blocks [total number of blocks = 119]. These blocks are manually classified into voiced, unvoiced and silent blocks by plotting on a graphics VDU of the computer. The algorithm is implemented on the PC using Turbo Pascal routines. To compute the STZCR the number of zerocrossings in each block of length 256 samples is found as 'zccount'. Now the STZCR is computed using the formula

$$\text{STZCR} = \text{round}\left(\frac{f_s \text{ x zccount}}{256}\right)$$

where f_s is the sampling rate.

To find STE, the maximum amplitude value in the speech material on test is found and the signal is normalized. The variance of each block of length 256 samples of the normalized speech data is computed as STE.

Table 4.1 gives the computed values of STZCR, STE and D for few segments which are initially manually classified into voiced-unvoiced and silent segments. All segments of the training sequence are correctly classified by this detection algorithm when n = 12 and k = 150. The validity of this algorithm is studied using a speech data base outside the training sequence.

Several utterances spoken by a male and a female speaker of duration 45 seconds are used to create the speech data base to validate the algorithm. They are denoted by the letters Fl, Ml etc. as shown in the table 4.2. These sentences were chosen since they are phonetically well balanced sentences. The speech waveform band limited to 4 kHz and digitized using a 12 bit A/D converter at a sampling rate of 8 kHz is stored in the data base. Each of the utterances in Table 4.2 is divided into 256 sample blocks. Each of these blocks is assigned as voiced

Table 4.1 Computed values of STZCR, STE and D for voiced, unvoiced and silent segments

Segment	No.	STZCR	STE	D
Voiced	1	563	0.00969	$5.81\text{x}10^4$
	2	2500	0.05011	$5.00\text{x}10^4$
	3	1406	0.01457	$9.65\text{x}10^4$
	4	719	0.06408	$1.12\text{x}10^4$
Unvoiced	1	4594	0.00031	$1.5\text{x}10^7$
	2	4469	0.00024	$1.89\text{x}10^7$
	3	3844	0.00013	$2.94\text{x}10^7$
	4	4500	0.00029	$1.55\text{x}10^7$
Silent	1	938	$< 10^{-5}$	$4.47\text{x}10^8$
	2	1313	$< 10^{-5}$	$4.75\text{x}10^8$
	3	1531	$< 10^{-5}$	$1.80\text{x}10^9$
	4	1250	$< 10^{-5}$	$2.10\text{x}10^9$

Table 4.2

Utterance	Male	Female
An icy wind racked the beach	M1	F1
The pipe began to rust while new	M2	F2
cats and dogs hate each the other	M3	F3
Oak is strong and also gives shade	M4	F4
Thieves who rob friends deserve jail	M5	F5
Open the crate but do not break the glass	M6	F6
Add the sum to the product of these three	M7	F7
Joe brought a young girl	M8	F8
A lathe is a big tool	M9	F9

or unvoiced by manual inspection of the waveform on the VDU of the computer. V/UV detection is then carried out using the algorithm and the results are compared with that of manual classification. It is observed that among the 1391 sample blocks, 38 sample blocks are wrongly classified. Therefore the measured total error probability of the algorithm = $\frac{38}{1391}$x 100 = 2.73%.

To summarize, a simple time-domain algorithm for V/UV detection of speech is presented in this section. The algorithm is a simple threshold detection procedure. An adaptive threshold is employed, as the algorithm takes into account the dynamic range of the speech signal and the type of speakers. This in turn reduces the error probability of this detection algorithm.

4.3 ALGORITHM BASED ON THE SECOND ORDER ATTRACTOR DIMENSION AND SECOND ORDER KOLMOGOROV ENTROPY OF SPEECH SIGNALS

Recently much effort has been devoted to the study of the chaotic or turbulent behaviour seen in physical systems. There is a growing interest in the modelling and explanation of apparently stochastic phenomena by the deterministic mechanism known as strange attractors and deterministic chaos [H.Atmanspacher and H.Scheingraber, 1986], [R.H.T.Bates and A.R.Murch, 1987].

The dynamics of a system can be experimentally studied by extracting two invariant parameters from the experimental time series data [H.Atmanspacher and H.Scheingraber, 1986]. They are: (1) The dimension of the attractor of the system in phase space, and (2) The K entropy which is connected with the evolution of the system in phase space.

These invariants are meant to be temporal invariants under constant boundary conditions. They may change if some control parameter of the signal is varied. Atmanspacher and Scheingraber [H.Atmanspacher and H.Scheingraber, 1986] have reported a method to determine these two invariants from the measurements of the time series of a single variable of the system in the context of the study of dynamical instabilities in multimode CW dye laser. The same method is extended to study the nature of the attractor underlying the production of speech signals in the following sections. Based on the experimental results a new distance measure is introduced for voiced/unvoiced classification.

4.3.1 Estimation of dimensions and entropies from a time series data

Consider a time series data sequence,

$$\{X(n)\} = [X(1), X(2), \ldots, X(N)] \tag{4.2}$$

The extraction of the second order dimension and the second order entropy from $\{X(n)\}$ is possible using a correlation integral [H.Atmanspacher and H.Scheingraber, 1986] which is defined as

$$C(r) = \lim_{N \to \infty} \frac{1}{N^2} \sum_{i,j=1}^{N} H(r-|X_i - X_j|) \tag{4.3}$$

where H is the Heaviside function, $H(x) = 0$ for $x \leq 0$ and $H(x) = 1$ for $x > 0$. The function $C(r)$ counts the number of pairs of points with a distance $|X_i - X_j|$ smaller than 'r'. Therefore when the distance between all the pairs of points is less than 'r' then $C(r) = 1$.

The method of estimating the second order attractor dimension and the second order entropy using the correlation integral (4.3) can be easily realized, if we construct 'd' additional data sets from the original time series data sequence $\{X(n)\}$ by introducing a time delay $d \Delta t$.

For the uniformly sampled time series data, with sampling period T, Δt = T. From the resulting data sets, 'd' dimensional phase space can be constructed such that 'd' is greater than the dimension of the actual phase space.

If each data set contains (N+d) values spaced by a time increment T, then the following data sets can be obtained for various values of 'd',

$$
\begin{array}{l}
X(1),\ X(2), \ldots\ldots\ldots\ldots\ldots X(N) \\
X(2),\ X(3), \ldots\ldots\ldots\ldots\ldots X(N+1) \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
\vdots \\
X(d+1), X(d+2), \ldots\ldots\ldots\ldots X(N+d)
\end{array}
$$

This will yield N data vectors of the type,

$$
\begin{array}{lcl}
\mathbf{X}_1 & = & [X(1)\ X(2) \ldots\ldots\ldots\ldots X(d+1)]^T \\
\mathbf{X}_2 & = & [X(2)\ X(3) \ldots\ldots\ldots\ldots X(d+2)]^T \\
\ldots\ldots & & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
\vdots & & \\
\mathbf{X}_N & = & [X(N)\ X(N+1) \ldots\ldots\ldots\ldots X(N+d)]^T
\end{array}
$$

Or a vector set $\mathbf{X}_1, \mathbf{X}_2, \ldots\ldots \mathbf{X}_N$ (4.4)

Now, for each j we can take a point X_j from (4.4) and determine the distance $|X_i - X_j|$ which is the usual Euclidean norm. In this manner we can determine the number of pairs of points whose distance is smaller than a given distance 'r'. Using this result we can directly compute the correlation integral C(r) given by (4.3). C(r) is the basic quantity needed for the further determination of the attractor.

The second order attractor dimension is defined as,

$$D_2 = \lim_{r \to o} [\log C(r)/\log r] \qquad (4.5)$$

When D_2 is an integer the system is regular, and when it is fractal the system is chaotic, while $D_2 \longrightarrow d$, the dimension of the constructed phase space, the system behaviour is stochastic.

Similar to the second order dimension D_2, a second order entropy K_2 can be defined as

$$K_2 = \lim_{r \to 0, d \to \infty} \frac{1}{T} \log[C_d(r)/C_{d+1}(r)] \qquad (4.6)$$

where the logarithms are to the base 2. K_2 is the more sensitive parameter than D_2. $K_2 = 0$ characterises a regular system and $K_2 > 0$, corresponds to a chaotic system while $K_2 \rightarrow \infty$ describes a completely stochastic system.

4.3.2 Simulation experiment

Turbo Pascal routines were implemented on a 3AT6 computer to compute C(r). Here too speech data blocks of 256 samples that are manually classified into voiced/ unvoiced segments are used for simulation. Using the speech data the correlation coefficients C(r) for several 'r' with respect to each particular dimension 'd' is computed. Fig.4.8 shows a log-log plot of C(r) vs. r. It may be noted that with the increase in dimensions the slope at the linear portions of these curves converge to a limiting value. Fig.4.9 shows the slope 'm' in the linear range of the different curves vs. the dimension 'd'. The limiting value of the slope corresponds to the second order dimension D_2 of the attractor.

Fig.4.10 represents the plot showing mean value of $\log[C_d(r)/C_{d+1}(r)]$ obtained from the linear range of the curves in Fig.4.8, vs. 'd'. The second order entropy K_2 is obtained as the product of the limiting value of $\log[C_d(r)/C_{d+1}(r)]$ and (1/T). Now we can define a distance

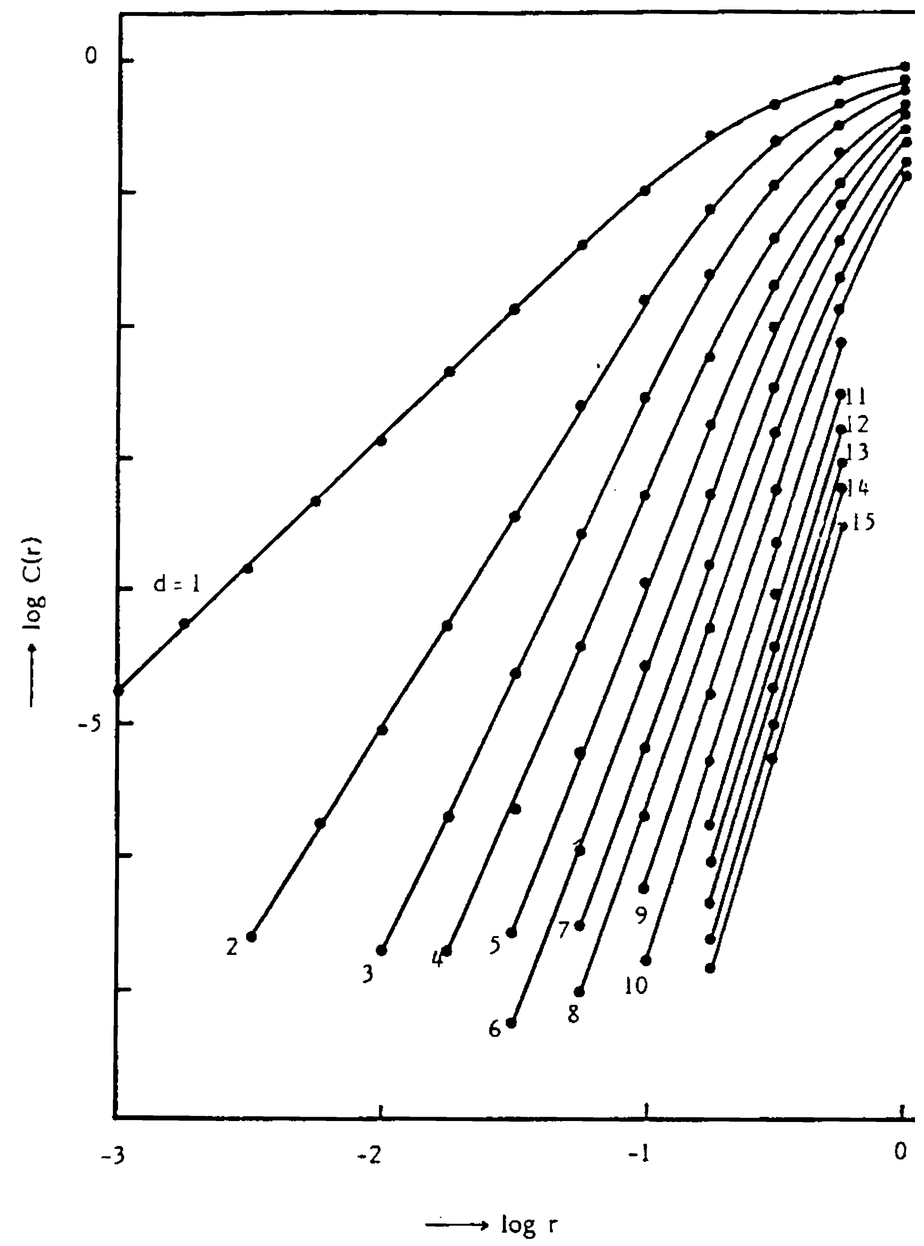


Fig. 4.8 Log-Log plot of the correlation integral C(r) versus the distance r for a speech segment.

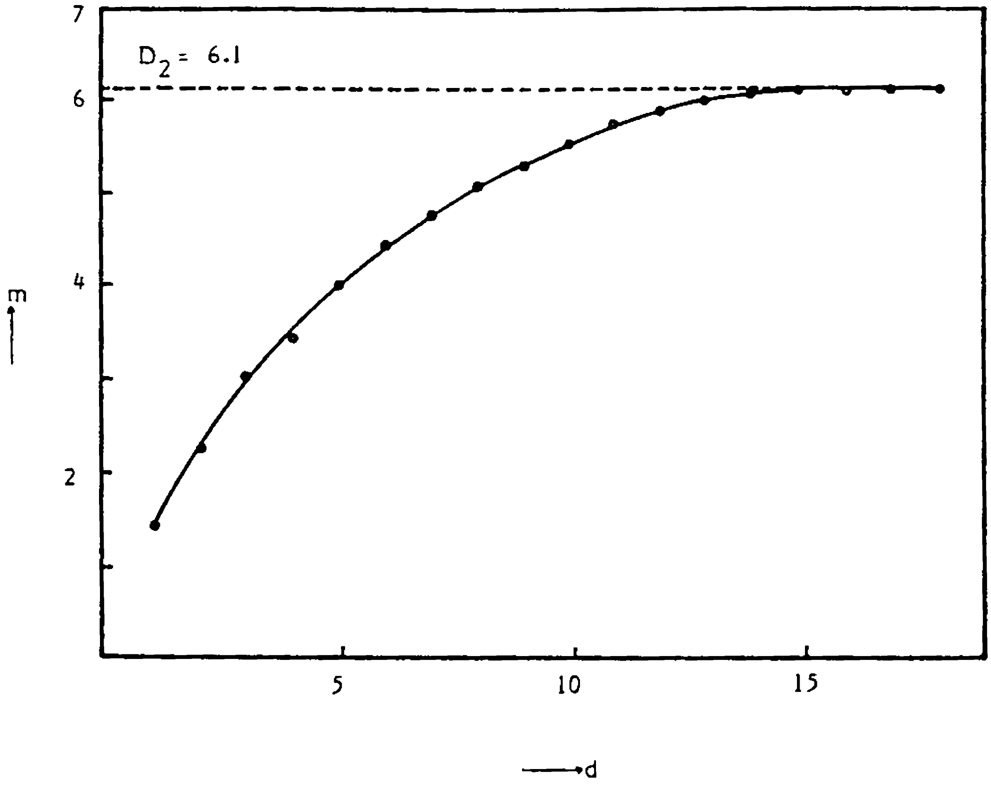


Fig. 4.9 The slope 'm' in the linear range of the different curves (in Fig. 4.8) versus the dimension 'd'.

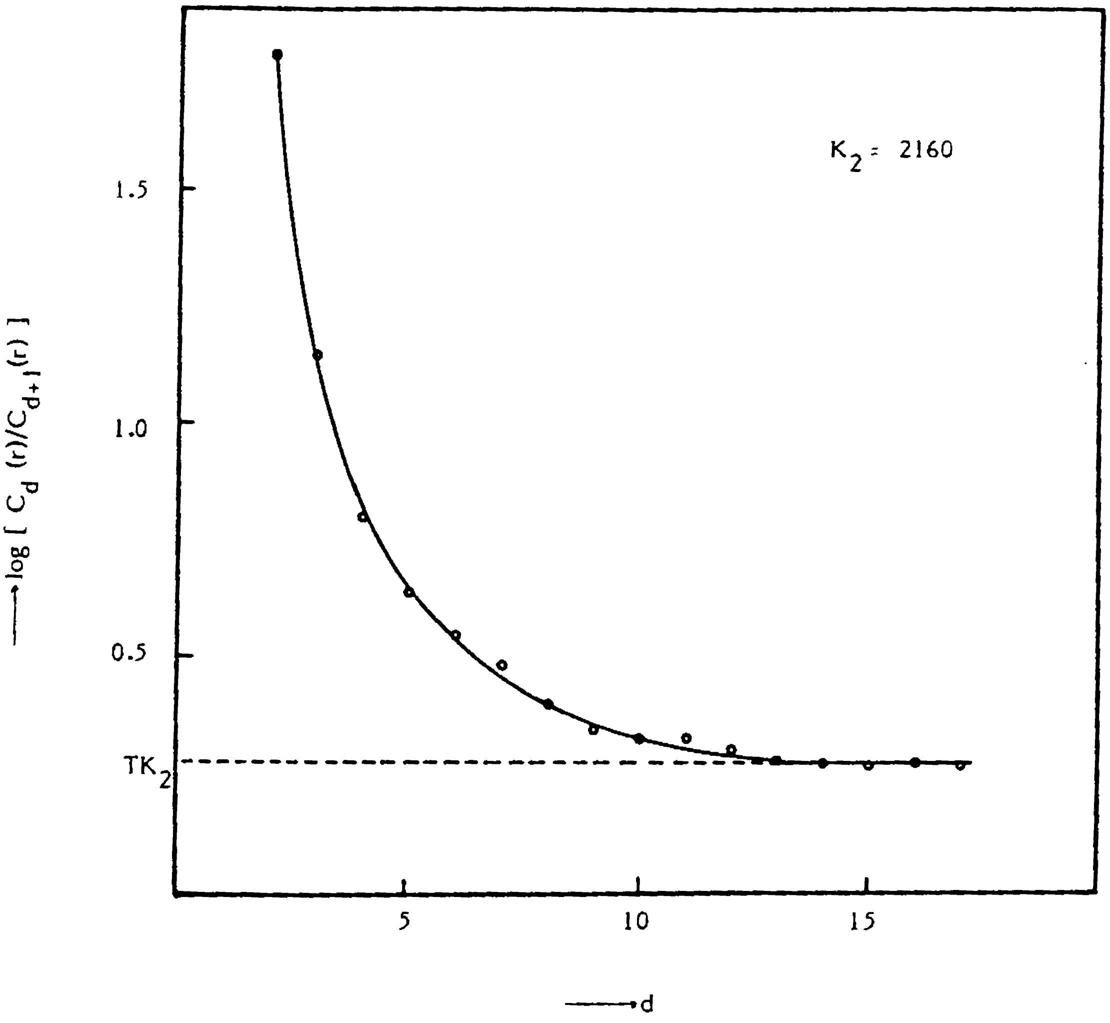


Fig. 4.10 Mean value of log [C_d (r)/C_{d+1} (r)] as a function of d, obtained from the linear range of the curves in Fig. 4.8

measure known as the chaotic distance (D_{CH}) as the product of D_2 and K_2

i.e., $$D_{CH} = D_2 \times K_2 \quad (4.7)$$

Table 4.3 gives the values of the second order attractor dimension D_2, second order Kolmogorov entropy K_2 and the chaotic distance D_{CH} for few segments of speech which are manually classified into voiced and unvoiced segments. It may be noted that there exists a large difference between the values of D_{CH} for voiced and unvoiced segment. Therefore a decision threshold for D_{CH} can be fixed for V/UV classification by inspecting a large and, phonetically balanced speech data base. For this the 1510 sample blocks in the two sets of speech samples used in section 4.2.1 (i.e., the training sequence and validating sequence) are used. The values of D_2, K_2 and D_{CH} for all the segments are computed. A threshold T_{CH} is said to be optimum if the resulting detection procedure minimizes the number of wrong classification. It is found that minimum number of segments are wrongly classified when T_{CH} = 27000. At this threshold, the number of segments wrongly classified was only 24, while the number was larger when the threshold

Table 4.3

Segment	No.	D_2	K_2	D_{CH}
Voiced	1	6.10	2160	13176
	2	5.03	1890	9507
	3	5.95	2010	11960
	4	7.05	2350	16568
	5	6.90	2270	15663
Unvoiced	1	10.11	3752	37933
	2	10.54	3870	40790
	3	9.65	3520	33968
	4	11.32	3980	45054
	5	10.68	3915	41812

was either increased or decreased. Therefore the measured total error probability of this algorithm = $\frac{24x100}{1510}$ = 1.59%.

In conclusion, for voiced/unvoiced classification the time domain method of measuring the newly proposed distance measure can be efficiently used. The measured accuracy of the former method is 97.22% while that of the latter method is 98.41%. However the first method is computationally simple and requires only lesser computation time compared to the second.

Chapter 5

SELECTION OF ENCODER AND CLASSIFIER

5.1 INTRODUCTION

In chapter 3 we have introduced a new method for estimating speech samples from their zerocrossings. In chapter 4 we have presented two methods for voiced/ unvoiced classification. One of the methods was based on the short time zerocrossing rate of speech signal. The other method was based on two relatively new concepts in theoretical physics--attractor dimension and Kolmogorov entropy. In this chapter we will evaluate the applicability of zerocrossing information and the attractor dimension and entropy for low bit rate coding based on the results obtained in the previous chapters.

5.2 APPLICABILITY OF ZEROCROSSING INFORMATION FOR LOW BIT RATE CODING

To use zerocrossing information for speech coding is not a new idea. This is because the extraction of zerocrossings is very easy and only simple digital circuits are necessary for this. Many researchers have attempted but not succeeded so far in this venture, to the best of our knowledge. The only result obtained favourably by earlier researchers is the reconstruction of low frequency noisy sinusoid from its composite zerocrossings by Kay and Sudhakar. This method is not suitable for speech signals.

The method we have presented in chapter 3 for reconstruction of speech signal from composite zerocrossings is theoretically supported with a good deal of approximation. In the reconstructed signal the shape of the original waveform is preserved, which is the essential characteristic of a waveform coder. However, the SNR obtained for the reconstructed speech is poor compared to standard PCM coding. But for higher number of bits for zerocrossing quantization--8 bits and above the reconstructed signal is of good quality. That is, we require an increased number of bits for zerocrossing quantization for obtaining reconstructed signal of better quality. This means that low bit rate coding of speech is not possible by the mere use of zerocrossing informations.

However, the results of our study is useful for improving the existing waveform coders. The zerocrossing based sample reconstruction method developed and presented in earlier chapter can be used to design new types of low complexity coders.

A different form of application of the zerocrossing information for speech coding is studied in chapter 4. The main aim of this study was to explore the potentiality of zerocrossing information for voiced/unvoiced classification. A simple time-domain algorithm that utilizes the zerocrossing informations for voiced/unvoiced detection is developed and studied. This algorithm is a simple threshold detection procedure. An adaptive threshold that takes into account the speaker dependency and the dynamic range of speech signal is employed in this algorithm.

The voiced/unvoiced classification is very useful in recognition and low bit rate coding. The voiced/unvoiced classification algorithm developed in chapter 4 is used to design an Adaptive Switching Transform Coder in chapter 7. It is shown that the coder definitely improves the encoded speech quality. Thus it is found that even though the zerocrossing information is not directly useful for bit rate reduction, it can be used to improve the existing waveform coders--both for reduced complexity and improved quality.

5.3 APPLICABILITY OF ATTRACTOR DIMENSION AND KOLMOGOROV ENTROPY FOR LOW BIT RATE CODING

In chapter 4, we have studied the dynamical instabilities and deterministic chaos with respect to speech signal with the prime aim to utilize the results for low bit rate coding. We have obtained the second order attractor dimension D_2, second order Kolmogorov entropy K_2 and the chaotic distance D_{CH} for segments of speech, and found that the values of D_{CH} differ considerably for voiced and unvoiced segments. This implies that we can conduct voiced/ unvoiced classification by extracting the parameter D_{CH} from the signal. And this knowledge can be used to improve the quality of existing waveform coders as discussed in section 5.2. This method can also be utilized for extracting invariant parameters that help recognition, but is beyond the purpose of the present work.

Chapter 6

MODIFIED ADAPTIVE TRANSFORM CODER

6.1 INTRODUCTION

Transform coding is a 'frequency domain' approach like sub-band coding. The efficiency of a transform coding system will depend on the type of linear transform used and the nature of bit allocation for quantizing the transform coefficients. Most practical systems are based on sub-optimal approaches for transform operation as well as bit allocation. To achieve coding efficiency the transform is chosen so that it decorrelates the input samples. And also more bits are assigned to more important transform coefficients and fewer bits to less important coefficients. This is the basic idea in Adaptive Transform Coding.

A number of transformations can be used in transform coding, such as Fourier Transform, Karhunen-Loeve Transform (KLT), Discrete Cosine Transform (DCT), Discrete Walsh-Hadamard Transform (DWHT) and others. KLT is an optimal transform but the difficulties in determining the statistical behaviour of speech makes it impractical.

DCT is a sub-optimal transform in the sense that it is asymptotically equivalent to KLT and is not a data dependent transform. DWHT is simple to implement because it requires only additions and subtractions.

In this chapter a modified adaptive transform coding scheme is studied. Performance of the coder by using both DCT and DWHT are evaluated.

The chapter starts with the description of DCT and DWHT, followed by the evaluation of the theoretical transform coding gain. The proposed modifications on the adaptive transform coding scheme is presented next. Finally, simulation results are given which shows that the coder achieves better performance due to these modifications.

6.2 DISCRETE COSINE TRANSFORM

The DCT of a data sequence $\{X(m)\}$, $m = 0,1,2,..,N-1$ and its inverse are defined as

$$Y_c(k) = \frac{2C(k)}{N} \sum_{m=0}^{N-1} X(m) \cos\left[\frac{(2m+1)k\pi}{2N}\right], \qquad k=0,1,2,...,N-1$$

and

$$X(m) = \sum_{k=0}^{N-1} C(k)Y_c(k) \cos\left[\frac{(2m+1)k\pi}{2N}\right]$$

$$m = 0,1,2,\ldots,N-1$$

respectively where

$$C(k) = \begin{cases} 1/\sqrt{2} & \text{for } k = 0 \\ 1, & \text{for } k = 1,2,\ldots,N-1 \end{cases}$$

$Y_c(k)$, $k = 0,1,2,\ldots,N-1$ is the DCT sequence.

The set of basis functions

$1/\sqrt{2}$, $\cos[(2m+1)k\pi/2N]$ is a class of discrete Chebyshev polynomials.

The DCT was originally proposed in [N.Ahmed, T.Natarajan and K.R.Rao, 1974]. For any finite transform size the DCT is always closer to the optimal KLT [M.Hamidi and J.Pearl, 1976]. Various algorithms have been proposed for the computation of DCT [N.Ahmed, T.Natarajan and K.R.Rao, 1974], [R.M.Haralick, 1976], [W.H.Chen, C.H.Smith and S.C.Fralick, 1977], [M.J.Narasimha and A.M.Peterson, 1978], [J.Makhoul, 1980]. In this work we implement the DCT operation by the algorithm proposed by N.Ahmed, T.Natarajan and K.R.Rao, since it enables a direct utilization of available FFT routines. Essentially, the N-dimensional data block is extended to a 2N block by either appending N zeros to it or by concatenating the original block with its mirror image. Then, a 2N DFT is performed on the extended block and the DCT coefficients are extracted from the first N components of the 2N transform block.

The DCT has been extensively used in image and speech transform coding systems because of its effectiveness in removing the correlation from highly correlated sources. In conventional transform coding systems the decorrelation process is the only task of the DCT and the efficiency of the coder is directly proportional to the degree of decorrelation, measured by the transform gain [N.S.Jayant and P.Noll, 1984, chap.12].

6.3 DISCRETE WALSH-HADAMARD TRANSFORM

The Walsh functions, named after J.L.Walsh, who introduced them in 1923, are the basis functions for the Walsh-Hadamard transform (WHT). They form a complete orthogonal set over a unit interval and can be developed from the Rademacher functions [D.F.Elliott and K.R.Rao, 1982].

Because the Walsh functions are binary valued (±1), their generation and implementation is simple. Fast algorithms have been developed based on sparse matrix factoring of (WHT) matrices [Y.Tadokoro and T.Higuchi, 1979], [C.K.Yuen, 1975]. These algorithms, however, require only addition (subtraction), as compared to the complex arithmetic operations (multiplication and/or addition) required for the FFT. Uniform sampling of the Walsh function of any ordering results in the Walsh-Hadamard matrices of corresponding order. The rows of these matrices represent the Walsh functions in a unique manner. These matrices can be generated using the following recurrence relation:

$$H_h(k) = \begin{bmatrix} H_h(k-1) & H_h(k-1) \\ H_h(k-1) & -H_h(k-1) \end{bmatrix} , \quad k=1,2,\ldots,L$$

where $H_h(0) = 1$ and $L = \log_2 N$.

For example, with k = 1, and k = 2 this yields

$$H_h(1) = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} , \quad H_h(2) = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

Let $\mathbf{X}^T = [X(0),X(1),\ldots,X(N-1)]$ denote an N-periodic data sequence of finite valued real numbers. The discrete Walsh-Hadamard transform and its inverse, respectively, can be defined as

$$\mathbf{Y}_w = (1/N)[H_h(L)]\ \mathbf{X} \text{ and } \mathbf{X} = [H_h(L)]\mathbf{Y}_w$$

where the transform sequence is denoted as

$$Y_w^T = [Y_w(0),\ Y_w(1),\ Y_w(2),\ldots,\ Y_w(N-1)]$$

The transform component $Y_w(m)$ represents the amplitude of $\text{Wal}_w(m,t)$ in a Walsh function series expansion for $\mathbf{X}$. The first component $Y_w(0)$ is the average or mean of X, and the succeeding components represent Walsh functions of increasing sequence.

6.4 TRANSFORM CODING GAIN

The transform coding gain (G_{TC}) indicates the objective improvement achieved by the transform in reducing the number of coefficients required to completely specify the signal being transformed. The transform coding gain G_{TC}, when the transform coefficients $Y(k)$, $k = 1,2,3,\ldots,N$

are quantized independently is defined as [R.Zelinski and P.Noll, 1977]

$$G_{TC} = \frac{Y^2}{\left[\prod_{k=1}^{N} Y^2(k)\right]^{1/N}} \qquad (6.1)$$

where Y^2 is the variance.

We have computed the G_{TC} of Cosine and Walsh-Hadamard transforms using the real speech data mentioned in section 4.2.1. [The data base contains a male and female utterance "An icy wind racked the beach". This is a 4 kHz band. limited noise free signal sampled at 8 kHz rate and quantized to 16 bits]. Figs.6.1(a-e) shows a comparative plot for theoretical G_{TC} for different data block length for both the transforms. It may be noted that the theoretical gain G_{TC} differs for different sounds. Most of the time the G_{TC} for DCT is better than that for DWHT. Few instants are noted for which the Walsh-Hadamard transform gives better G_{TC} than cosine transform.

6.5 ADAPTIVE TRANSFORM CODING (ATC)

The basic block schematic diagram of the Adaptive

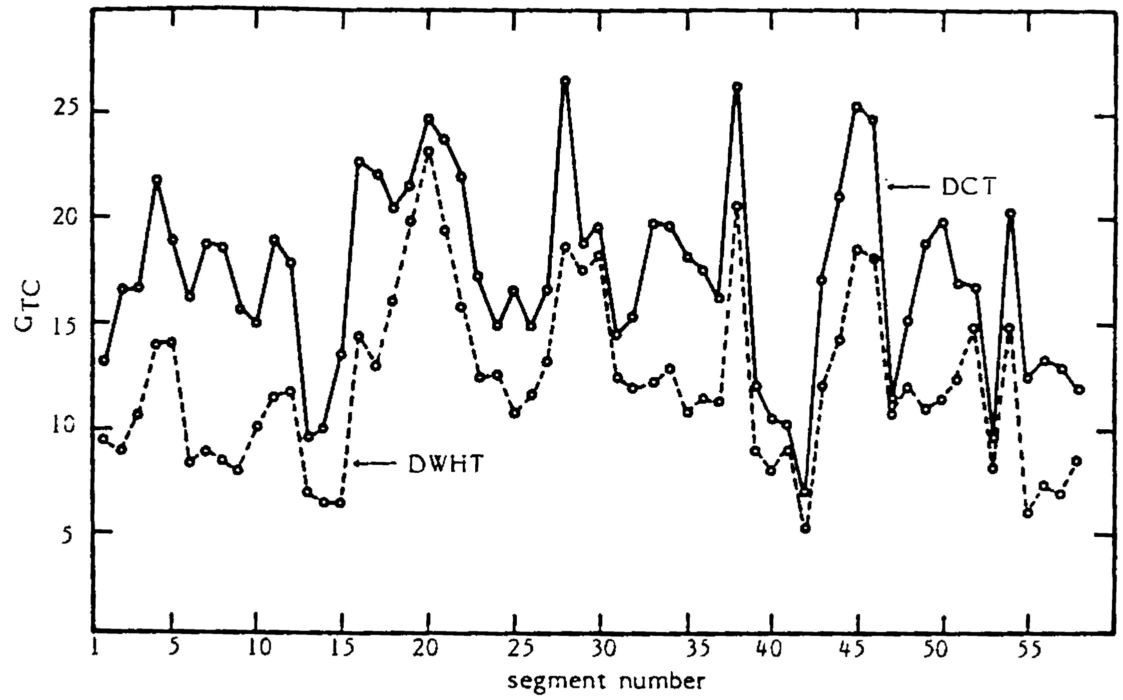


Fig. 6.1(a) Comparative plot for theoretical G_{TC} for DCT and DWHT (data block length = 256)

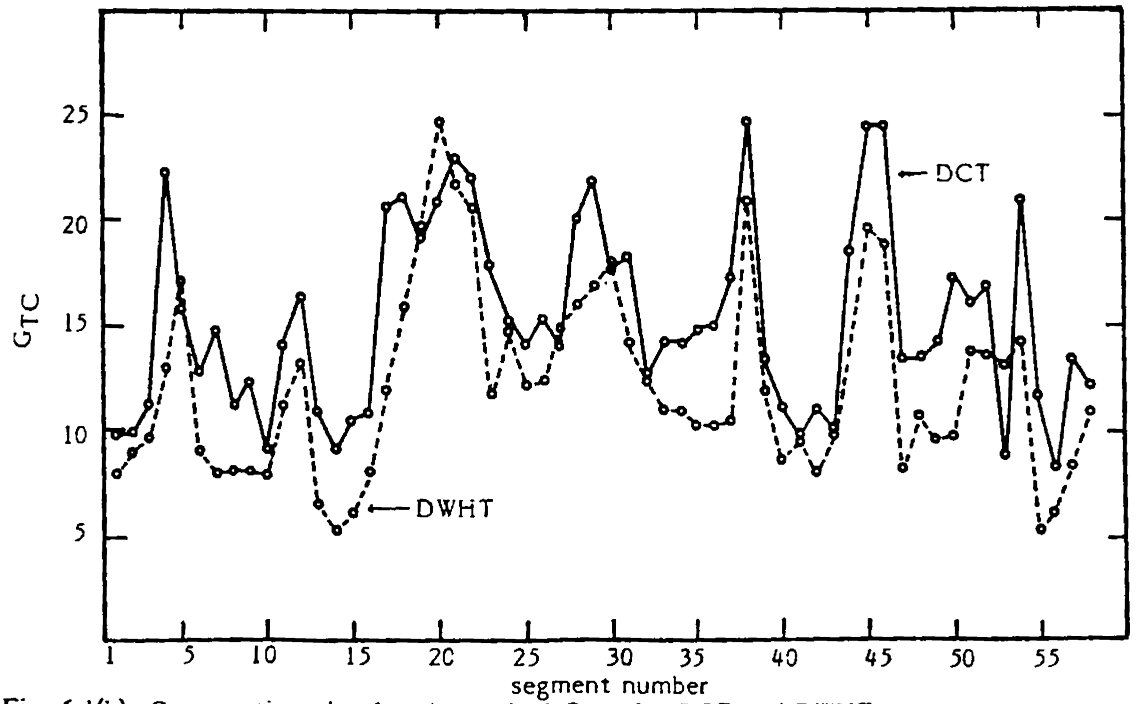


Fig. 6.1(b) Comparative plot for theoretical G_{TC} for DCT and DWHT (data block length = 128)

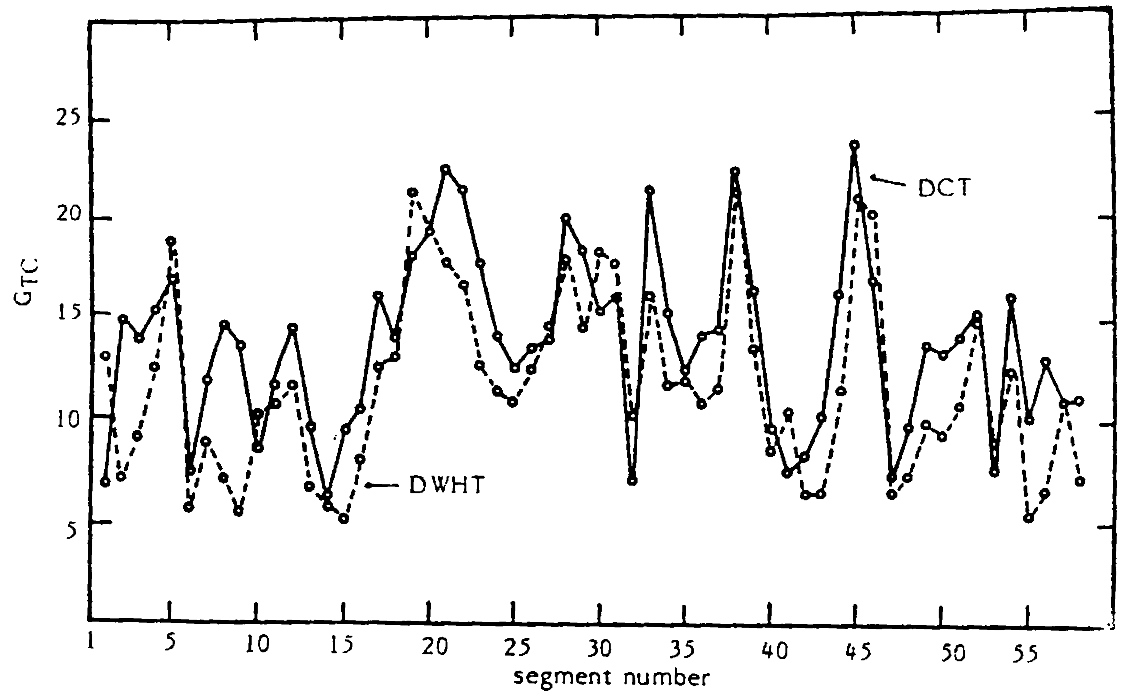


Fig. 6.1(c) Comparative plot for theoretical G_{TC} for DCT and DWHT (data block length = 64)

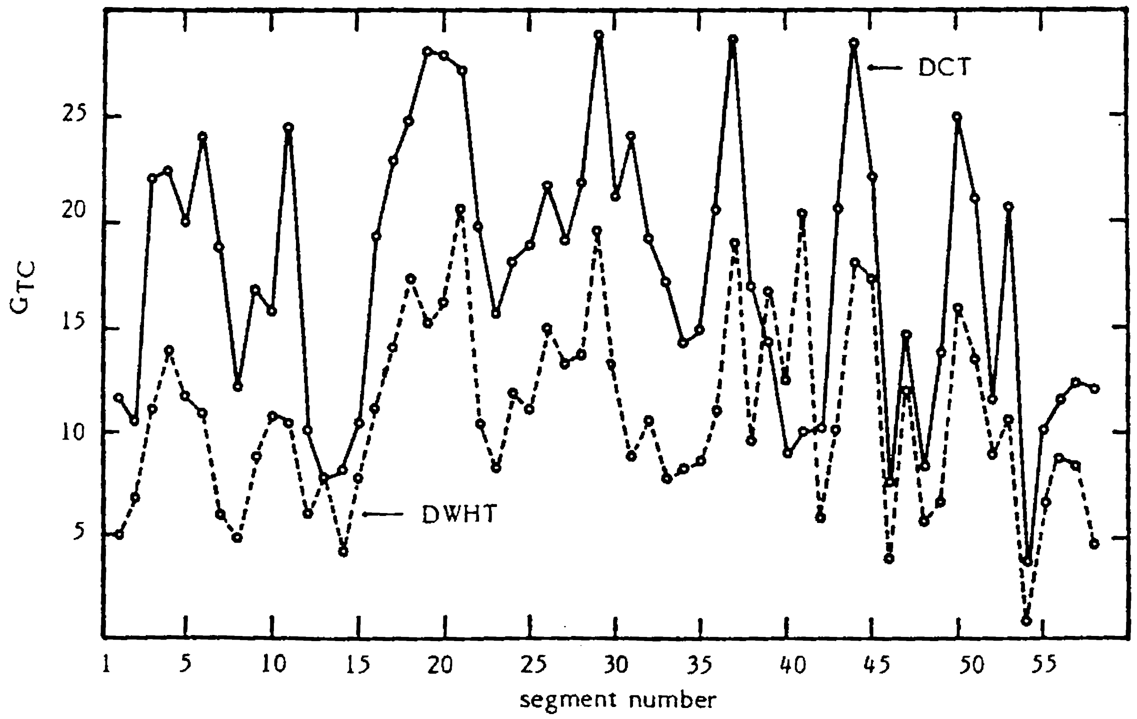


Fig. 6.1(d) Comparative plot for theoretical G_{TC} for DCT and DWHT (data block length = 32)

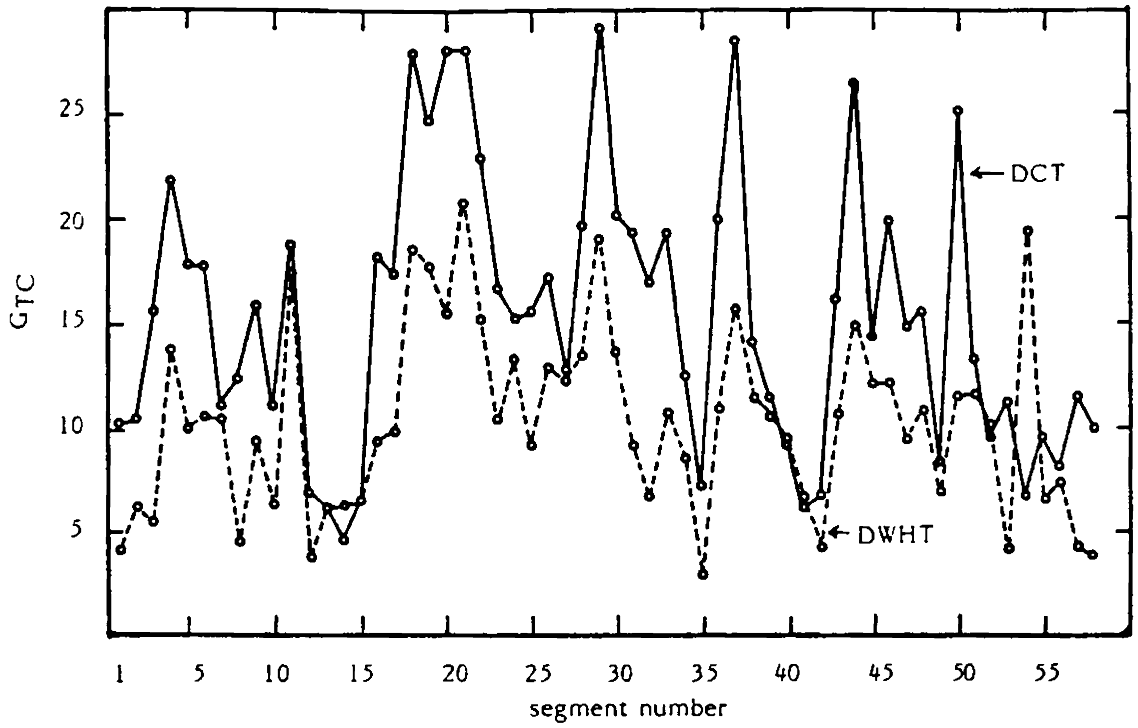


Fig. 6.1(e) Comparative plot for theoretical G_{TC} for DCT and DWHT (data block length = 16)

Transform Coder (ATC) proposed by Zelinski and Noll [1979] is shown in Fig.6.2. The input speech is buffered into successive blocks of size N = 128-256 each. This input block $\{X(n)\}$ is transformed into $\{Y(n)\}$. The transformed coefficients $Y(n)$ are then adaptively quantized and transmitted to the receiver At the receiver they are decoded and inverse transformed into blocks $\{\hat{X}(n)\}$. These blocks are then used to synthesize the output speech signal by a concatenation of the blocks.

A major concern in designing a good ATC coder is the adaptation of the bit allocation and the quantizer stepsize to the changing statistics of the transform coefficients. It is well-known that the distribution of the transform coefficient variances is a crucial factor in determining the transform gain and hence, the coder performance. The potential transform gain can only be realized if an accurate description of the variance distribution is available to both the transmitter and the receiver. This however, requires a significant amount of side-information whose information rate depends on how well the variance patterns are coded. To reduce the side information rate, the designers of conventional ATC's have usually chosen to transmit only a rather crude description of the variance patterns, thereby

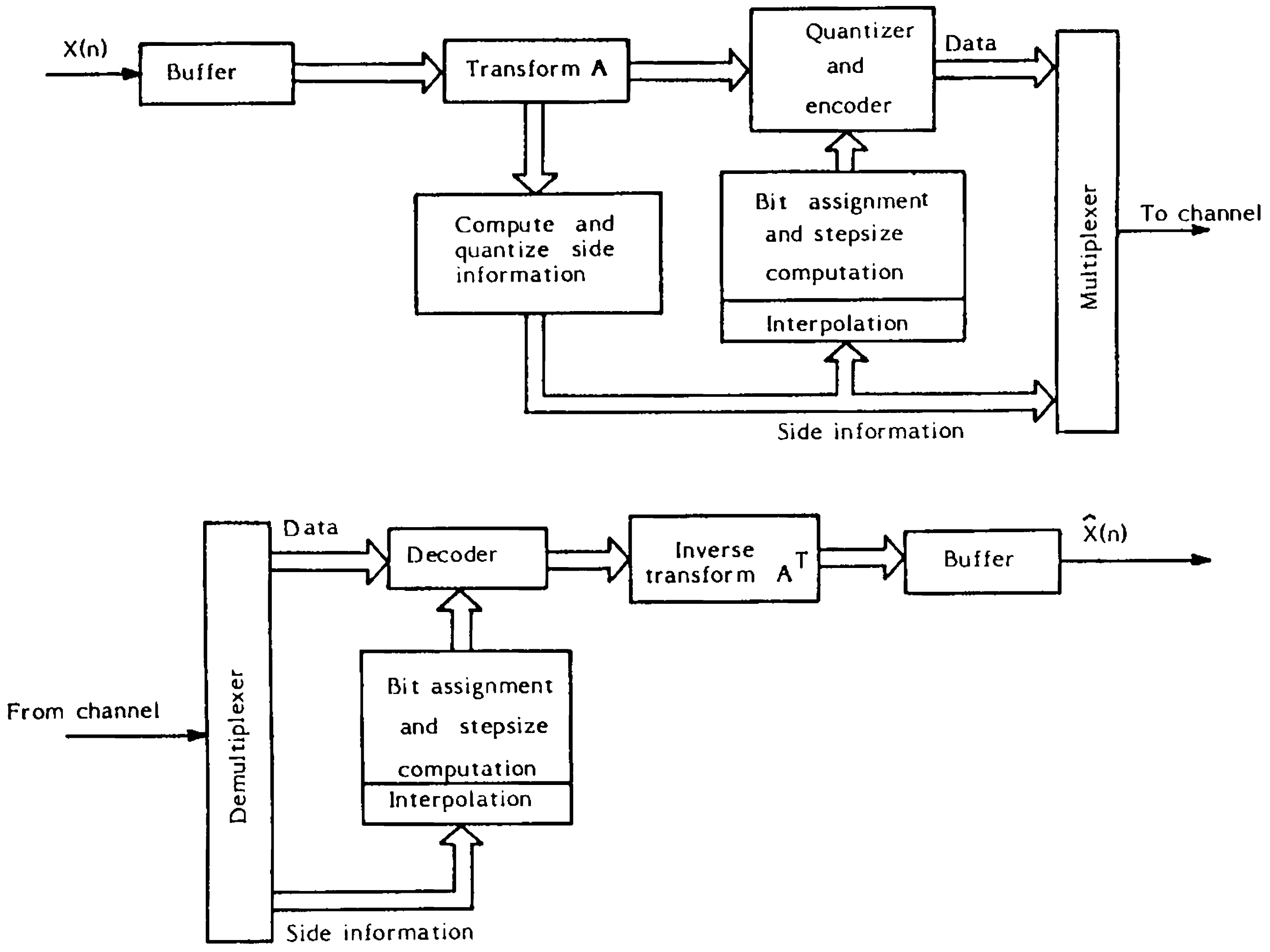


Fig. 6.2 Block diagram of an adaptive transform coder

losing the fine structure of the variance spectrum and, inevitably, a good part of the transform gain [Yair Shoham, 1985].

Various techniques have been tried in order to efficiently code the spectral side information [R.E.Crochiere and J.M.Tribolet, 1979], [J.M.Tribolet and R.E. Crochiere, 1980], [R.Zelinski and P.Noll, 1977], [R.Zelinski and P.Noll, 1979], and [R.V.Cox and R.E.Crochiere, 1981]. In all these schemes the total number of bits for each frame is the same, but the distribution of the bits to the transform coefficients changes from frame to frame according to the changing speech statistics. If R is the number of the total available bits to be distributed to the transform coefficients, then the best bit assignment rule is given by the equation [J.Huang and P.Schultheiss, 1963],

$$R(n) = \frac{R}{N} + \frac{1}{2}\left[\log_2 \hat{Y}^2(n) - \frac{1}{N}\sum_{n=0}^{N-1} \log_2(\hat{Y}^2(n))\right] \quad \text{bits/sample} \qquad (6.2)$$

where $\hat{Y}(n)$ is the estimated variance of the k^{th} transform coefficient. An estimate of the transform coefficient is transmitted as "side information" which is used by the transmitter and receiver for step-size adaptation and bit allocation.

6.5.1 Spectral Parameterization

Since speech is a quasi-stationary process the spectral variance are not known apriori and must therefore, be estimated, encoded, and transmitted to the receiver. This information about the spectral variance is often referred to as "side information".

One of the basic adaptation techniques for transform coding of speech proposed by Zelinski and Noll [1977] is illustrated in Fig.6.3. The spectrum of the speech signal is represented by a reduced set of (typically 16 to 24) equally spaced samples of the spectral estimate. These samples are computed by a local averaging of the logarithm of the square of N/L coefficients around a sample coefficient (The variable L represents the number of transform coefficients we would like to transmit as "side information"). The sample values are quantized and encoded for transmission to the receiver as side information. These quantities represent the estimate of some of the Y(k). The rest of the Y(k) are computed from these values by interpolation, as shown in Fig.6.3(c). They are also decoded and used in the transmitter so that the step-size and bit allocation

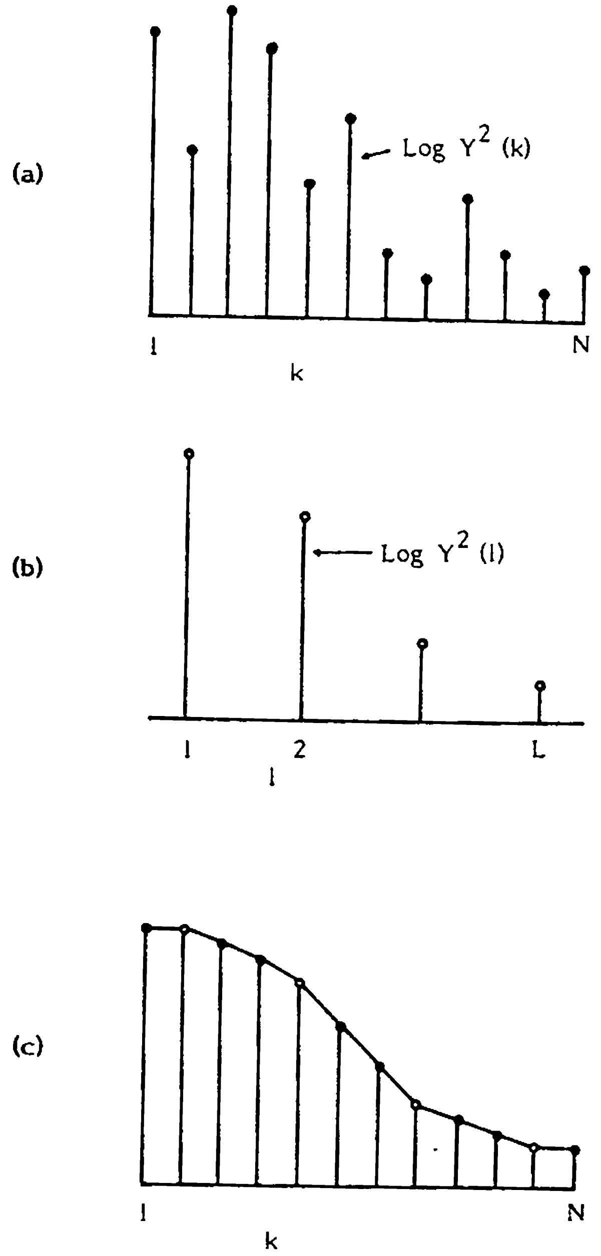


Fig. 6.3 Adaptive bit allocation in transform coding of speech: (a) transform coefficients: (b) result of averaging N/L values in (a): (c) spectral envelope obtained by straight-line interpolation (N= 12, L= 4)

computation is exactly duplicated in the transmitter and receiver. The encoding of the side information requires approximately 2 kbits/s.

This simple algorithm has been referred to as "non-speech specific" since it does not take into account the dynamical properties of speech production. This adaptation technique is however quite appropriate for speech transmission at or above 16 kbits/s, since there are enough bits to allow accurate representation of the fine structure of the spectrum. But it becomes increasingly more difficult to accurately encode the fine structure at rates below 16 kbits/s. Therefore the signal is degraded with distortion at these bit rates.

J.M.Tribolet and R.E.Crochiere [1978] proposed a more appropriate algorithm for lower bit rates. This is a more complex, "speech specific" adaptation algorithm which utilizes the traditional model of speech production to predict the spectral coefficients. This algorithm is based on an all pole model of the formant structure of speech and a pitch model to represent the fine structure (pitch striation) in the speech spectrum. The resulting algorithm is referred to as a "vocoder-driven" adaptation strategy due to the close relationship of this spectral estimate to a vocoder model. A similar technique is to extract a homomorphic representation of the spectrum, extract the pitch, and then, synthesize an estimate for the spectrum [R.V.Cox and R.E.Crochiere, 1981].

These complex "speech specific" techniques have been shown to improve the ATC performance at bit rates from 9.6 to 16 kbits/s. However, the spectral estimates sometimes poorly represent the original spectrum due to inadequate formant tracking and pitch estimation.

Another major drawback of this speech specific ATC is its high computational complexity. In the present work our aim is to improve the simple algorithm proposed by Zelinski and Noll without introducing much complexity. The remaining part of this chapter concentrates on this point.

6.5.2 Description of the Proposed Modification

A. Maximum amplitude of the transform coefficient as a side information

In the side information extraction process, the estimated coefficients are obtained by locally averaging the logarithms of the square of the transform coefficients. The remaining coefficients are obtained by linear interpolation. This will give only an approximation about the spectral envelope. The linear smoothing is used based on the assumption that speech spectrum varies slowly. But often this is not true. Some of the speech transform coefficients may be very much predominant compared to its neighbours as illustrated in Fig.6.4 (Fig.6.4 is a DCT spectrum of a speech segment of length 256). When locally averaged, the identity of such coefficients is lost and they are never faithfully recovered at the receiver. This causes error in step size computation in both transmitter and receiver. This causes deterioration in the ATC performance.

As a solution to this problem the maximum amplitude of the transform coefficient is also considered as a side information. The step size is computed using this maximum amplitude both in transmitter and receiver. The steps involved in this modified scheme is as follows.

1. The logarithm of the square of transform coefficients is computed, and they are grouped in sets of N/L and averaged together. (L is the number of side information coefficients).

DCT spectrum of a segment of speech

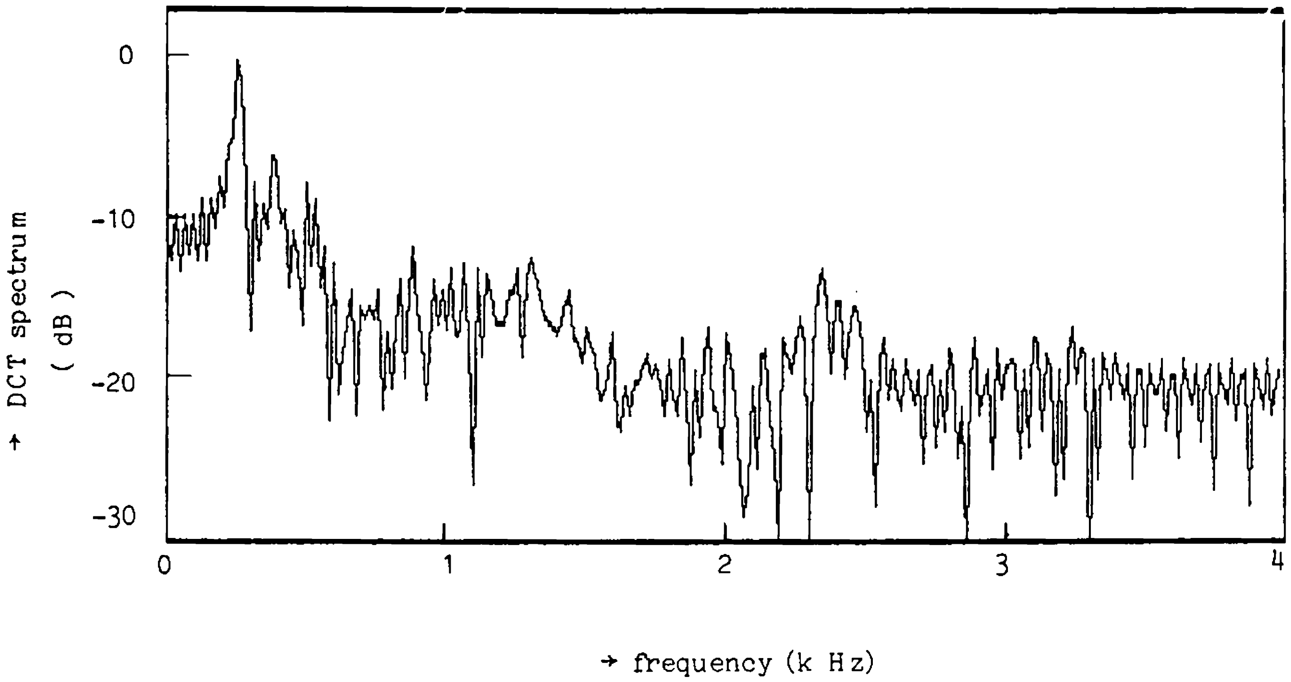


Fig. 6.4

2. The rest of the coefficients are estimated by linear interpolation as in Fig.6.3(c).

3. Using equation (6.2) the number of bits available for quantization of each coefficient are computed.

4. The maximum value of the number of bits assigned is obtained as b_{max}.

5. The maximum value of transform coefficients is obtained as Y_{max}.

6. Step size is computed as $\Delta_s = \frac{Y_{max}}{\left[2^{(b_{max}-1)} - 1\right]}$

(One bit used for sign information is subtracted from b_{max}).

This modified scheme enables recovery of the transform coefficients with higher power values, more accurately. Along with the 'L' averaged coefficients, the maximum amplitude coefficient is also transmitted to the receiver as 'side information' so that in the receiver, the step size computation is exactly duplicated as in the transmitter.

B. Modified bit assignment

The optimum bit assignment scheme is given by the equation (6.2) as

$$R(n) = \frac{R}{N} + \frac{1}{2} [\log_2 \hat{Y}^2(n) - \frac{1}{N} \sum_{n=0}^{N-1} \log_2 \hat{Y}^2(n)] \text{ (bits/sample)}$$

where $R(n)$, $\frac{R}{N}$, and $\hat{Y}^2(n)$ are the bits assigned to n^{th} coefficient, the average bits available per sample and the estimated and quantized variance of the n^{th} transform coefficient respectively. Generally the computed value of $R(n)$ is not an integer. Also some of the $R(n)$'s can be negative. It is meaningless that the bits assigned is noninteger and negative. Various sub-optimal solutions to this problem have been studied by S.Krishnan and K.K.Paliwal [1987]. The bit assignment with 'largest variance first' is reported to provide better SNR performance in their study. However this method require involved computation. A computationally simple and efficient bit reassignment method is proposed here. Firstly, if the number of bits assigned to a coefficient is less than zero, then it is reassigned as zero. If only one bit is available for a coefficient, that coefficient, is not coded and transmitted since it will encode only the sign information. Such bits are reassigned one bit each for the coefficients in the lower frequency band. The steps involved in this scheme may be written mathematically as follows.

1. If $R(n) < 0$, then $R(n)_r = 0$; $(1 \leq n \leq N)$

2. If $0 < R(n) \leq 1$ then $R(n)_r = 0$ and

   $$\text{bit balance} = \sum_{i=1}^{N} R(i), (R(i) = 0 \text{ if } R(i) > 1).$$

3. $R(n)_r = R(n) + 1$ for $(1 \leq n \leq \text{bitbalance})$. Here $R(n)_r$ and bitbalance are respectively the reassigned bits and the sum of bits that falls between 0 and 1.

6.6 EXPERIMENTAL RESULTS

This section describes a computer simulation study of the ATC coder described in the previous section.

The coders simulated are of the following types.

Type 1: ATC coder proposed by Zelinski and Noll, without bit reassignment.

Type 2: ATC coder proposed by Zelinski and Noll, with modified bit reassignment.

Type 3: Type 1 modified by 'maximum amplitude coefficient as side information'.

Type 4: Type 3 with modified bit reassignment.

All the four types are simulated and studied with different data block length (64, 128 and 256) and different bit rates (8, 9.6, 12, 16, 24, 32 kbits/s). Both DCT and DWHT are used in the simulation.

All the simulation programs are developed in Turbo Pascal and implemented on an IBM PC/AT. The coder performance are measured in terms of SNR value.

6.6.1 Comparison of the coder performance

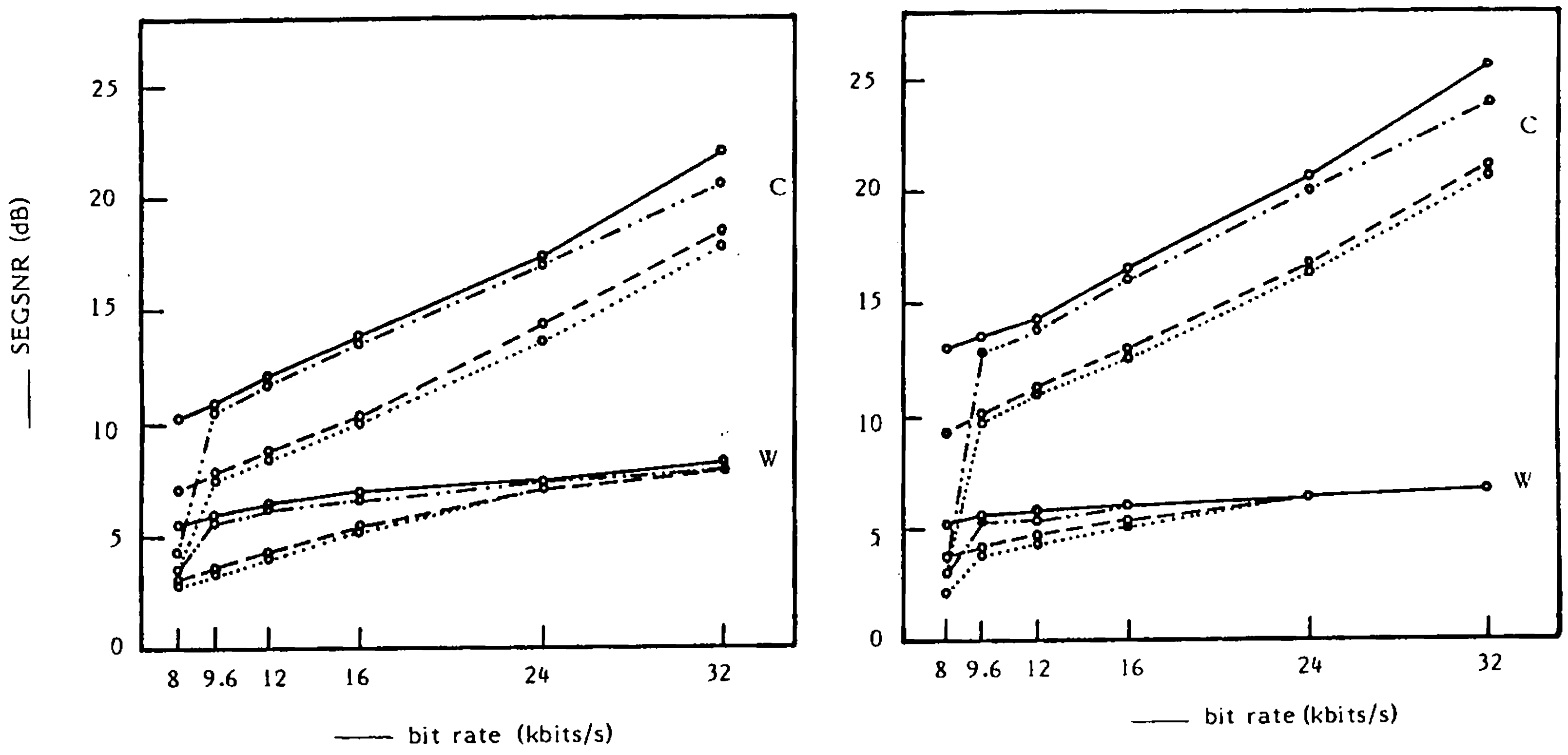
The speech signal is characterized by its time varying nature. To take into account this fact, the performance of a coding system is measured in terms of segmental SNR which is denoted as SEGSNR. To compute the SEGSNR we divide the speech signal into segments of 64-256 sample length, and compute the SNR(m) dB where m = 1,2,...,M for each block of a particular block length. Then the segmental

SNR is defined by

$$\text{SEGSNR} = \frac{1}{M} \sum_{m=1}^{M} \text{SNR}(m) \text{ dB.}$$

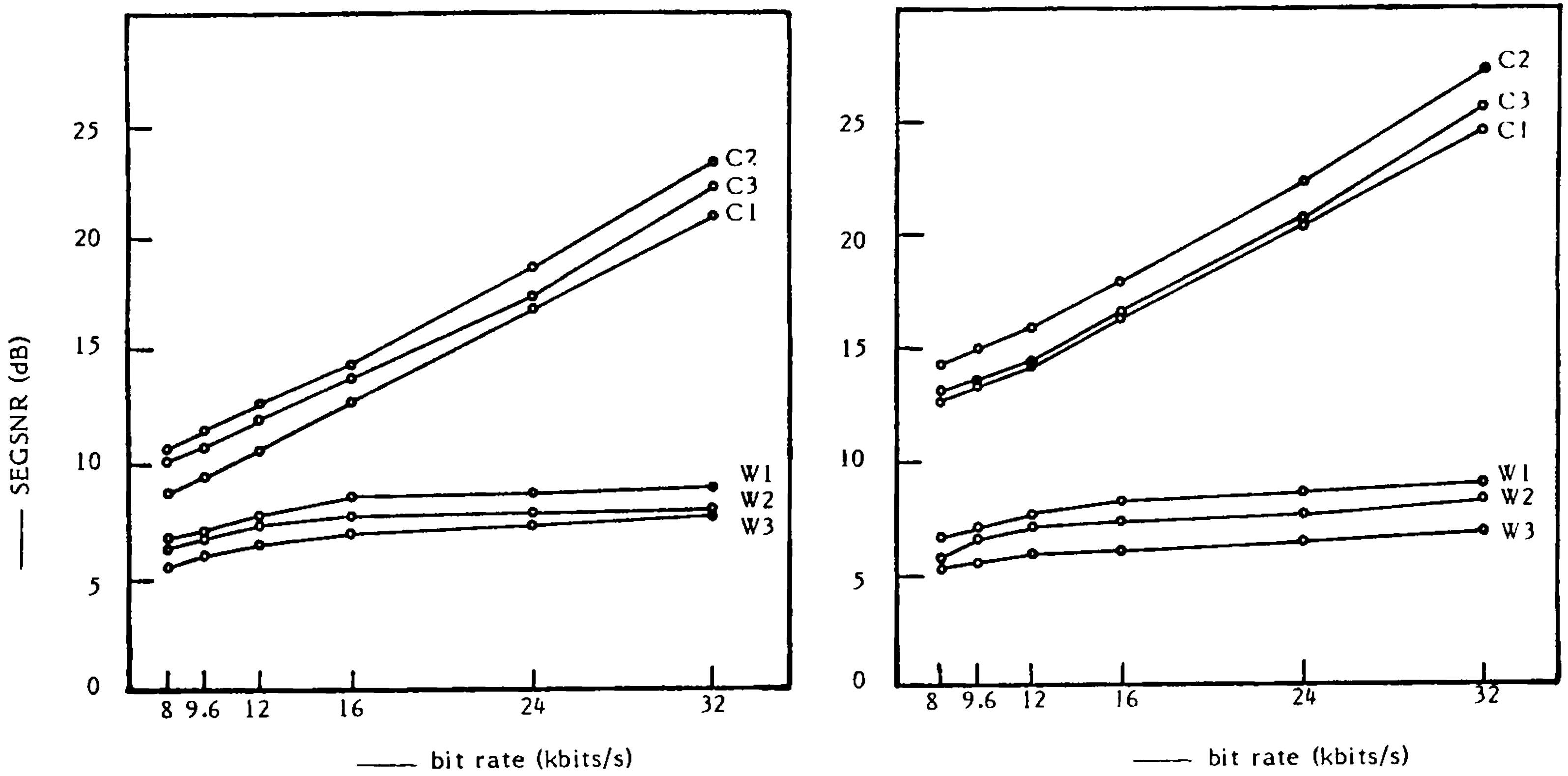
Fig.6.5(a) and (b) represent the SEGSNR of the four types of ATC coders simulated with data segment length of 256 samples each vs. bit rate for two speakers, a male and a female (mentioned in section 4.2.1). It may be noted that type 4 gives better performance compared to the other three. In the SEGSNR sense, the DCT based ATC is far superior compared to DWHT based one. There is a slight difference in performance for the two speakers (male and female).

Fig.6.6(a) and (b) give a comparison of the SEGSNR of the coder type 4 for different data block length (64, 128 and 256) vs. bit rate. When DCT is used, the 128 segment length gives better performance compared to 256 and 64 segment sizes. When DWHT is used, 64 segment size performs better compared to the other two. The SEGSNR obtained at 16 kbits/s using DCT was 14.42 dB for male speech and 17.8 dB for female speech at 128 segment length.



(a) (b)

Fig. 6.5(a&b) SEGSNR for the four types of coders versus bit rate (a - male, b - female, C: with DCT, W: with DWHT, ······· type 1, -·-·· type 2, ----- type 3,——— type 4)



(a)

(b)

Fig. 6.6(a&b) SEGSNR of type 4 coder for different data block length versus bit rate (a - male, b - female,
C1, C2, C3: using DCT with block lengths 64, 128, 256 respectively,
W1, W2, W3: using DWHT with block lengths 64,128, 256 respectively)

Figs.6.7, 6.8, 6.9 and 6.10 present the variation of the signal to noise ratio from segment to segment obtained by the ATC coder of type 1, 2, 3 and 4 respectively over a bit rate of 8, 16 and 32 kbits/s (simulation was carried out at bit rates 9.6, 12 and 24 kbits/s also) respectively for a frame of speech containing 58 consecutive segments of 256 samples each at 8 kHz sampling frequency. The continuous line represents the performance of the coder simulated using DCT and the broken line that using DWHT. At 32 kbits/s bit rate the SNR of DCT ATC is much better than DWHT ATC, for almost all the segments, for the four types of coders. For some segments the SNR of DCT ATC falls considerably, especially when the coder is designed for bit rates below 16 kbits/s. These segments are identified as unvoiced segments. (Fig.6.11 (a) and (b) represents the waveform of two such segments, segment 13 and 14). It may be noted that at lower bit rates the DWHT ATC gives better SNR for these segments of speech. Among the four types of coders the type 4 performs better even at a bit rate of 8 kbits/s. Figs.6.12, 6.13 and 6.14 illustrate the comparison of the coded voiced waveform using DCT ATC and DWHT ATC, while Figs.6.15, 6.16 and 6.17 illustrate the comparison of the coded unvoiced waveform using DCT ATC and DWHT ATC (both for the type 4 coder) at a bit rate of 8 kbits/s.

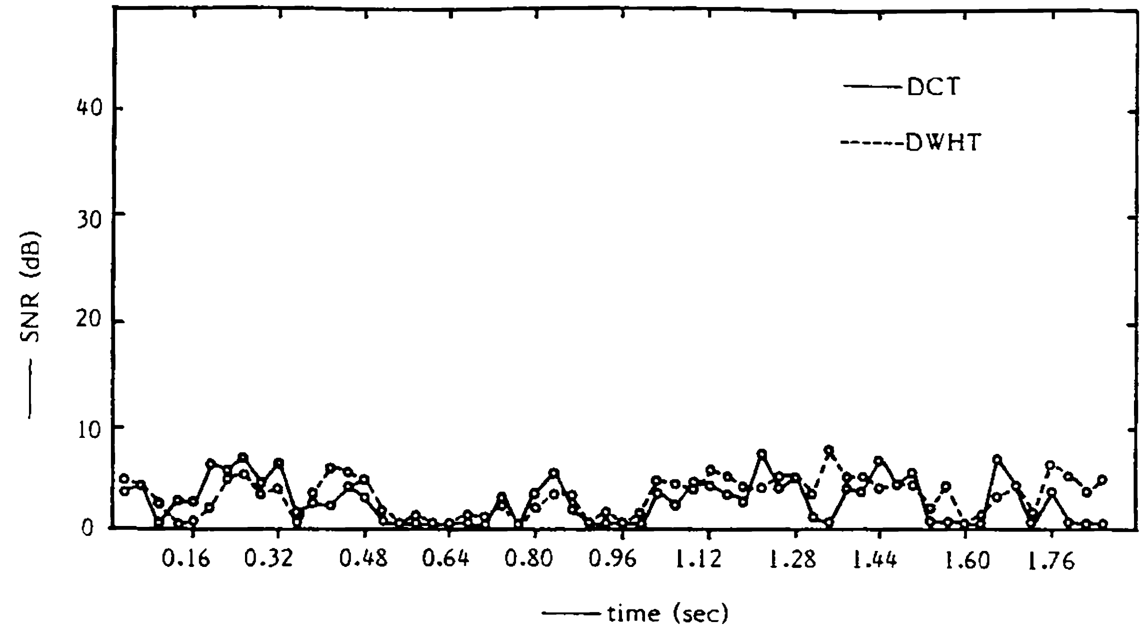


Fig. 6.7(a) Time dependance of SNR (dB) of coded speech (type 1. 8 kbits/s).

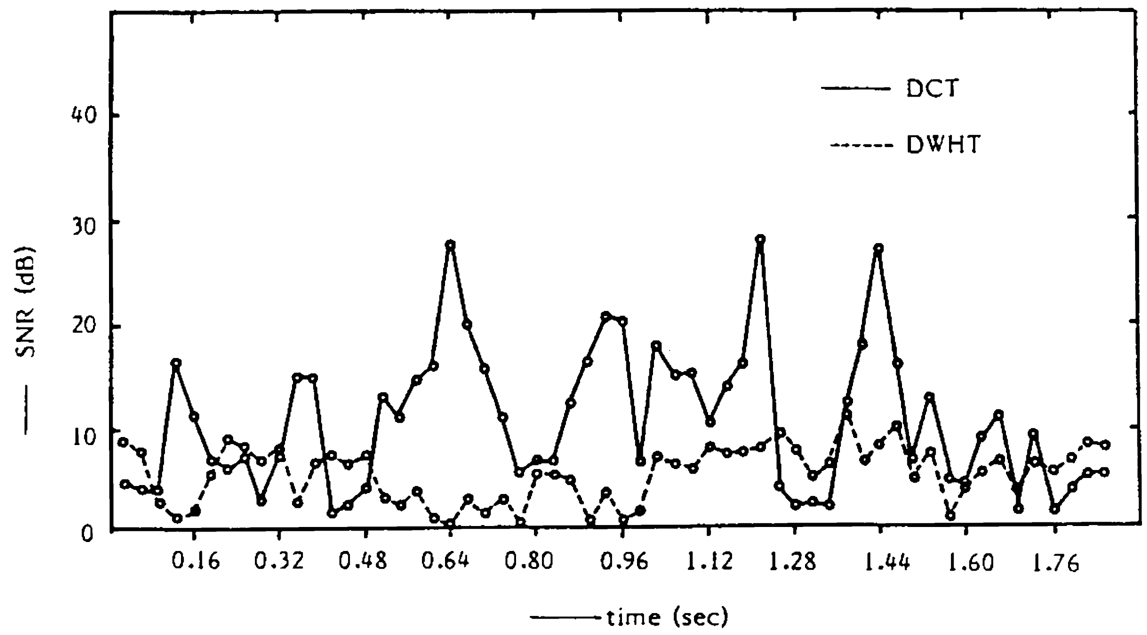


Fig. 6.7(b) Time dependance of SNR (dB) of coded speech (type 1. 16 kbits/s).

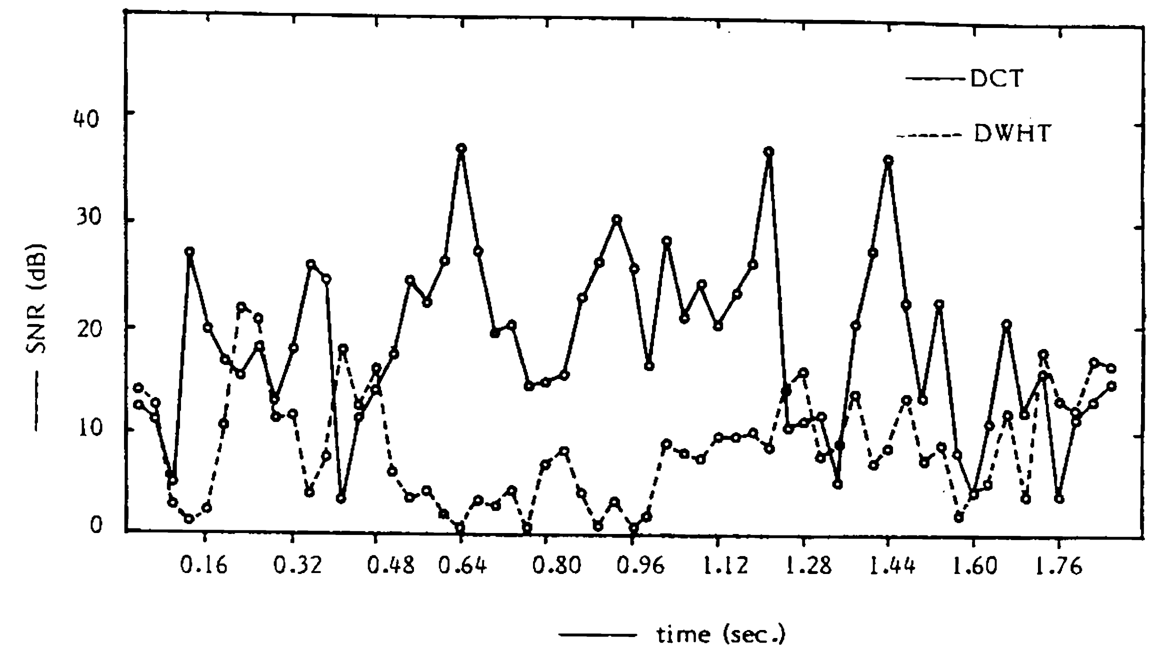


Fig. 6.7(c) Time dependance of SNR (dB) of coded speech (type 1. 32 kbits/s).

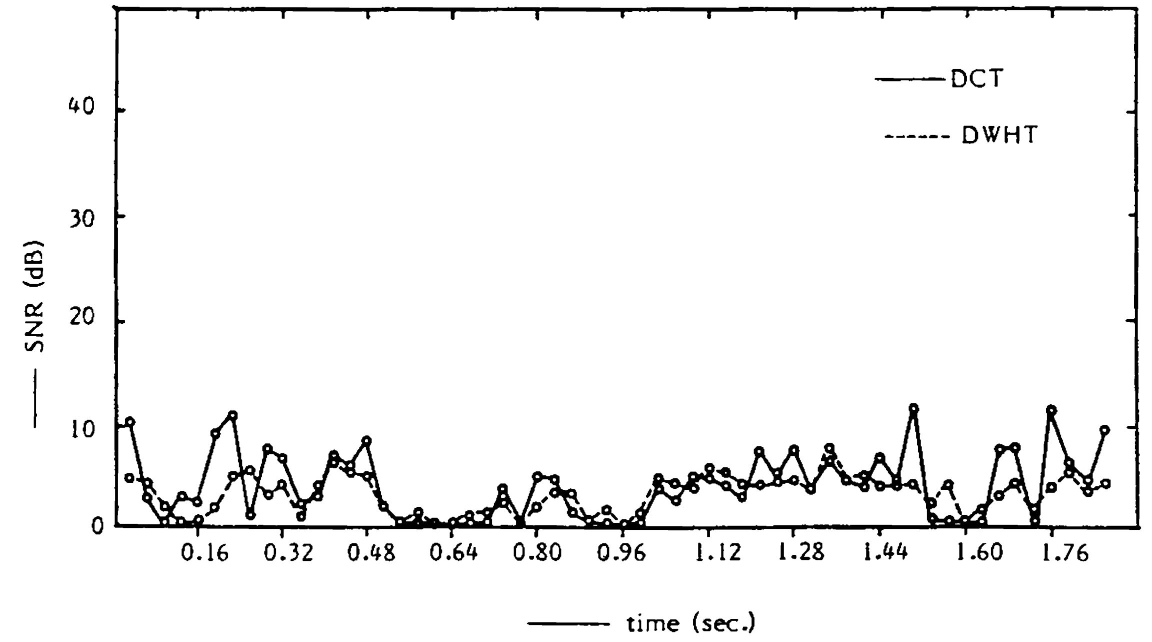


Fig. 6.8(a) Time dependance of SNR (dB) of coded speech (type 2. 8 kbits/s).

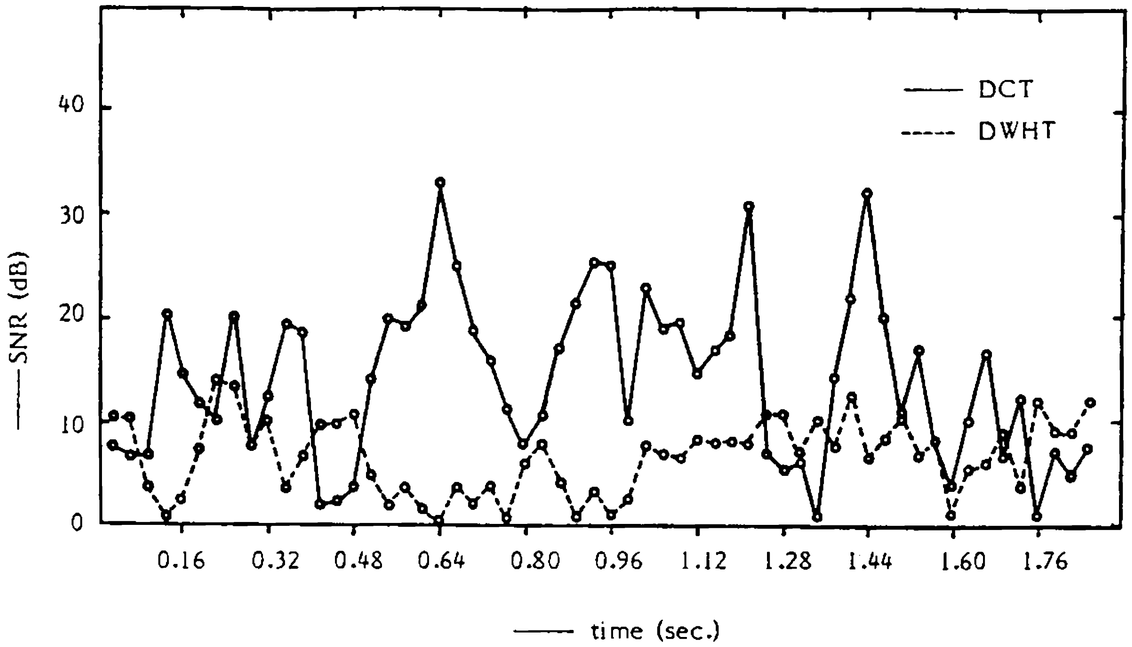


Fig. 6.8(b) Time dependance of SNR (dB) of coded speech (type 2. 16 kbits/s)

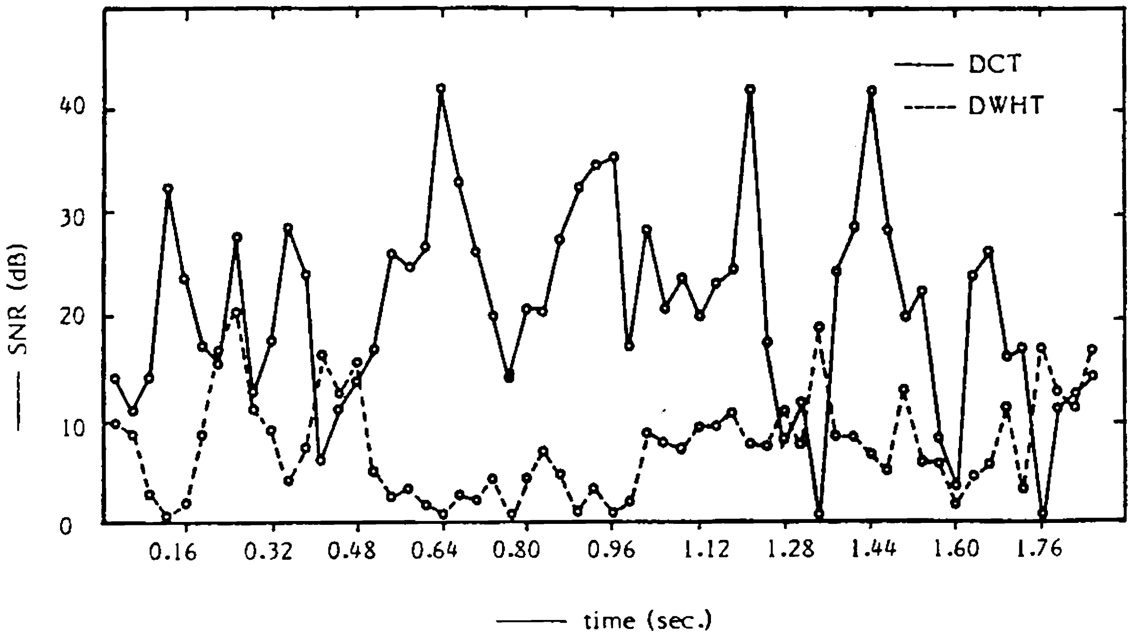


Fig.6.8(c) Time dependance of SNR (dB) of coded speech (type 2. 32 kbits/s)

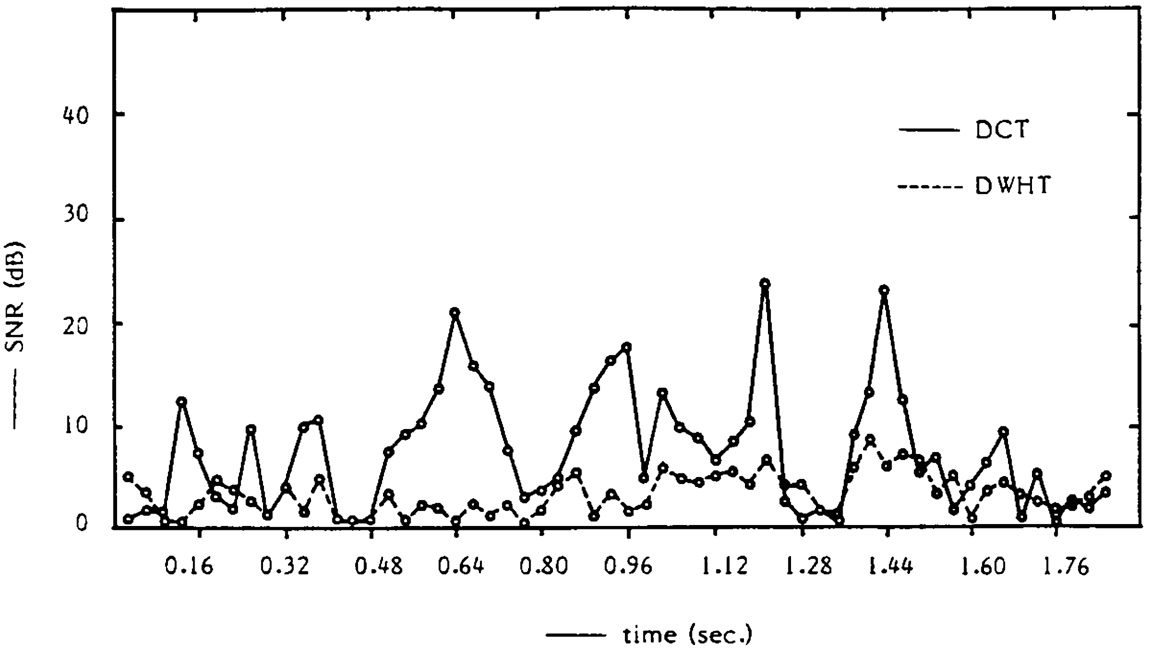


Fig. 6.9(a) Time dependance of SNR (dB) of coded speech (type 3. 8 kbits/s)

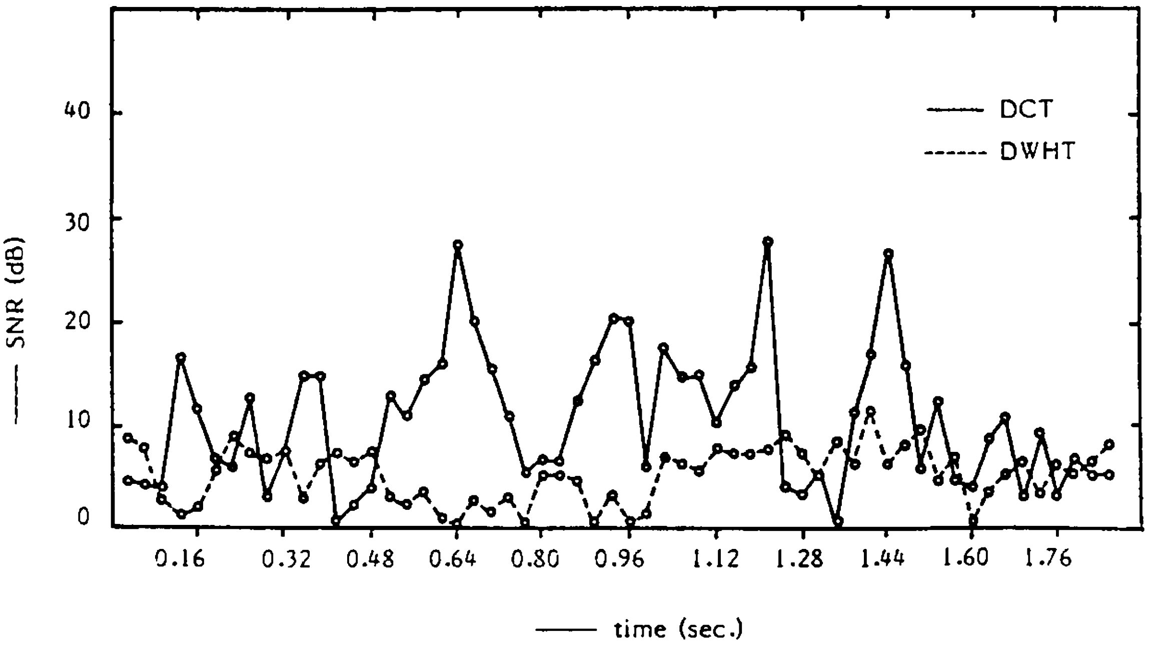


Fig. 6.9(b) Time dependance of SNR (dB) of coded speech (type 3. 16 kbits/s)

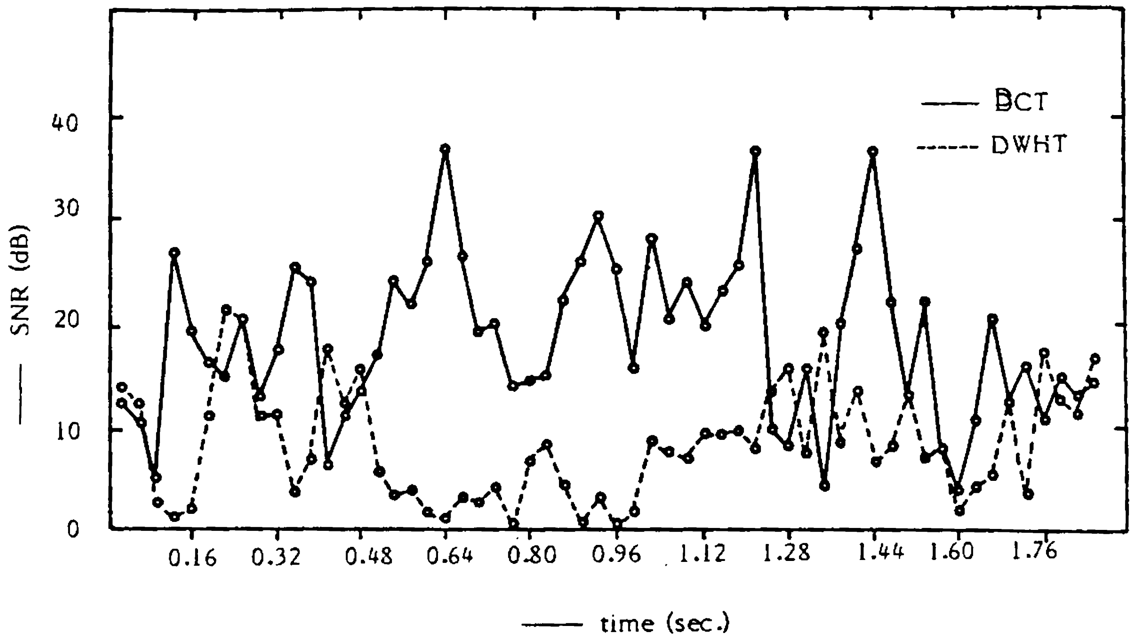


Fig. 6.9(c) Time dependance of SNR (dB) of coded speech (type 3. 32 kbits/s)

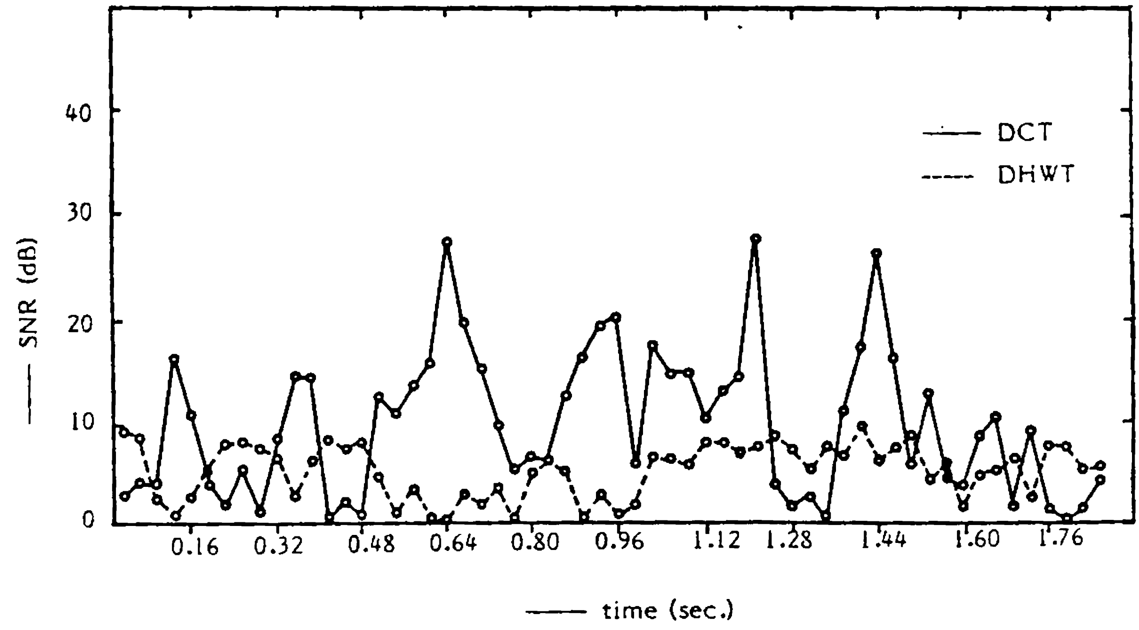


Fig. 6.10(a) Time dependance of SNR (dB) of coded speech (type 4. 8 kbits/s)

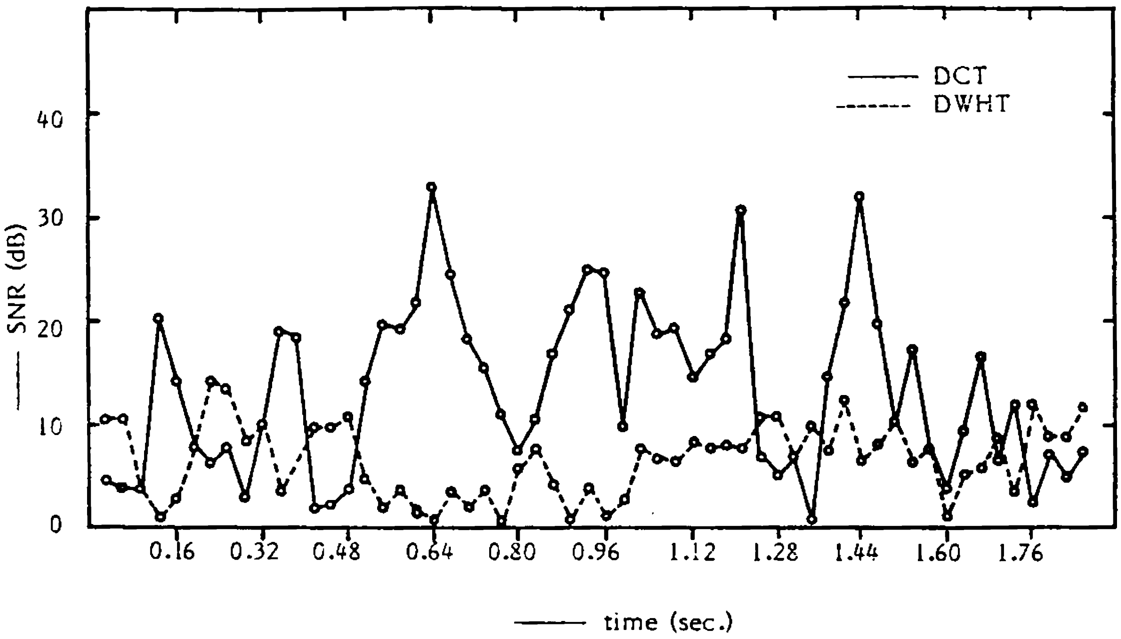


Fig. 6.10(b) Time dependance of SNR (dB) of coded speect (type 4, 16 kbits/s)

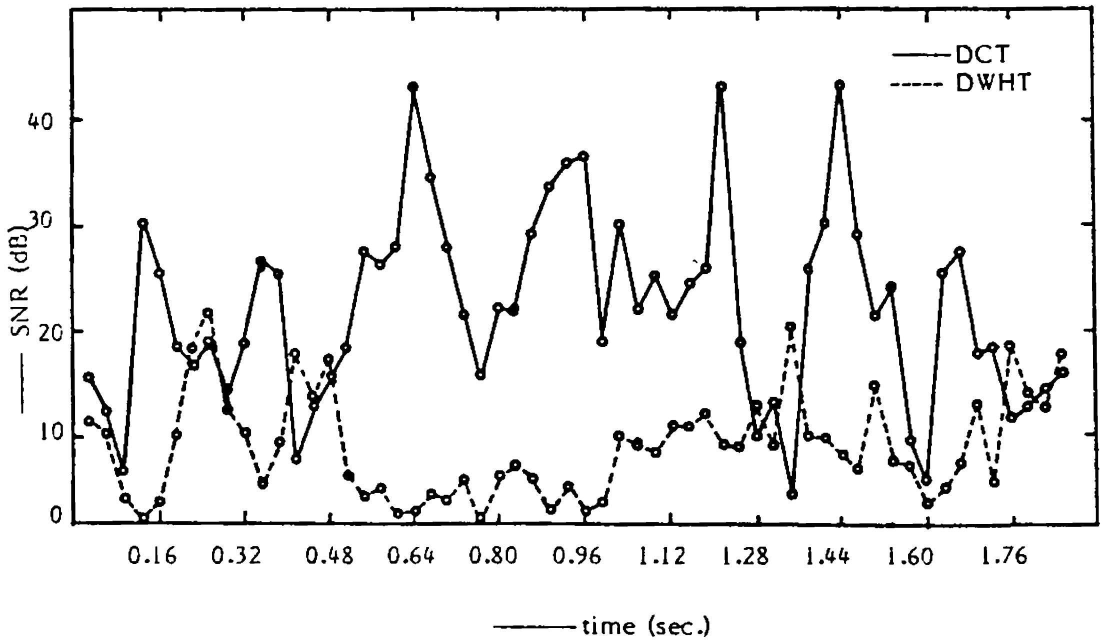


Fig. 6.10(c) Time dependance of SNR (dB) of coded speech (type 4, 32 kbits/s)

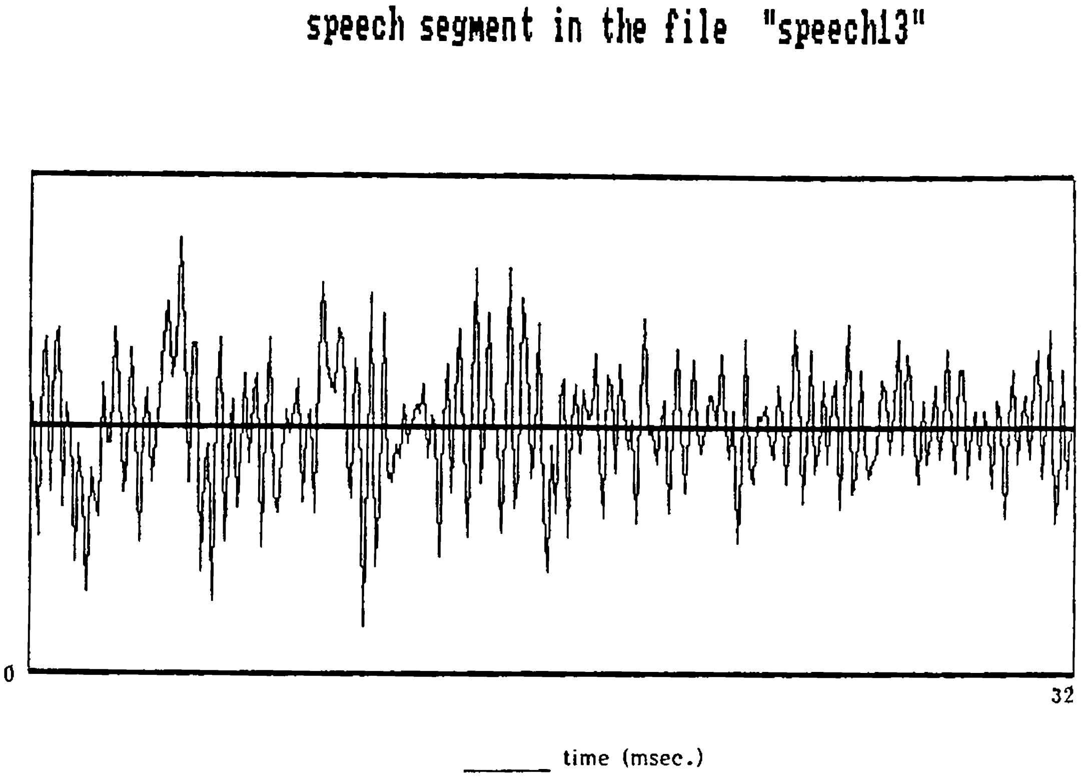


Fig.6.11(a)

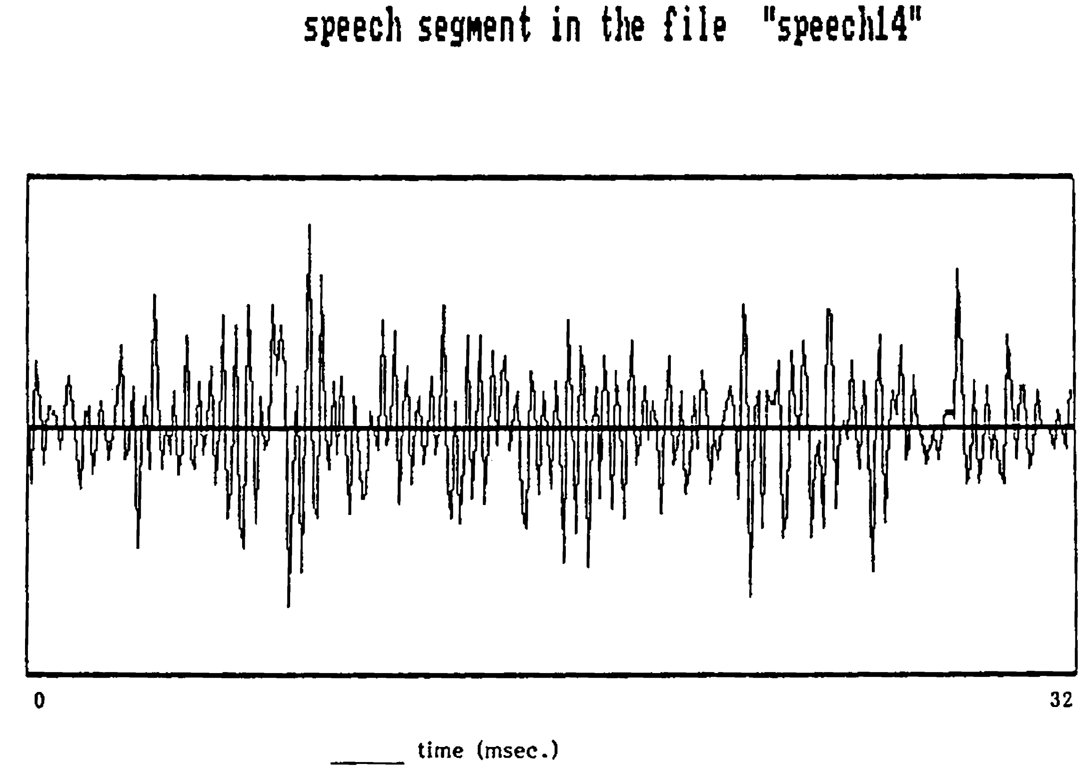


Fig.6.11(b)

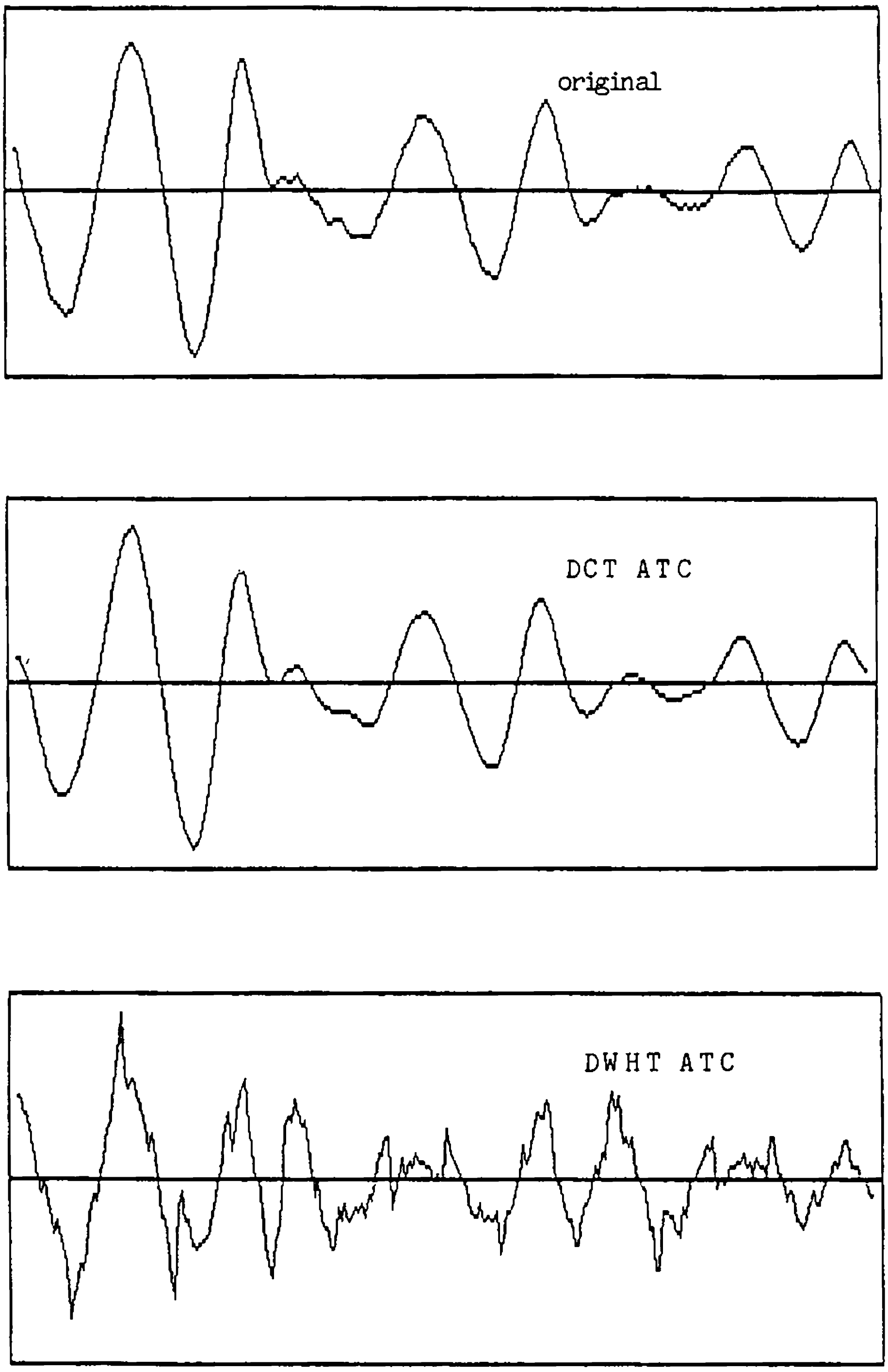


Fig. 6.12 Comparison of coded voiced waveform using DCT ATC and DWHT ATC

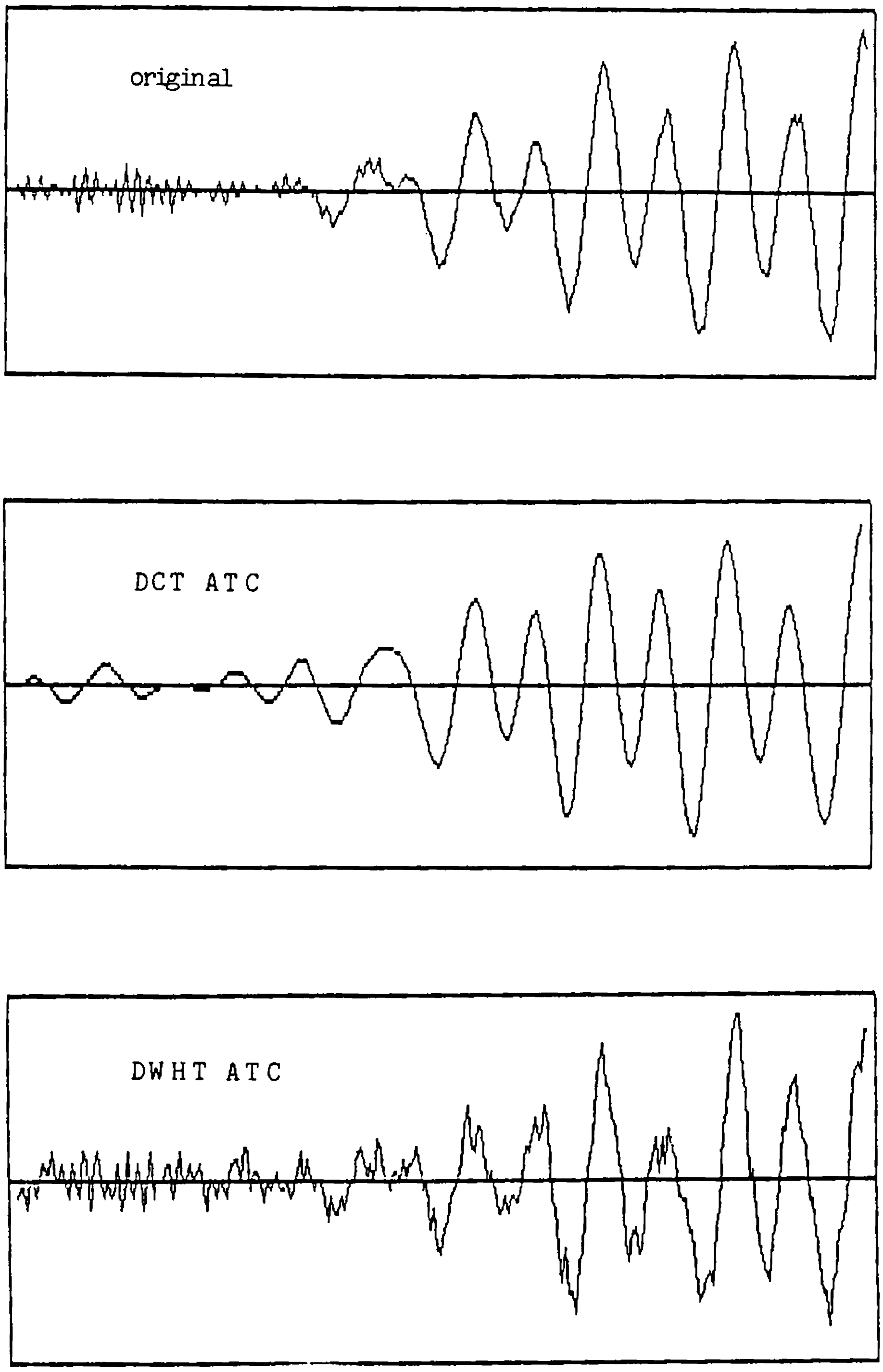


Fig. 6.13 Comparison of coded voiced waveform using DCT ATC and DWHT ATC

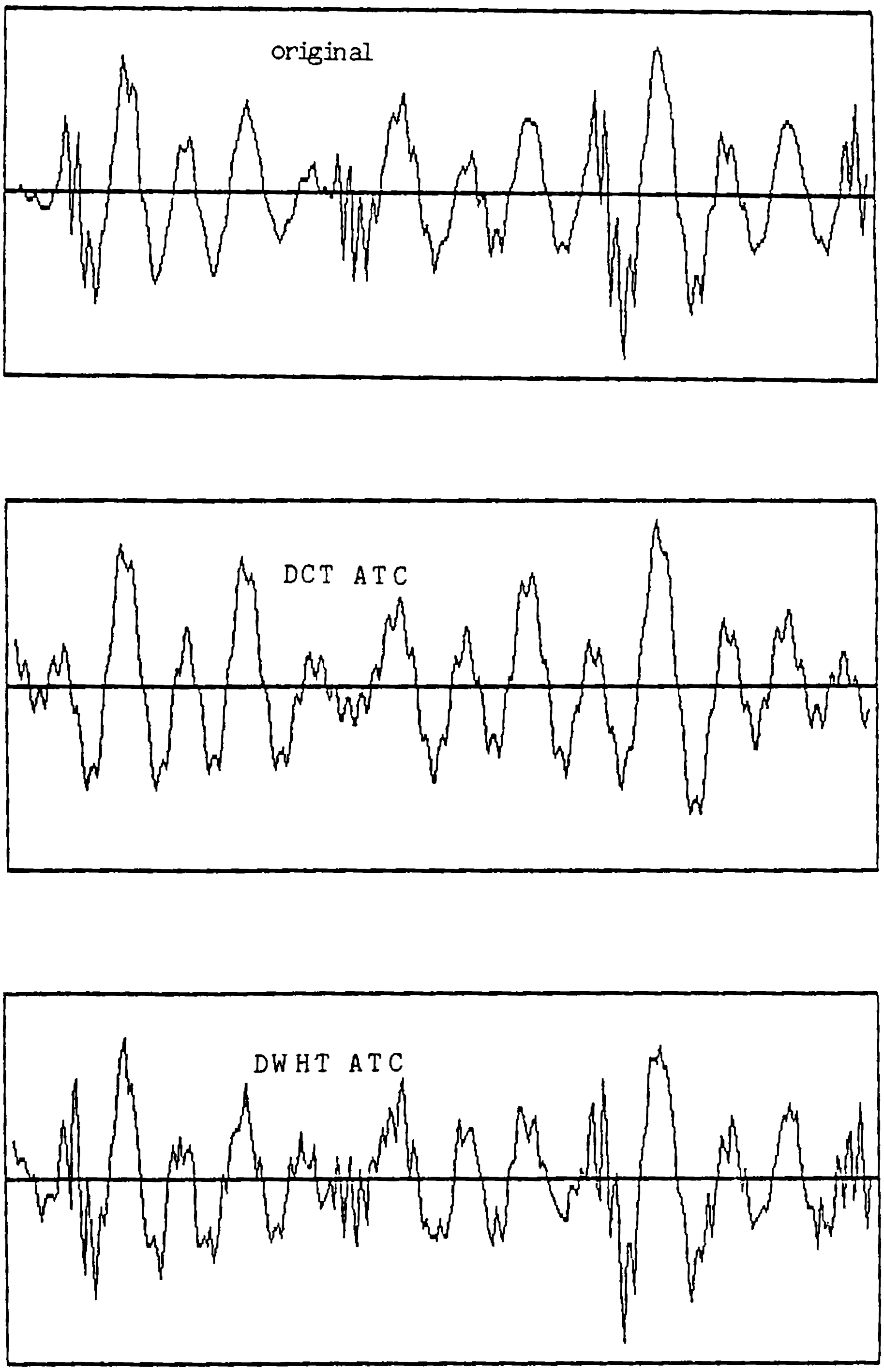


Fig. 6.14 Comparison of coded voiced waveform using DCT ATC and DWHT ATC

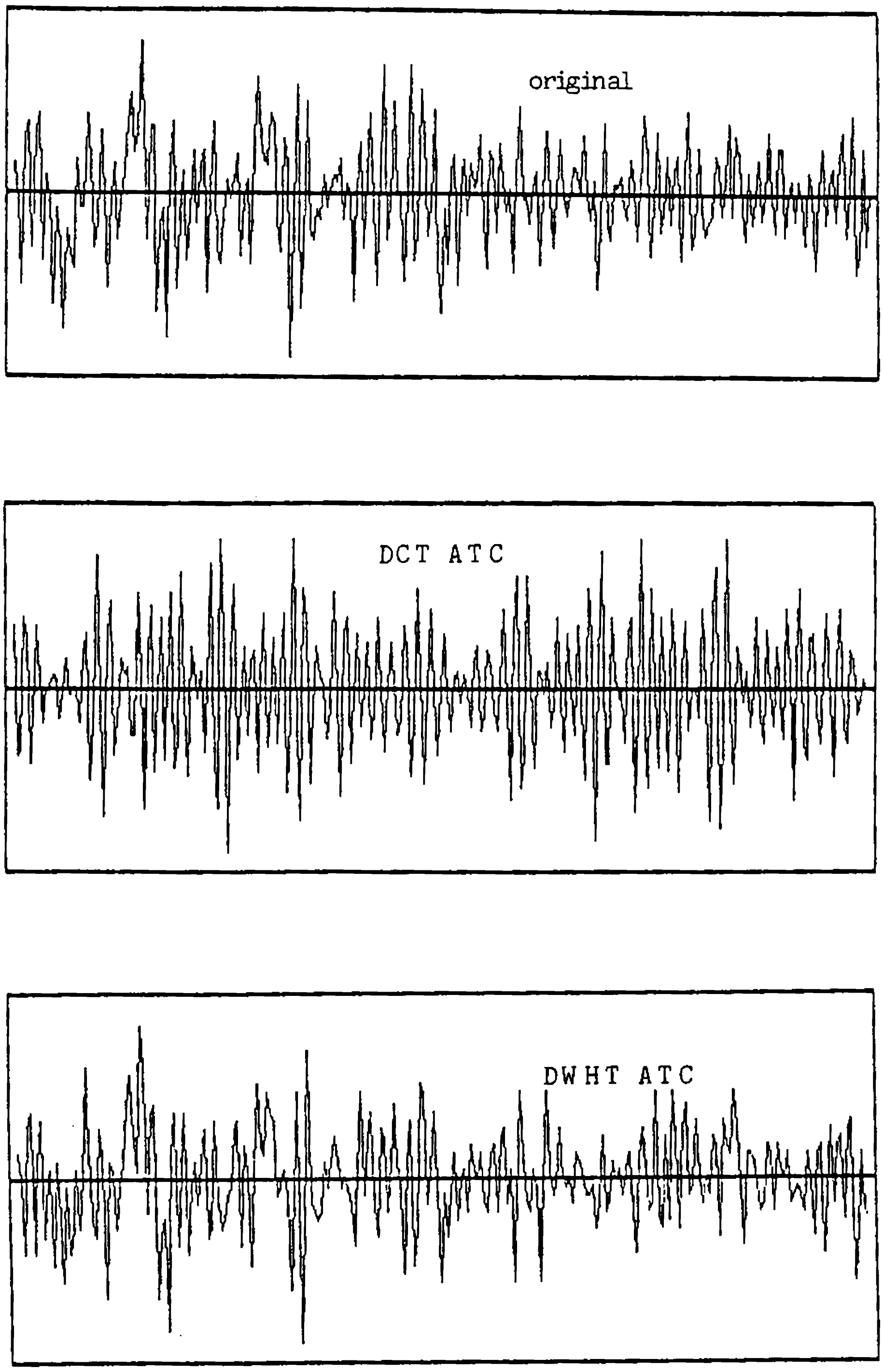


Fig. 6.15 Comparisom of coded unvoiced waveform using DCT ATC and DWHT ATC

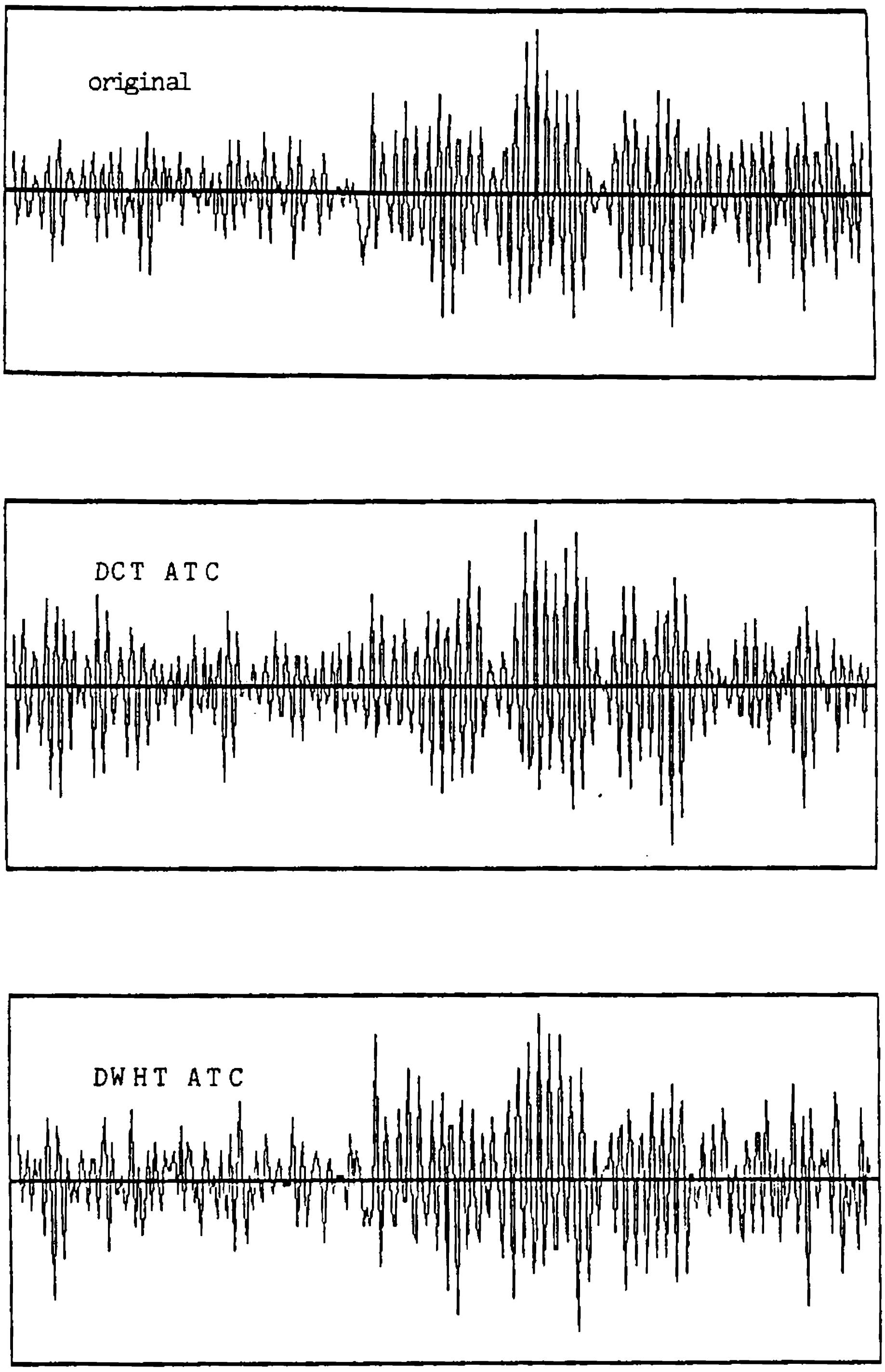


Fig. 6.16 Comparison of coded unvoiced waveform using DCT ATC and DWHT ATC

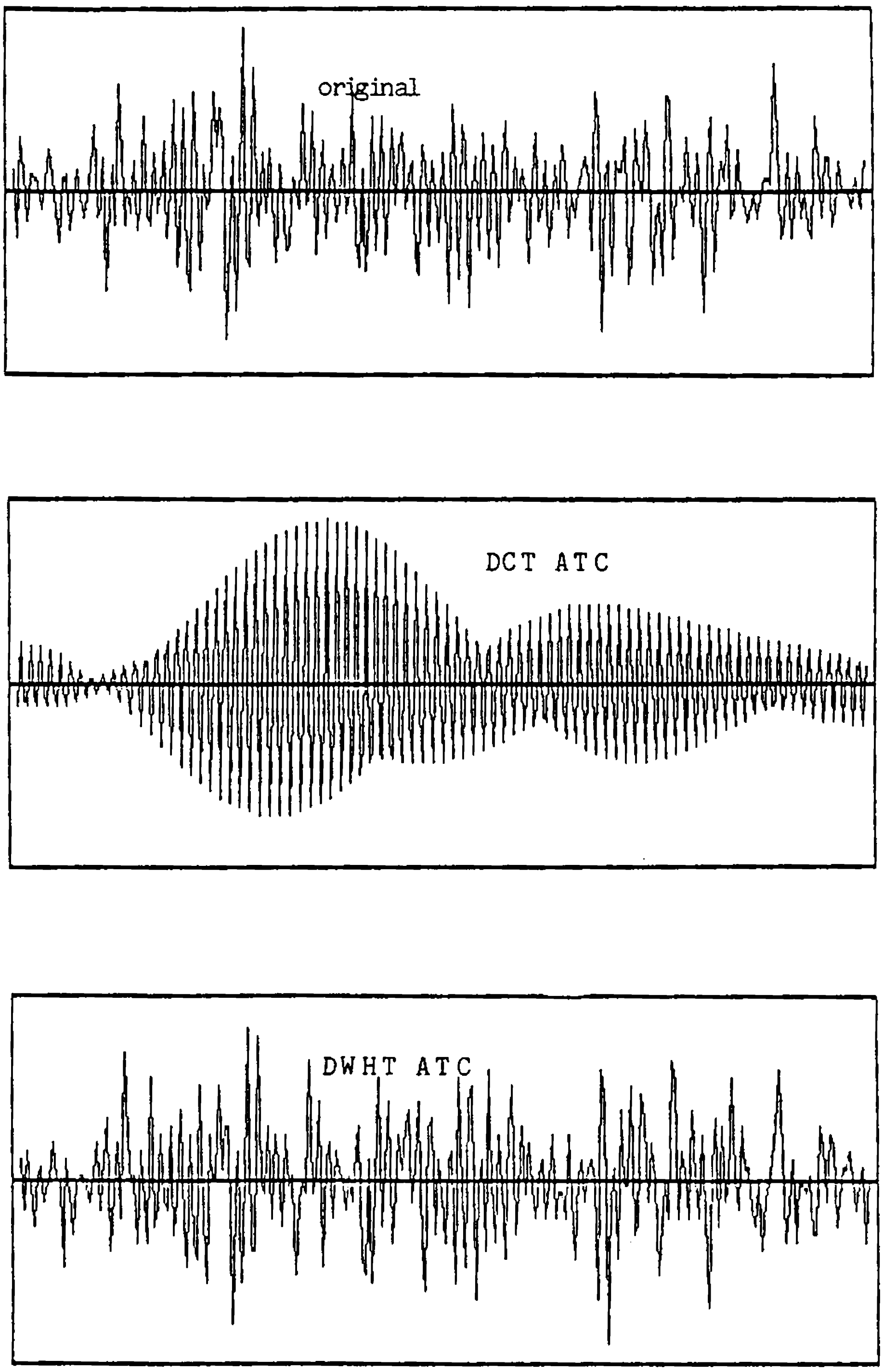


Fig. 6.17 Comparison of coded unvoiced waveform using DCT ATC and DWHT ATC

The main conclusion from these experiments is that the incorporation of the maximum amplitude transform coefficient as the side information and the modified bit reassignment method into an ATC scheme helps to achieve a better performance due to more efficient quantization of the side information and the data.

To summarize, this chapter presents the design method of a modified Adaptive Transform Coder. The results of a computer simulation study of the modified coder with DWHT and DCT applied to coding of speech waveforms are presented. Experimental results show that, the modified coder performs better significantly even at a bit rate of 8 kbits/s. Further it is seen that the DWHT will give better SNR for unvoiced speech segments than DCT. Based on this result an adaptive switching of transform is used for better speech quality which is discussed in the next chapter.

As a final remark, we have not tried to subjectively improve the performance of the coder by known techniques like block overlapping, noise shaping, post filtering etc. The main focus of this study was on the effectiveness of the proposed modifications as measured by a simple objective criterion like the SNR. Those techniques may be added to the proposed system to improve the subjective quality of the coded speech.

Chapter 7

ADAPTIVE SWITCHING TRANSFORM CODER

7.1 INTRODUCTION

In the previous chapter we discussed a relatively simple modification of the ATC scheme proposed by Zelinski and Noll. This chapter is devoted to a more sophisticated coding scheme in which the V/UV notion is used in combination with the earlier coding scheme to enhance the already established waveform coding method.

The modified coder in the previous chapter was studied using DCT and DWHT. The SNR of DWHT ATC was very much lower than DCT ATC for voiced speech. But for unvoiced speech the SNR of DCT ATC falls below that of DWHT ATC, especially when the coder was designed for a bit rate below 16 kbits/s. Therefore better speech quality is possible by way of an adaptive switching of transforms by performing V/UV tests on speech segments. These techniques have shown improvements in the ATC performance at bit rates ranging from 8 to 16 kbits/s.

The chapter starts with a general description of the coder's main building blocks: the transform used and the V/UV decision algorithm. Finally simulation results are given for 8 to 32 kbits/s coders, which show that the proposed coder indeed performs better than the conventional ATC systems.

7.2 DESCRIPTION OF THE ASTC CODER

The basic block diagram of the Adaptive Switching Transform Coder (ASTC) is shown in Fig.7.1. The input speech is segmented into successive blocks of size K = 256 data each. These blocks are stored in the buffer. The processor first computes the short time zerocrossing rate (STZCR) and short time energy (STE) of each input block **X** and determines whether the region is a voiced or an unvoiced segment. If voiced, the input block **X** undergoes a discrete cosine transform (DCT), which results in the corresponding block **Y** in the DCT domain. And if the input block is an unvoiced segment, then it undergoes a discrete Walsh-Hadamard transform (DWHT). At the next stage, the Maxamp detector finds the maximum amplitude of the transform coefficients. Maximum amplitude of the transform coefficient and DCT/DWHT decision are transmitted as side information together with the spectral information. The bit assignment and step size computation are performed as discussed in chapter 6. The quantized and encoded block $\hat{\mathbf{Y}}$ at the Quantizer output and the side information are multiplexed and transmitted to the channel. At the receiver this is demultiplexed and decoded into $\hat{Y}$. The decoder output is then inverse transformed to get $\hat{X}$, the final output block.

7.2.1 Selection of transforms

The selection of transforms for the coder is based on the results of the previous chapter. The modified ATC coder in chapter 6 gives better SNR performance for voiced speech when DCT is used. For unvoiced speech DWHT gave better SNR performance at bit rates below 16 kbits/s. Our aim in designing the present coder is to improve the speech quality at lower bit rates--i.e., below 16 kbits/s. The experimental results of the previous chapter point that DCT and DWHT are suitable to serve this purpose. The description of DCT and DWHT are given in sections 6.2 and 6.3 respectively.

7.2.2 Voiced/unvoiced classifier

In chapter 4, we have presented two methods for voiced/unvoiced classification. The first one, based on the short-time zerocrossing rate and short time energy of speech signal, is used here. The flow chart of the voiced/unvoiced classification algorithm used in the ASTC is shown in Fig.7.2.

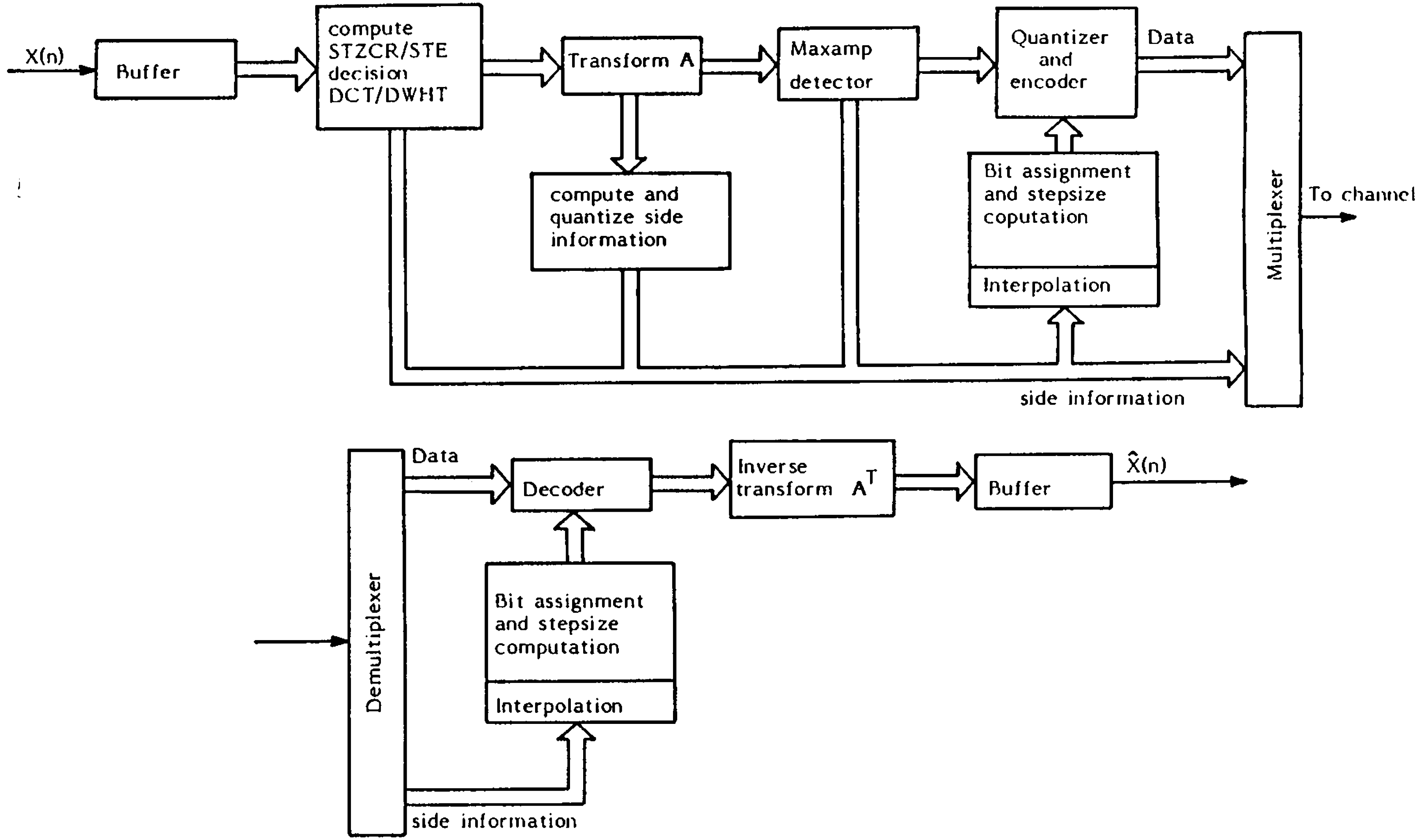


Fig. 7.1 Block diagram of Adaptive Switching Transform Coder

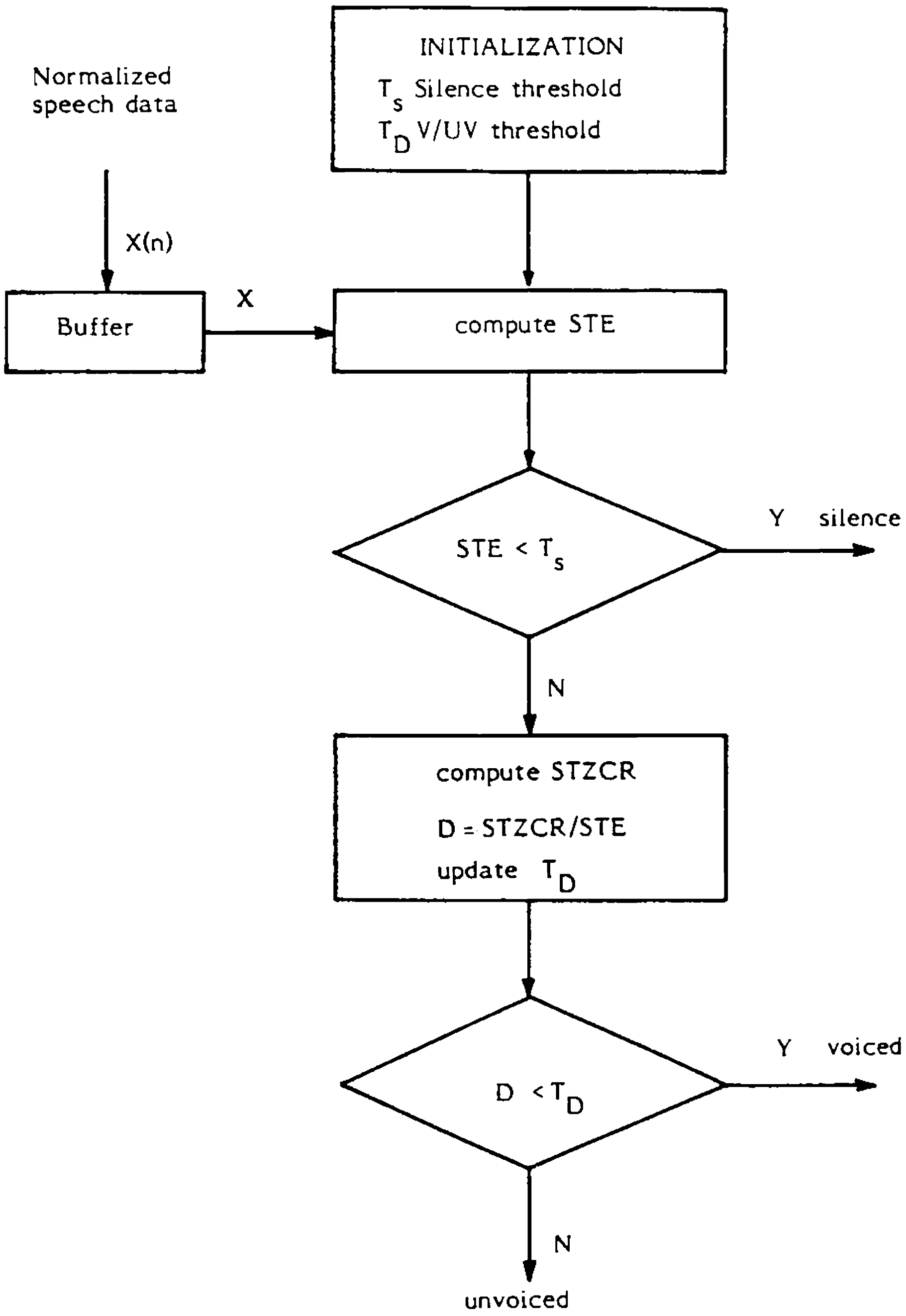


Fig. 7.2 Flow chart of the V/UV classification algorithm

The working of this algorithm is very simple and can be explained as follows. In the 'INITIALIZATION' process the algorithm chooses the initial values of the silence threshold T_s and the V/UV decision threshold T_D. (The initial values of T_s and T_D are experimentally obtained to be equal to 10^{-5} and $9x10^5$ respectively for the present data base). The block length is selected as N = 256.

Normalized speech data is segmented to consecutive blocks of 256 samples each. The processor first computes the short time energy of each segment and compares it with the silence threshold T_s. If the STE is less than T_s, then the segment is classified as silent segment. If STE is greater than T_s, then the short time zerocrossing rate is computed and hence the corresponding value of D is obtained. The processor then updates the threshold value T_D as 150 times the recent most minimum value (RMMV) of D. i.e.,$T_D = 150 \times \text{RMMV}$. If the D value for a particular segment exceeds T_D, then the segment is classified as unvoiced and otherwise as voiced.

7.3 THE DESIGN PROCEDURE FOR THE ADAPTIVE SWITCHING TRANSFORM CODING ALGORITHM

The complete design procedure for the adaptive switching transform coding (ASTC) algorithm can be summarized as follows.

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

2. If the segment is voiced, select the DCT.

3. If the segment is unvoiced, select the DWHT.

4. If the segment is silent then no transform.

5. Perform the Modified Adaptive Transform Coding (MATC).

For perfect decoding of the transmitted data at the receiver, the voiced/unvoiced/silent decision information is also considered as 'side information'. The flow chart of the ASTC design algorithm is shown in Fig.7.3.

7.4 EXPERIMENTAL RESULTS

This section describes the results of computer simulation of the ASTC coder developed in the previous sections.

The ASTC coder shown in Fig.7.1 is simulated on a 3AT6 computer with Turbo Pascal routines (the implementation details are discussed in section 7.2). Performance of the coder was evaluated using the two sets of speech

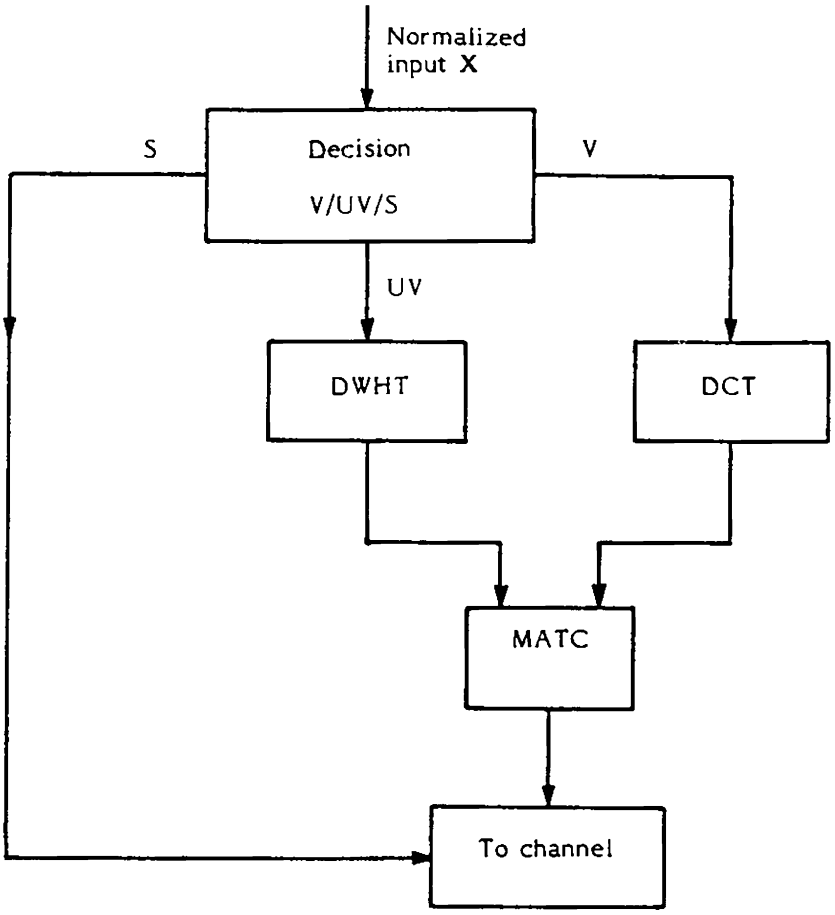


Fig. 7.3 Flow chart of the ASTC design algorithm

data base used in section 4.2.1. The total duration of this speech material is about 49 seconds. The speech material consisted of phonetically balanced sentences spoken by a male and a female speaker.

The coder is designed at six different bit rates. 8, 9.6, 12, 16, 24 and 32 kbits/s. The measured performance criteria are the SNR and the segmental SNR (SEGSNR) as defined in chapter 2 with segment size of 256 samples. Both the SNR and the SEGSNR are measured in the time domain, by calculating the distortion between the original and the coded speech segments.

Tables 7.1 and 7.2 present the SNR results for unvoiced segments, obtained using DCT based MATC and ASTC coders respectively. It may be noted that the SNR values of the unvoiced segments are increased by values ranging from 1.10 dB to 7.85 dB for the ASTC coder with respect to the MATC coder at a bit rate of 8 kbits/s. The SEGSNR results for all the speech material of 49 seconds duration, using the two coders MATC and ASTC are summarized in Table 7.3. About 0.7 dB increase in SEGSNR is noted at bit rate ranging from 8 to 16 kbits/s. The time dependence of the SNR values of the ASTC coder at 9.6 kbits/s is shown in Fig.7.4. The dotted line represents the SNR performance of the MATC coder.

Table 7.1 The SNR dB obtained for some UV segments using MATC coder for different bit rates

Unvoiced segment No.	Bit rate in kbits/s	8	9.6	12	16	24	32
1	SNR (dB)	0.4	0.4	0.53	2.14	5.37	6.91
2		2.31	2.31	2.38	2.51	4.95	11.03
3		0.77	1.63	2.88	3.85	7.80	13.74
4		0	0	1.46	7.05	8.65	11.46
5		1.59	2.78	3.86	4.80	7.81	12.95
6		4.53	6.16	7.62	7.62	9.04	14.55
7		1.54	1.88	2.79	4.74	6.44	8.71

Table 7.2 The SNR dB obtained for the same UV segments (table 7.1) using ASTC coder for different bit rates

Unvoiced segment No.	Bit rate in kbits/s	8	9.6	12	16	24	32
1	SNR (dB)	8.25	9.30	11.15	9.76	12.39	16.30
2		7.36	8.28	8.88	9.67	11.27	12.08
3		8.06	8.41	10.56	10.65	10.97	15.77
4		7.72	8.50	7.90	8.65	11.41	12.61
5		5.45	6.07	6.96	8.69	10.41	11.21
6		5.63	6.80	8.44	11.94	14.51	16.70
7		6.14	10.20	11.73	9.88	14.52	16.05

Table 7.3 Segmental - SNR in dB for MATC and ASTC coders at different bit rates

Coder type	Bit rate in kbits/s	8	9.6	12	16	24	32
MATC	SEGSNR (dB)	9.24	10.59	12.34	14.28	17.83	22.31
ASTC		9.95	11.38	13.10	14.96	18.21	22.59

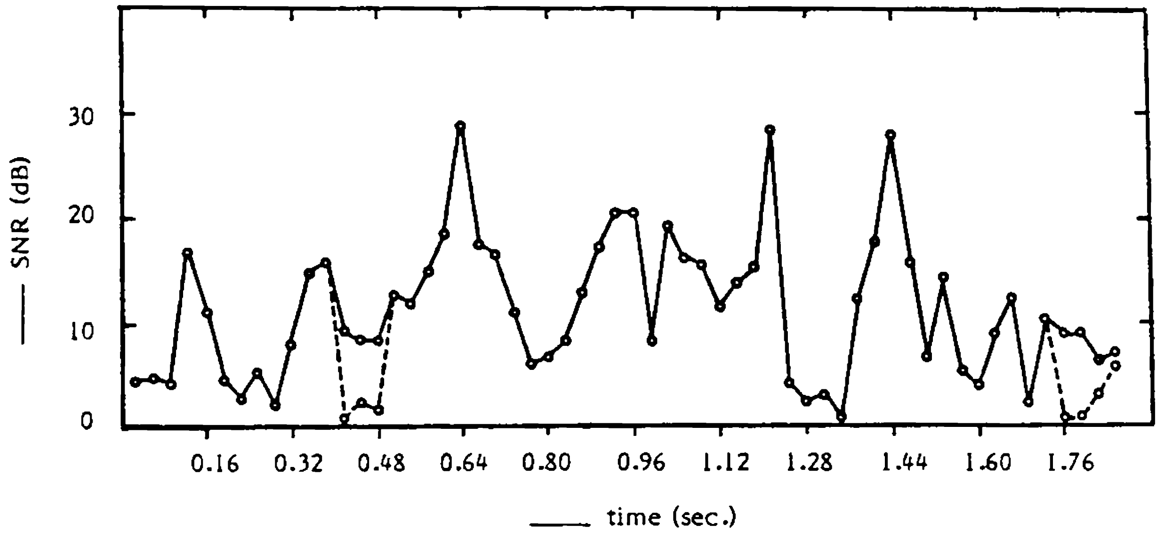


Fig. 7.4 Time dependence of the SNR values of the ASTC coder at 9.6 kbits/s (dotted line represents the SNR performance of the MATC coder).

The perceptual quality of the coded speech at 9.6 kbits/s encoding rate using MATC and ASTC is examined and compared by subjective listening tests. This is carried out with sixteen listeners. The nine sentences, each spoken by a male and a female, given in table 4.2, are examined.

The original speech, MATC coded speech and ASTC coded speech are presented to the listeners in that order. The test was blind for the listeners, as they did not know whether the material was MATC coded speech or ASTC coded speech. They are asked to evaluate the quality of the coded speech by comparing with the original on the basis of its clarity, crispness, hoarseness and warbling effect. They are also asked to give their preference in comparing the MATC coded speech with the ASTC coded speech.

To obtain the Mean Opinion Score (MOS), the listeners are asked to rate both the coded speech on an absolute scale, ranging between 1 and 5 by comparing with the original. The meaning of these grades are

5. excellent
4. good
3. fair
2. poor
1. bad.

The MOS rated for the MATC coded speech was 4.38 while that for ASTC coded speech was 4.62. Thus the 9.6 kbits/s coded speech using both the coders sounds very close to toll quality. All the listeners preferred the ASTC coded speech over the MATC coded speech. This is also indicated by the higher value of MOS rated for the ASTC coded speech.

The results of this simulation experiment indicate that an improvement in performance can be obtained by using the simple V/UV classifier and thereby adaptively switching for DCT and DWHT. Of course, this performance can be further improved by using a more complex classifier which would observe other parameters in addition to STZCR and STE. This approach could be extended to the point where a different transform is used for each acoustic class of speech sounds: fricatives, nasals, plossives, etc. But this will surely enhance the coder complexity in a considerable dimension. The development of such a classifier would be of interest to speech recognition research and is beyond the purpose of the present work.

To summarize, the use of the adaptive switching transform coder, based on a simple V/UV classification algorithm gives a notable improvement in performance at the expense of a moderate increase in coder complexity. The focus of this study is on the effectiveness of the ASTC coder as measured by the objective criterion like the signal to noise ratio. Therefore, we avoided using techniques which might prevent accurate determination of the contribution of the ASTC coder alone. Such techniques like block overlapping, noise shaping, post-filtering etc. may be added to the ASTC system to improve the subjective quality of the coded speech.

Chapter 8

CONCLUSIONS

In this work, we have introduced a new technique for estimating speech samples from their zerocrossings. This technique is particularly useful for designing low complexity digital communication systems. The conventional A/D converter circuitry for digitizing the analog speech signal can be replaced with simple digital circuits. This will enable the reduction of the cost of digital communication systems.

In addition to the use of zerocrossing information for speech sample estimation, we have studied its use for speech signal classification. A simple time domain algorithm that uses a distance measure based on zerocrossing rate and energy is developed. A nice feature of this algorithm is that it takes into account the dynamic range and speaker dependency of speech signal. This algorithm is used to design an enhanced Adaptive Transform Coder in this thesis.

The investigations on the attractor dimension and entropy of speech signal also provided another new method for speech signal classification. The knowledge about the attractor dimension and entropy can be utilized in recognition purpose. It is proposed to carry on further research with the intent of finding efficient set of these features for speech recognition. It is also of interest to investigate their use in other pattern recognition areas.

The new coding scheme, Adaptive Switching Transform Coder presented in chapter 7 is the result of the investigations to improve the performance of conventional Adaptive Transform Coding systems. This scheme uses the Discrete Cosine Transform to encode voiced speech and Discrete Walsh-Hadamard Transform for unvoiced speech. A notable improvement in the performance on the conventional Adaptive Transform Coding system is achieved in this work. An increase in SNR value ranging from 1.10 dB to 7.85 dB is obtained for unvoiced speech segments. This enables a reduction in the tonal distortion present in conventional Adaptive Transform Coding systems at bit rates below 16 kbits/s. This system performs well even at 8 kbits/s. In this coder the improvement in SNR is obtained by using the simple voiced unvoiced classifier developed in chapter 4 and thereby adaptively switching between Discrete Cosine Transform and Discrete Walsh-Hadamard Transform. Further research work may be conducted to improve this performance by using a more complex classifier which would observe other parameters in addition to the short-time zerocrossing rate and short time energy of speech signal. Different transforms can be used for each acoustic class of speech sounds. The development of the above mentioned classifier would be of interest to speech recognition research also.

APPENDIX I

PASCAL PROGRAMS DEVELOPED

1. Zerocrossing detection routines
2. TIF implementation
3. SNR (dB) computation
4. FFT computation
5. Histogram plotting
6. Signal plotting
7. FFT plotting
8. Gaussian noise generation
9. Chi-square test
10. Second order attractor dimension and entropy computation
11. STZCR and STE computation and V/UV classification
12. DCT (using 2N point FFT) computation
13. DWHT computation
14. ATC implementation
15. ASTC implementation

APPENDIX II

Speech waveforms of the utterances in Table 4.2

(a) - male, (b) - female)

Fig.1 An icy wind racked the beach

Fig.2 The pipe began to rust while new

Fig.3 Cats and dogs hate each the other

Fig.4 Oak is strong and also gives shade

Fig.5 Thieves who rob friends deserve jail

Fig.6 Open the crate but do not break the glass

Fig.7 Add the sum to the product of these three

Fig.8 Joe brought a young girl

Fig.9 A lathe is a big tool.

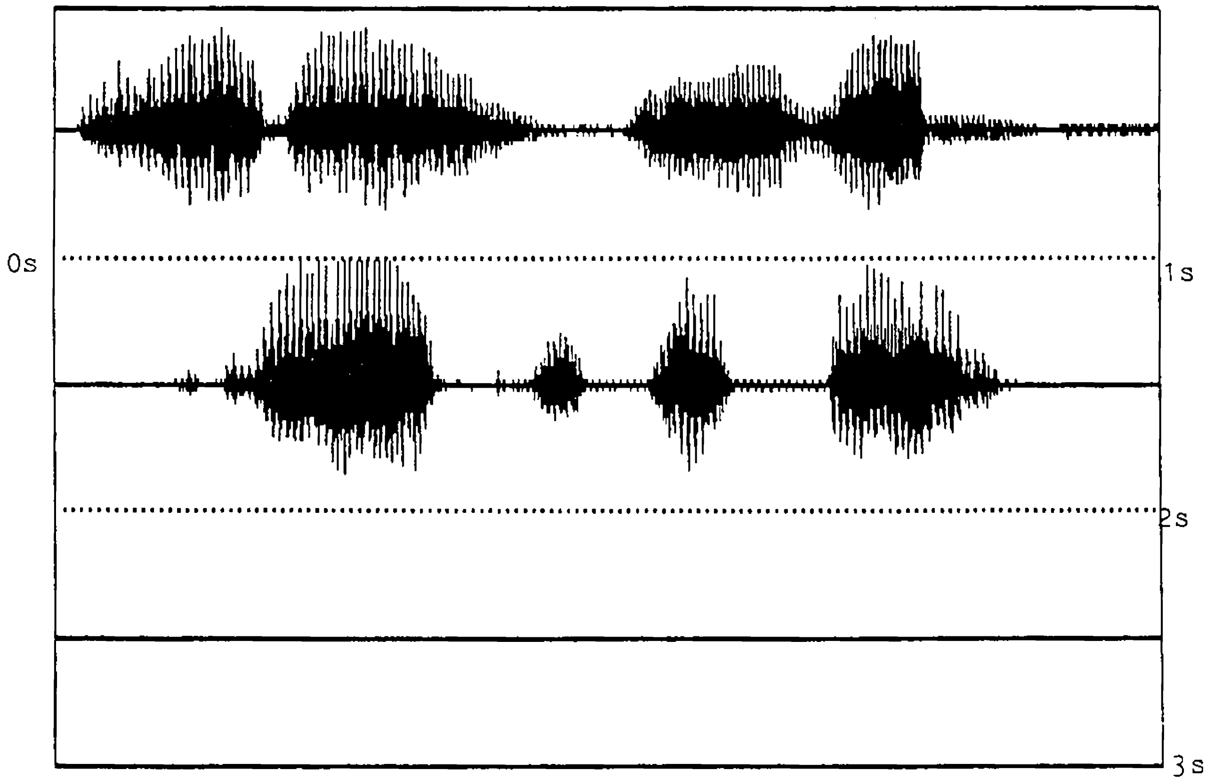


Fig.1 (a)

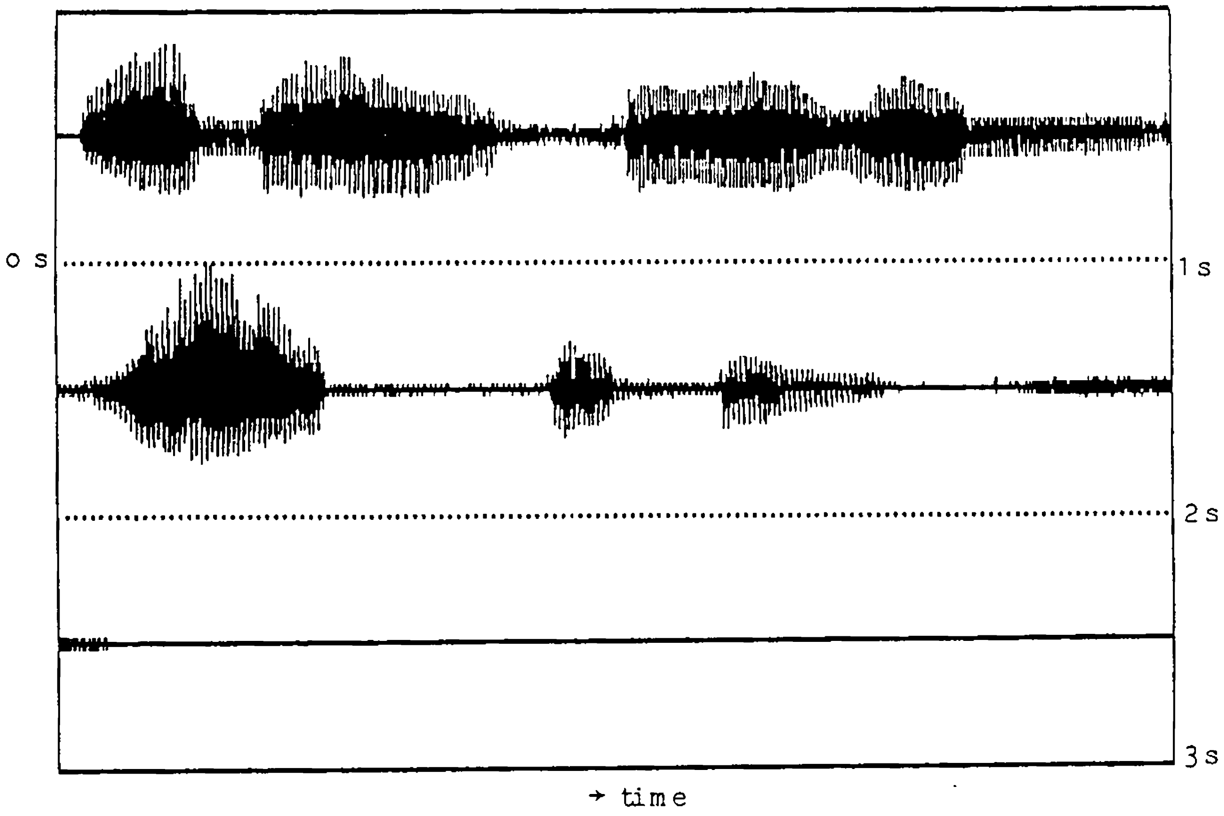


Fig. 1(b)

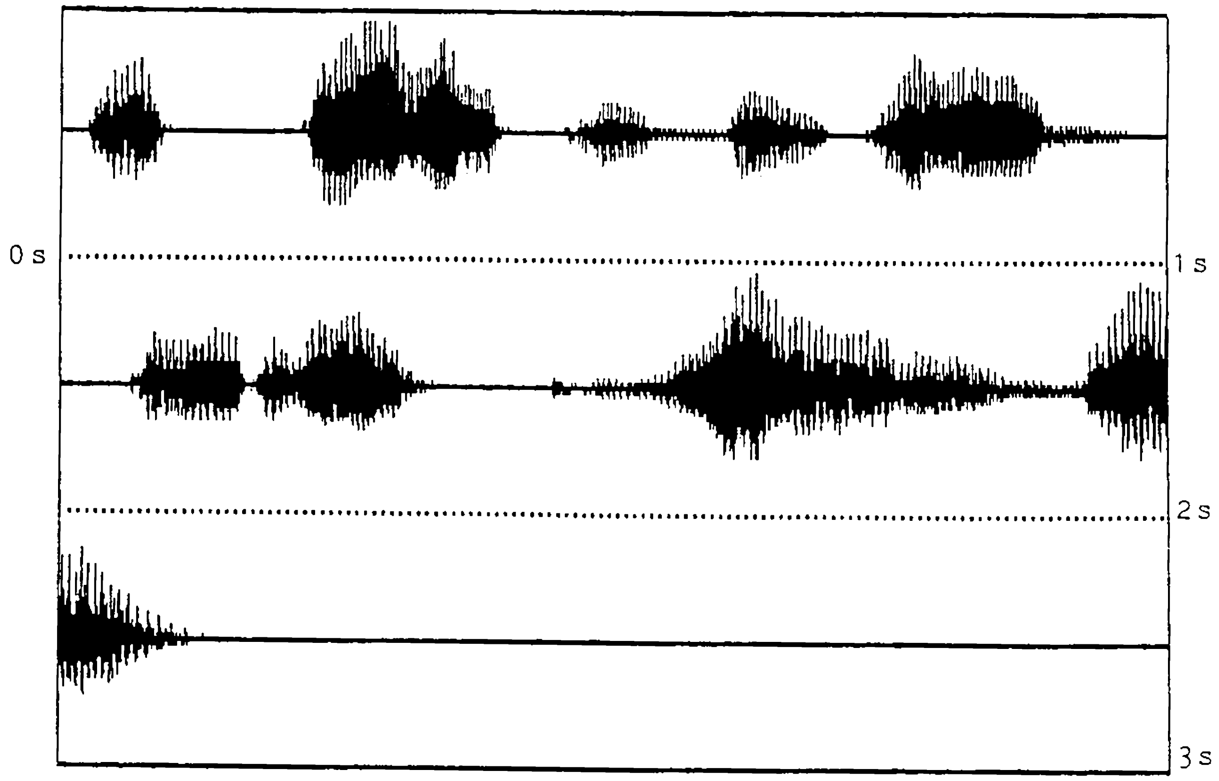


Fig. 2(a)

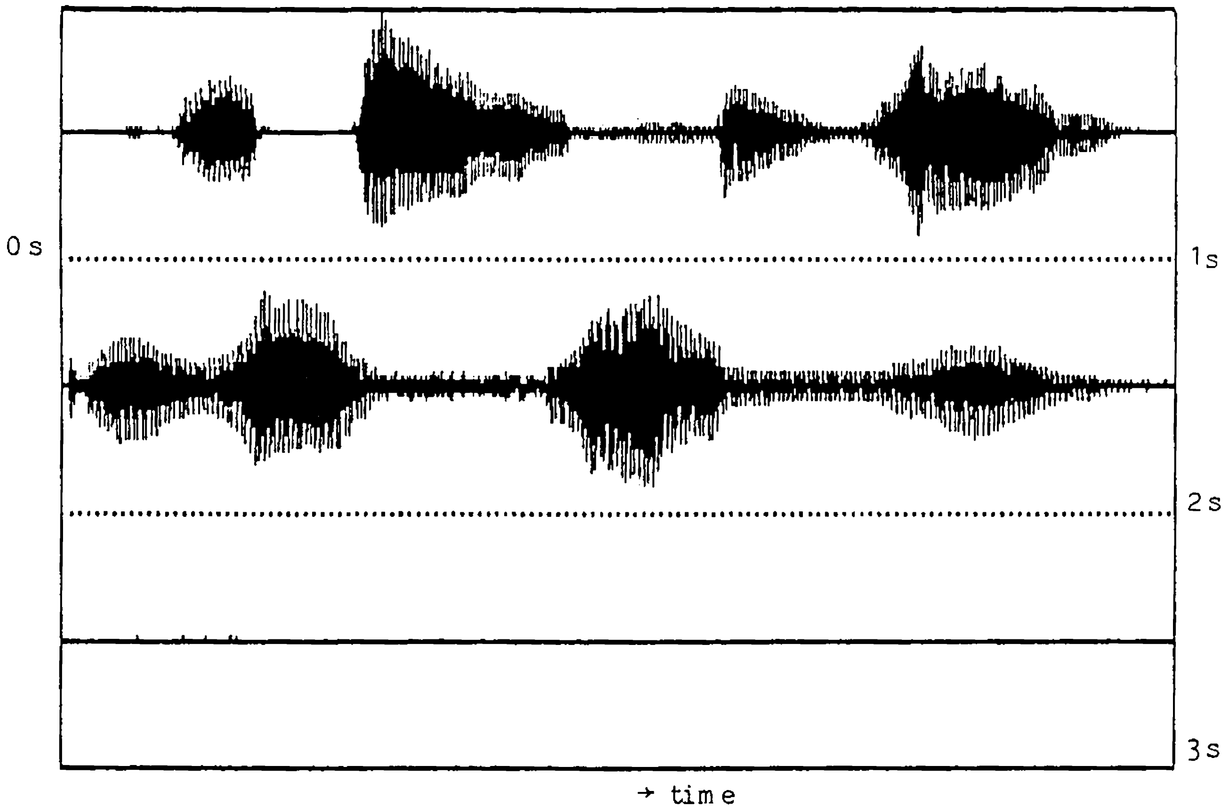


Fig. 2(b)

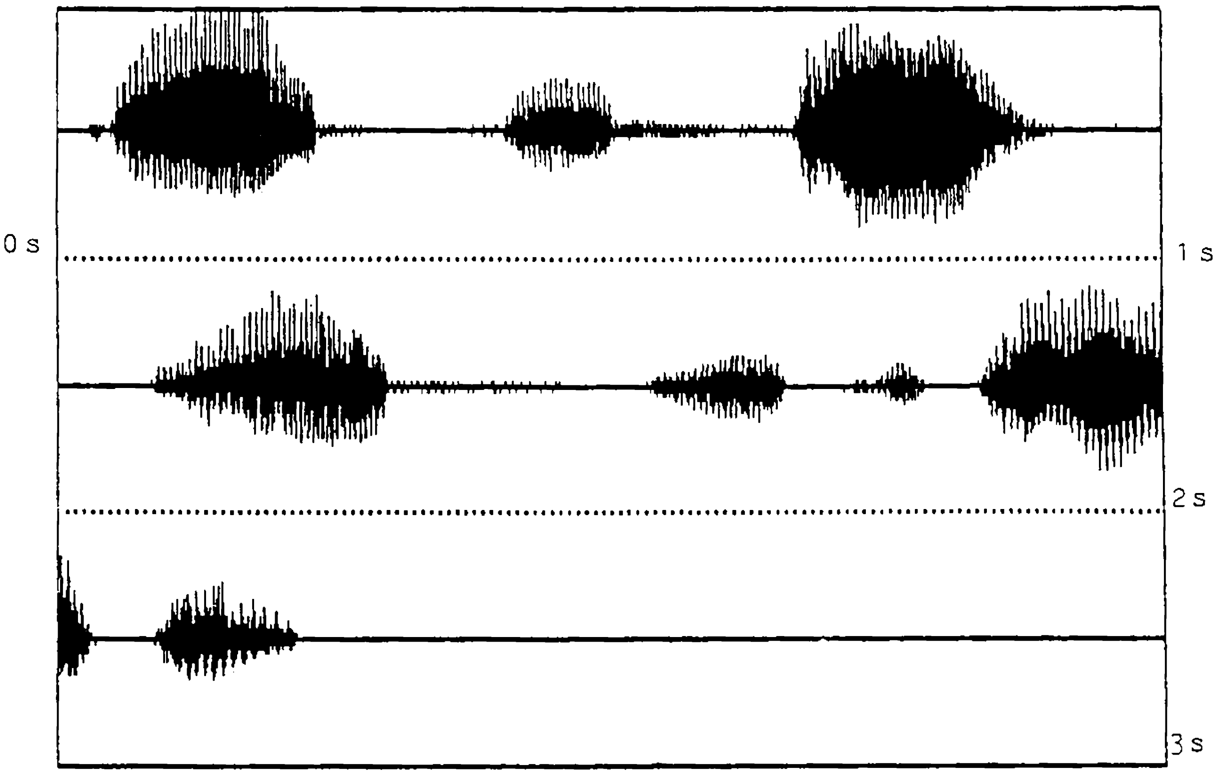


Fig. 3(a)

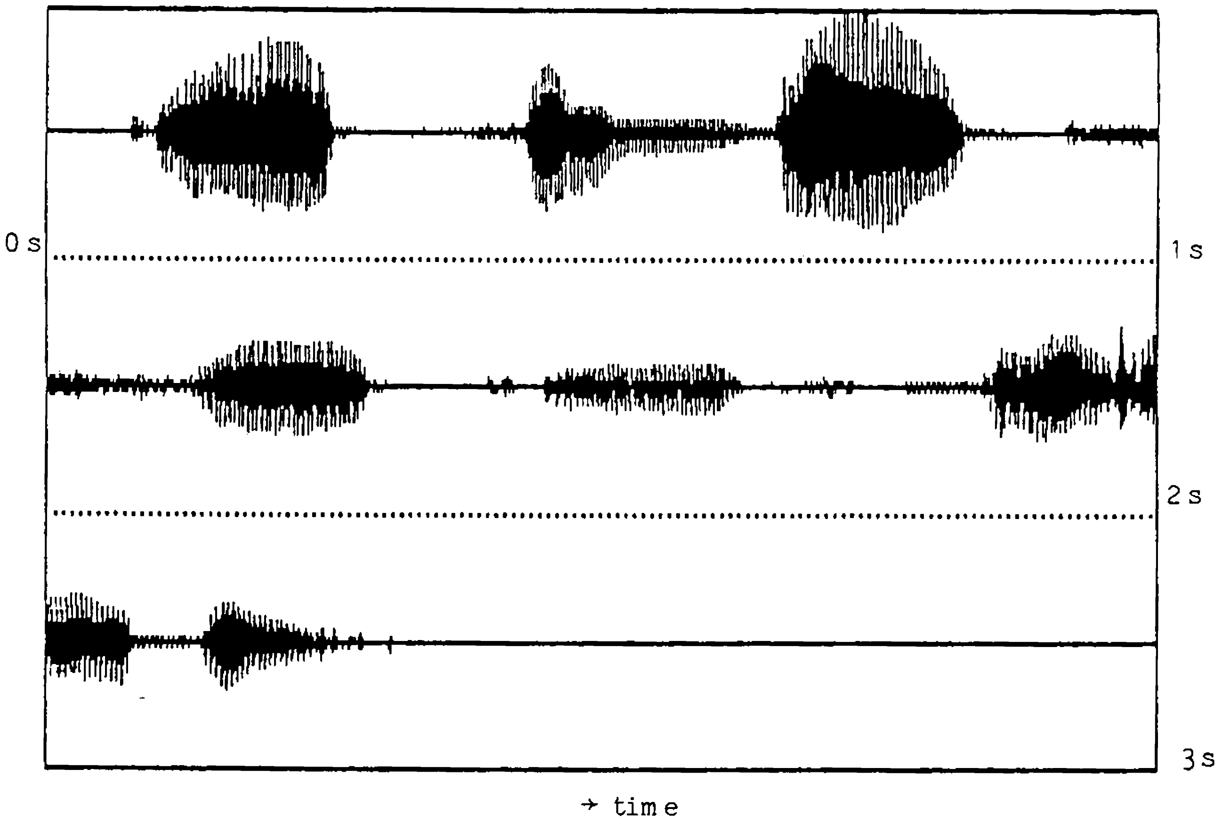


Fig. 3(b)

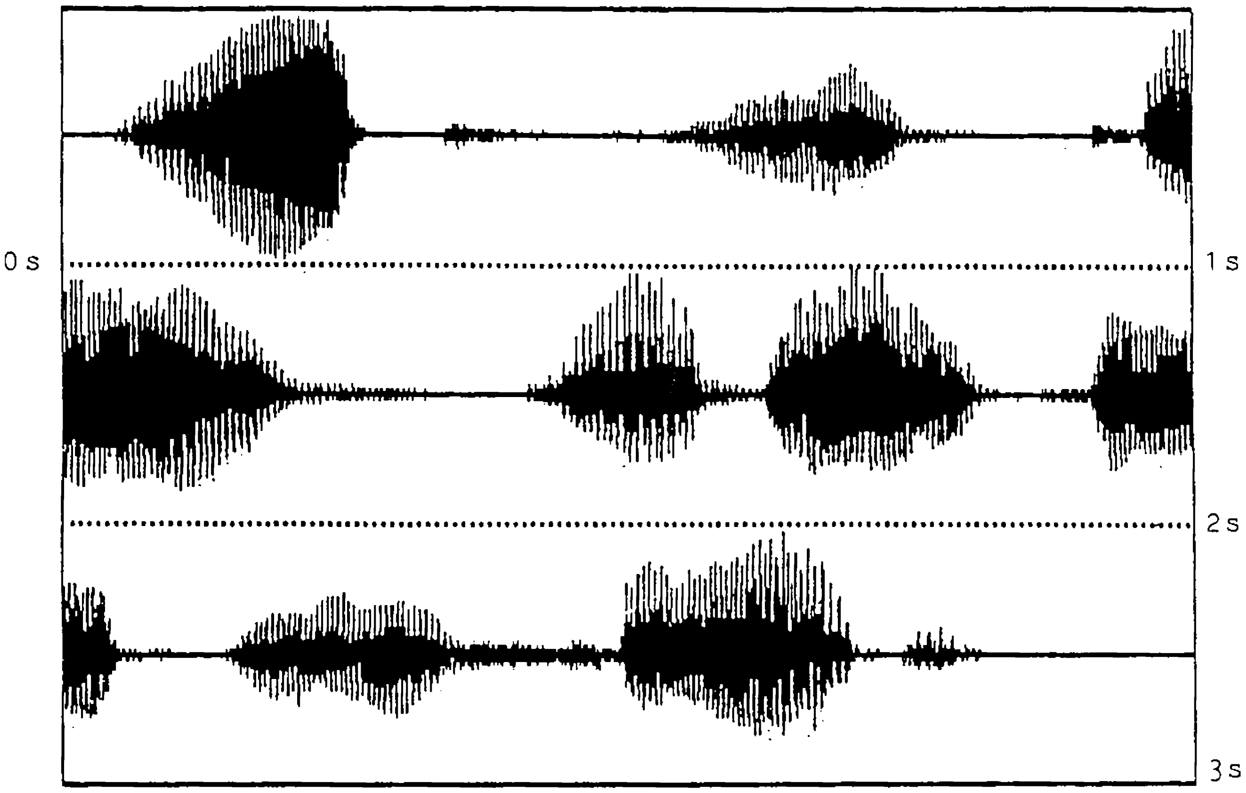
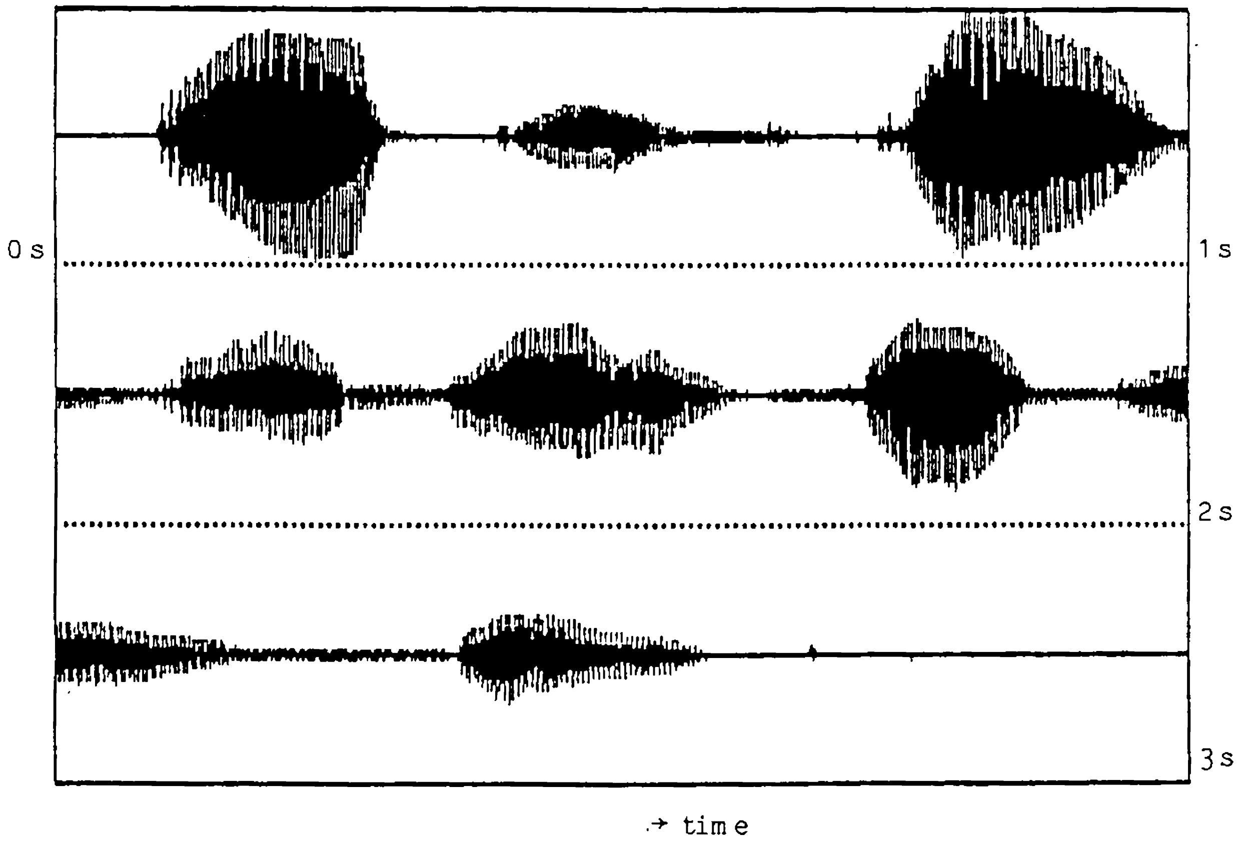


Fig. 4(a)



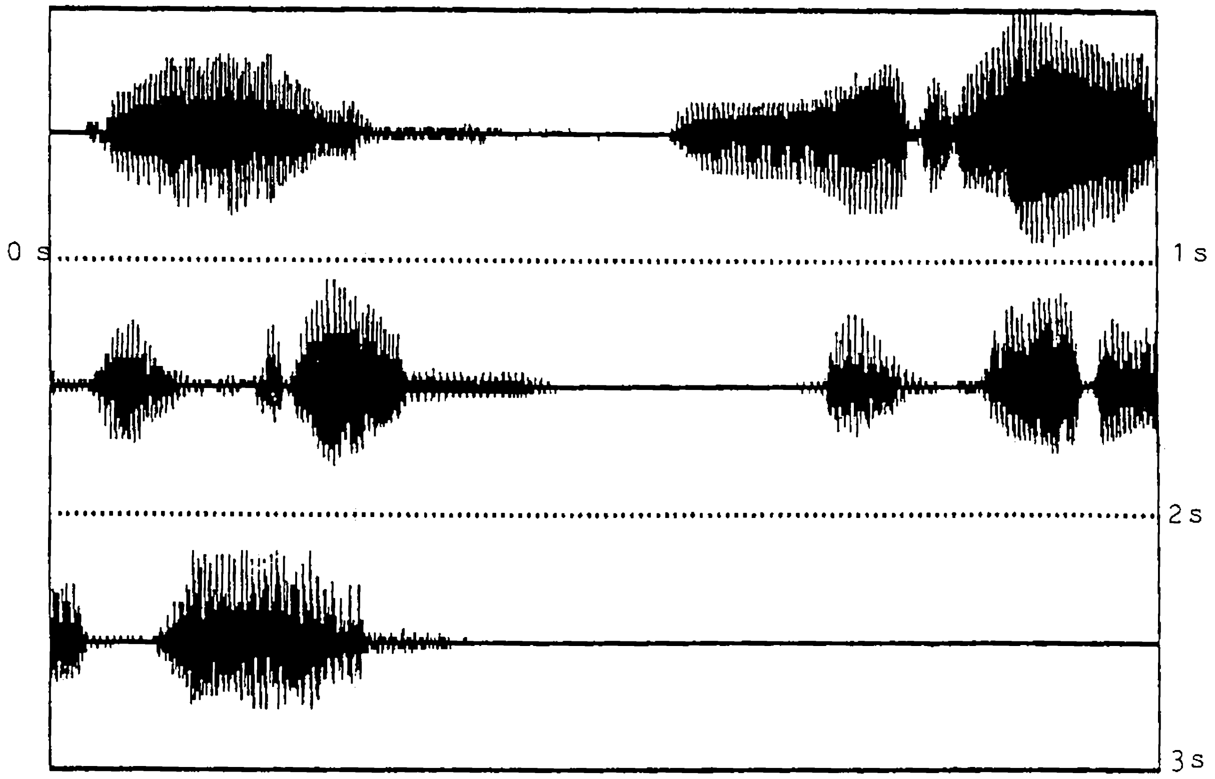
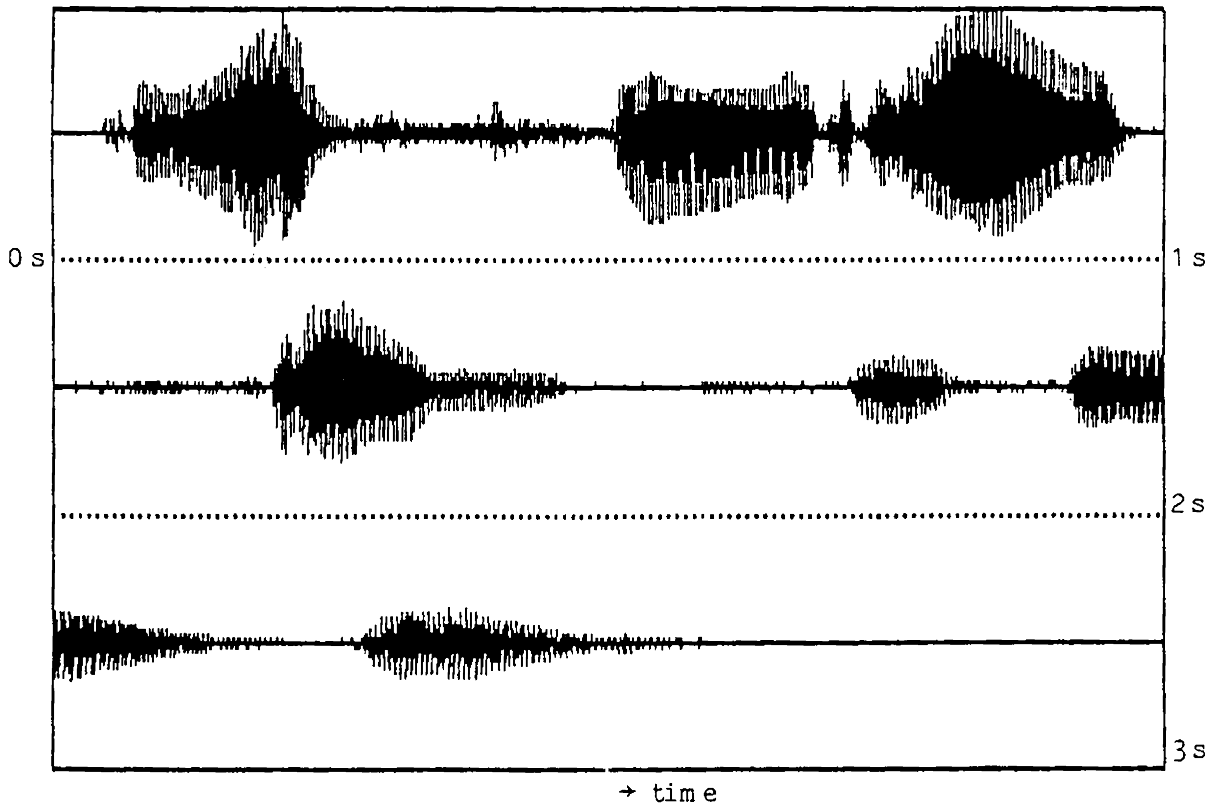


Fig. 5(a)



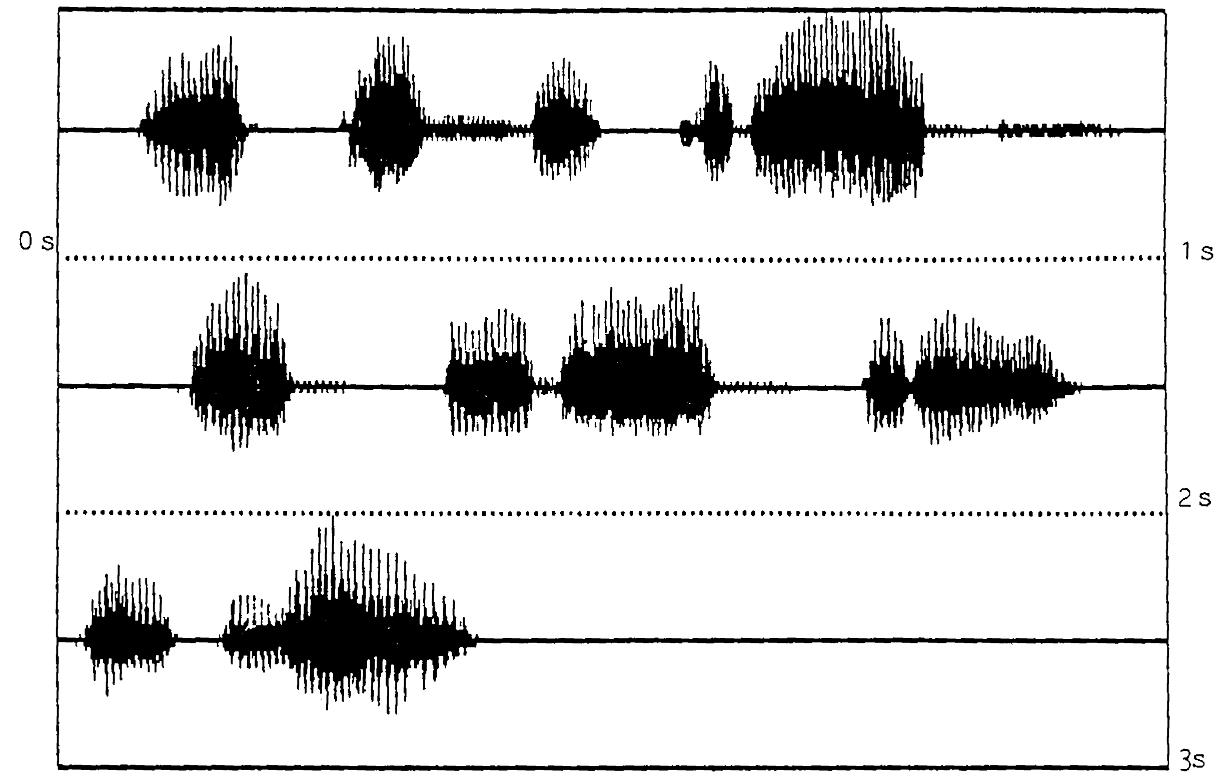
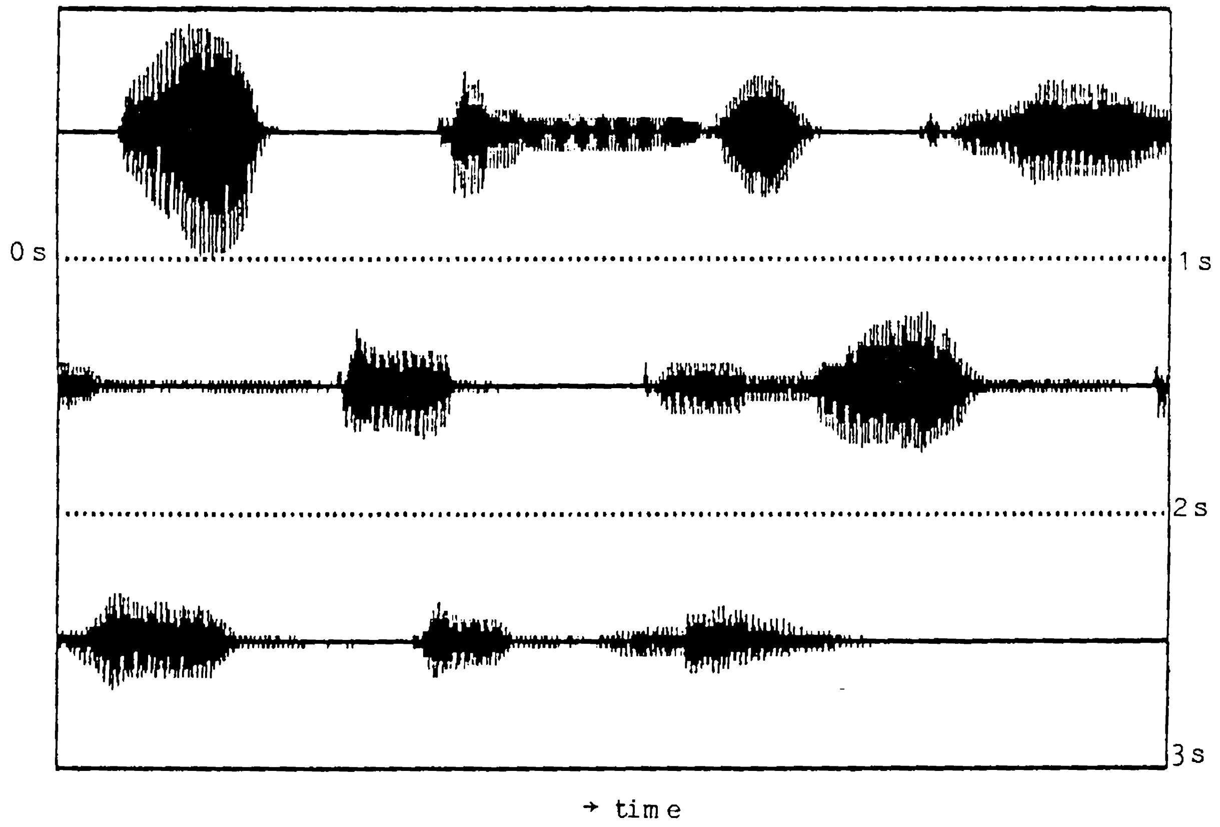


Fig. 6(a)



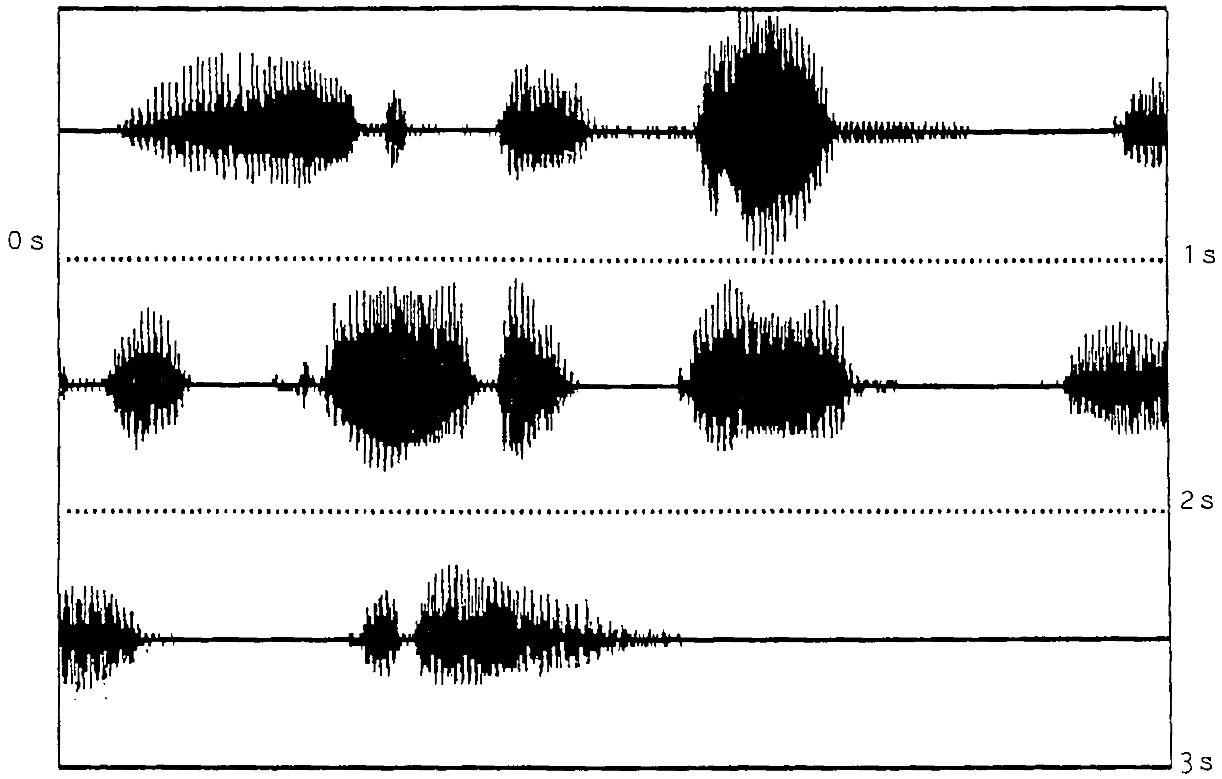


Fig. 7(a)

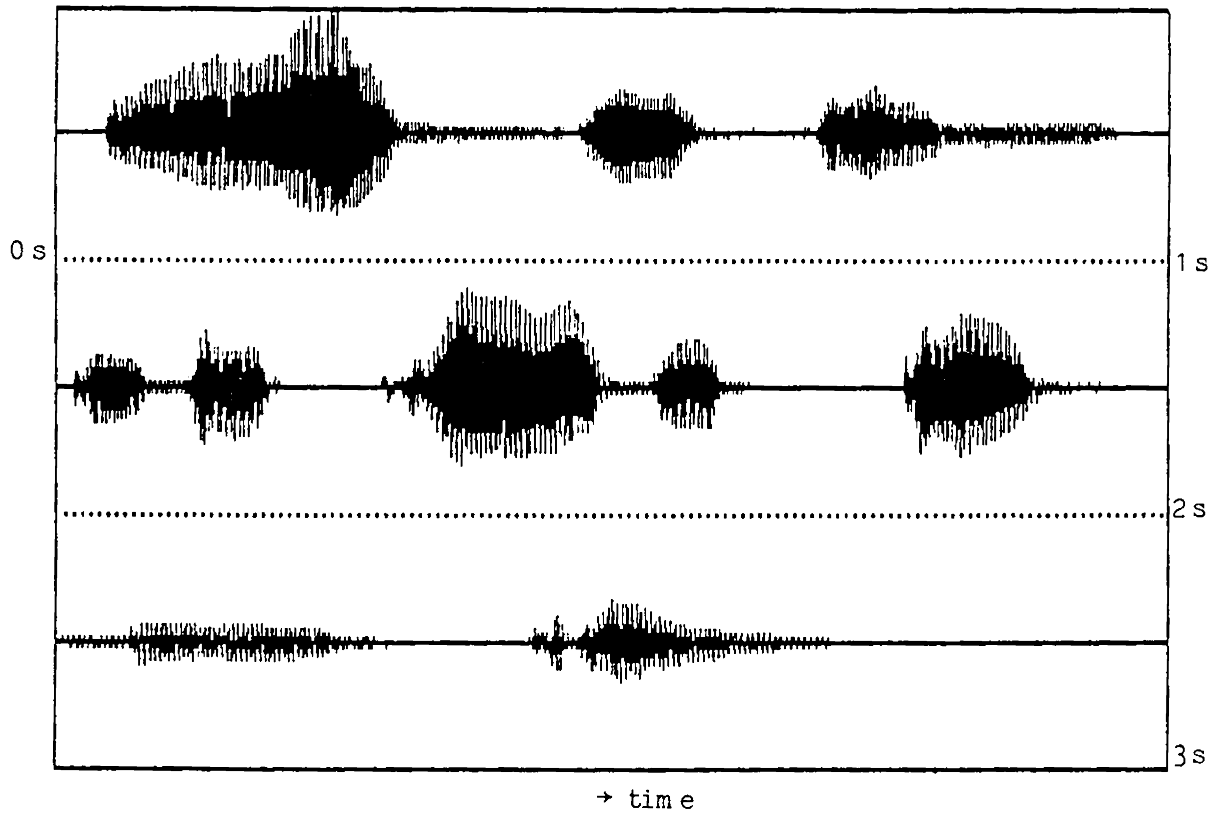


Fig. 7(b)

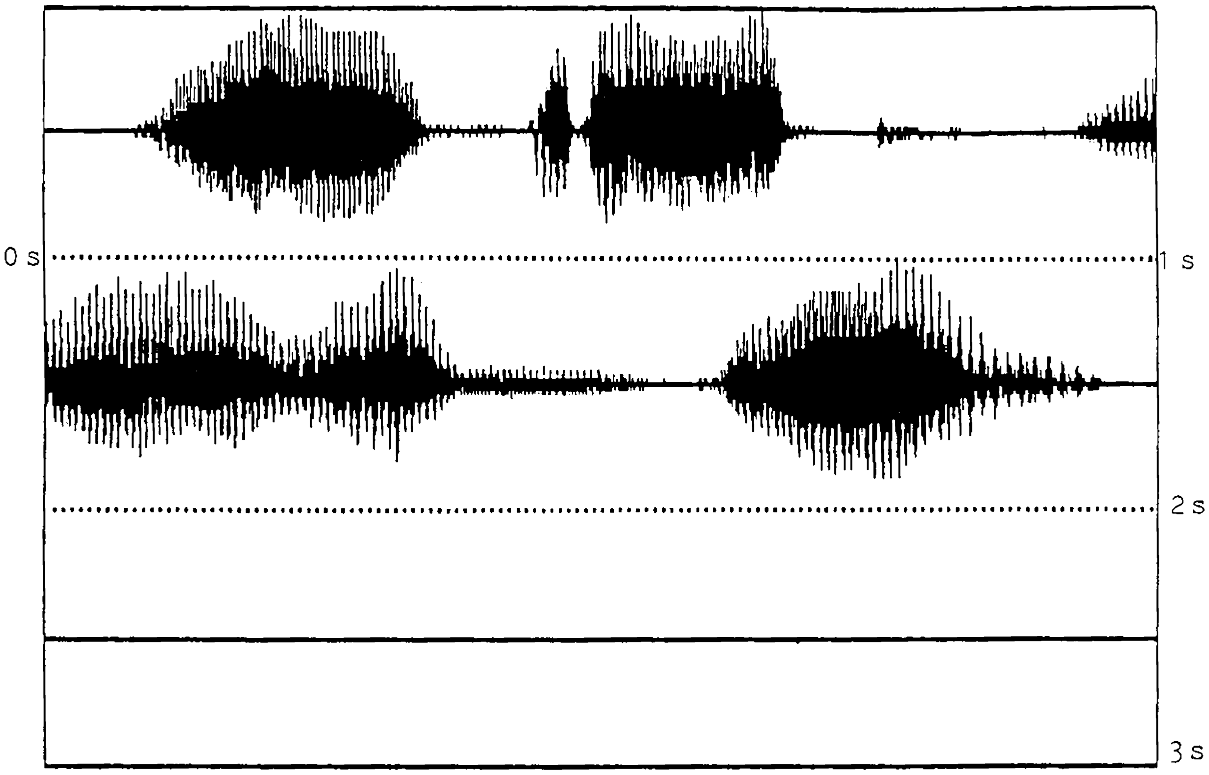
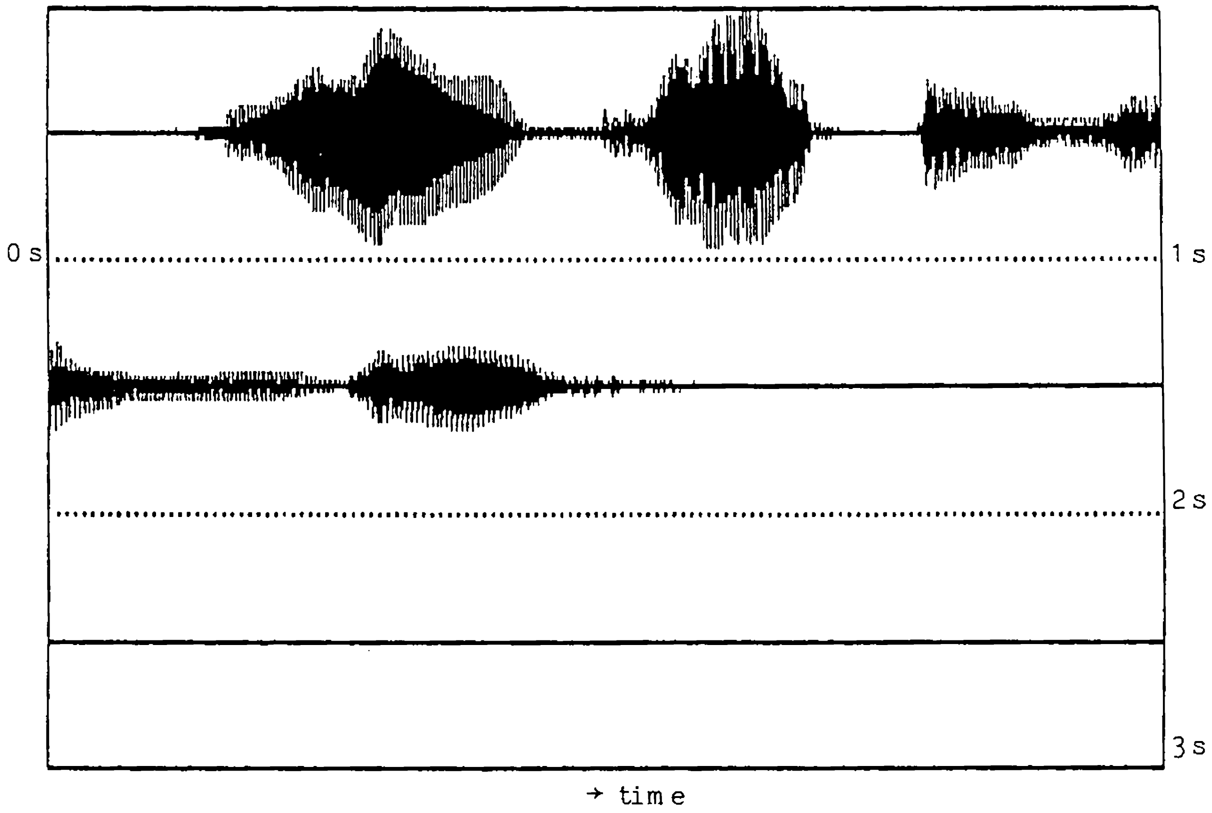


Fig. 8(a)



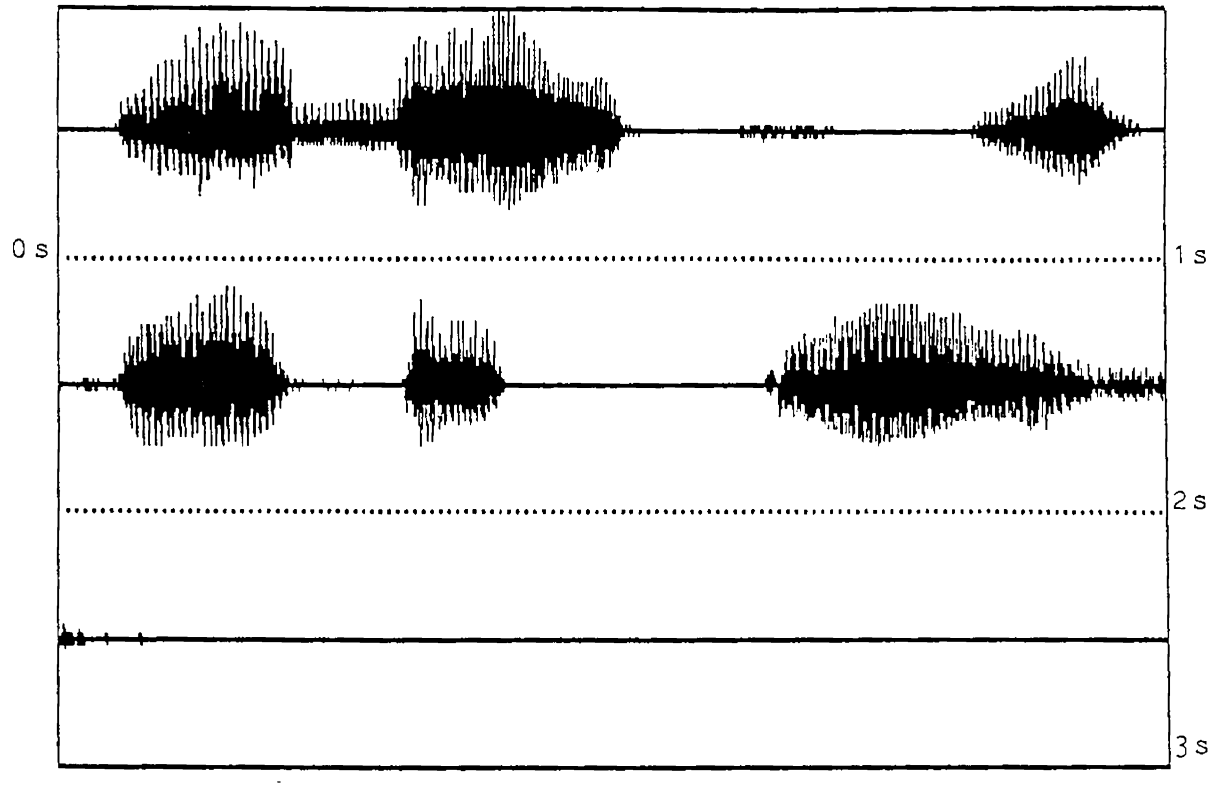
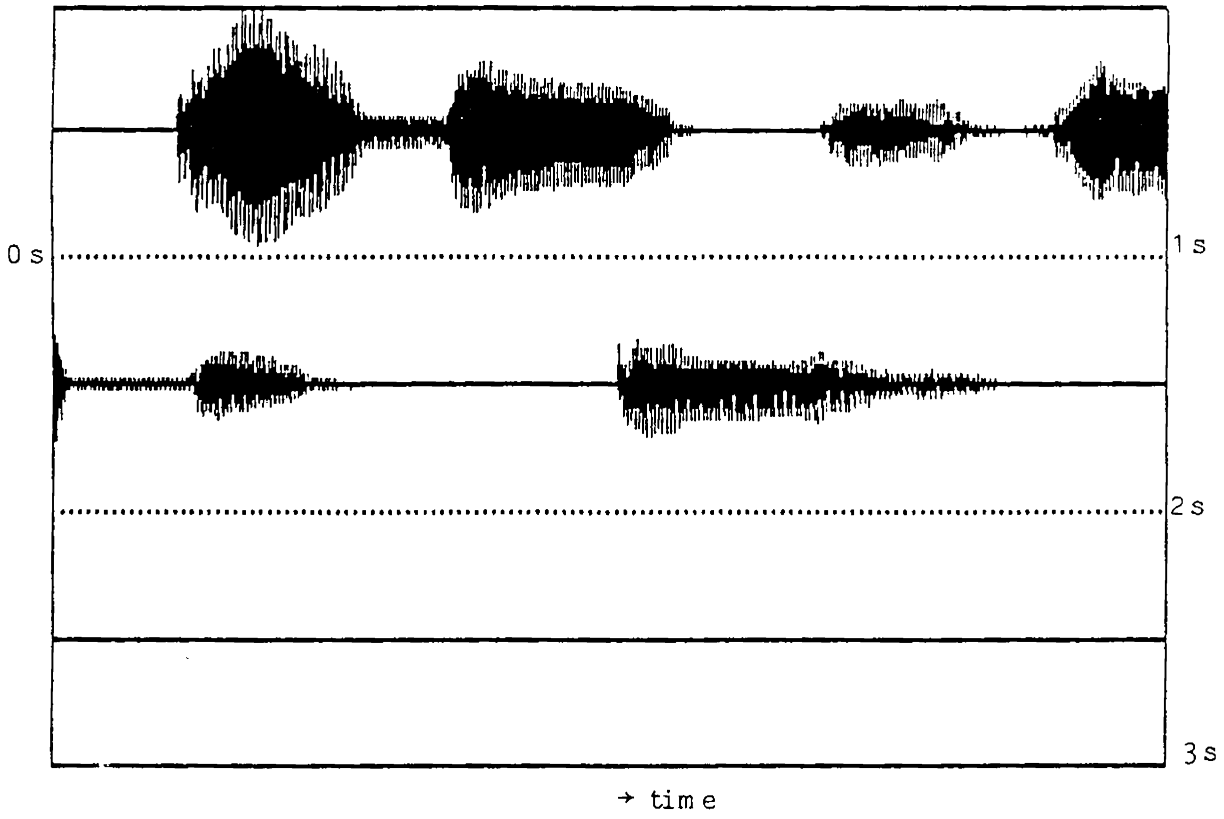


Fig. 9(a)



REFERENCES

1. H.Abut, R.M.Gray and G.Rebolledo, "Vector Quantization of Speech and Speech-Like Waveforms", IEEE Trans.Acoust., Speech, Signal Processing, Vol.ASSP-30, pp.423-436, 1982.

2. N.Ahmed, T.Natarajan and K.R.Rao, "Discrete Cosine Transform", IEEE Trans.Computers, Vol.C-23, pp.90-93, 1974.

3. W.A.Ainsworth, "Mechanisms of Speech Recognition", Pergamon Press, 1976.

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

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

6. H.Atmanspacher and H.Scheingraber, "Deterministic Chaos and Dynamical Instabilities in a Multimode CW Dye Laser", Physical Review A, Vol.34, No.1, pp.253-263, 1986.

7. Babu P.Anto, N.K.Narayanan and C.S.Sridhar, "Use of Zerocrossing Information for the Extraction of Speech Parameters", J.Acoust.Soc.Ind., Vol.15, pp.96-101, 1987.

8. R.H.T.Bates and A.R.Murch, "Deterministic-Chaotic Variably Coloured Noise", Electronics Letters, Vol.23, No.19, pp.995-996, 1987.

9. F.E.Bond and C.R.Cahn, "On Sampling the Zeros of Bandwidth Limited Signals", IRE Trans.Inform.Theory, Vol.IT-4, pp.110-113, 1958.

10. W.H.Chen, C.H.Smith and S.C.Fralick, "A Fast Computational Algorithm for the Discrete Cosine Transform", IEEE Trans. Commun., Vol.COM-25, No.9, pp.1004-1009, 1977.

11. R.V.Cox and R.E.Crochiere, "Real-Time Simulation of Adaptive Transform Coding", IEEE Trans.Acoust., Speech, Signal Processing, Vol.ASSP-29, pp.147-154, 1981.

12. R.E.Crochiere, S.A.Webber and J.L.Flanagan, "Digital Coding of Speech in Sub-bands", BSTJ, Vol.55, pp.1069-1085, 1976.

13. R.E.Crochiere and J.M.Tribolet, "Frequency Domain Techniques for Speech Coding", J.Acoustic.Soc.Amer., Vol. 66(6), pp.1642-1646, 1979.

7. Babu P.Anto, N.K.Narayanan and C.S.Sridhar, "Use of Zerocrossing Information for the Extraction of Speech Parameters", J.Acoust.Soc.Ind., Vol.15, pp.96-101, 1987.

8. R.H.T.Bates and A.R.Murch, "Deterministic-Chaotic Variably Coloured Noise", Electronics Letters, Vol.23, No.19, pp.995-996, 1987.

9. F.E.Bond and C.R.Cahn, "On Sampling the Zeros of Bandwidth Limited Signals", IRE Trans.Inform.Theory, Vol.IT-4, pp.110-113, 1958.

10. W.H.Chen, C.H.Smith and S.C.Fralick, "A Fast Computational Algorithm for the Discrete Cosine Transform", IEEE Trans. Commun., Vol.COM-25, No.9, pp.1004-1009, 1977.

11. R.V.Cox and R.E.Crochiere, "Real-Time Simulation of Adaptive Transform Coding", IEEE Trans.Acoust., Speech, Signal Processing, Vol.ASSP-29, pp.147-154, 1981.

12. R.E.Crochiere, S.A.Webber and J.L.Flanagan, "Digital Coding of Speech in Sub-bands", BSTJ, Vol.55, pp.1069-1085, 1976.

13. R.E.Crochiere and J.M.Tribolet, "Frequency Domain Techniques for Speech Coding", J.Acoustic.Soc.Amer., Vol. 66(6), pp.1642-1646, 1979.

14. V.Cuperman and A.Gersho, "Adaptive Differential Vector Coding of Speech", Conf.Rec., GLOBECOM 82, pp.1092-1096, 1982.

15. D.F.Elliott and K.R.Rao, "Fast Transforms-Algorithms, Analysis, Applications", Academic Press, 1982.

16. 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.

17. 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.

18. A.Gersho, T.Ramstad and I.Versvik, "Fully Vector-Quantized Sub-band Coding with Adaptive Codebook Allocation", Proc.IEEE Int.Conf., ICASSP-84, pp.10.7.1-10.7.4, 1984.

19. R.M.Gray and H.Abut, "Full Search and Tree Search Vector Quantization of Speech Waveforms", Proc.IEEE Int.Conf., ICASSP-82, pp.593-596, 1982.

20. S.Haavik, "The Conversion of Zeros of Noise", M.S.Thesis, Univ.Rochester, Rochester, NY, 1966.

21. M.Hamidi and J.Pearl, "Comparison of the Cosine and Fourier Transforms of Markov-1 Signals", IEEE Trans.

Acoust., Speech, Signal Processing, Vol.ASSP-24, pp.428-429, 1976.

22. R.M.Haralick, "A Storage Efficient Way to Implement the Discrete Cosine Transform", IEEE Trans.Computers, Vol.C-25, pp.764-765, 1976.

23. J.Huang and P.Schultheiss, "Block Quantization of Correlated Gaussian Random Variables", IEEE Trans.Commun.Sys., Vol.CS-11, pp.289-296, 1963.

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

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

26. N.S.Jayant and P.Noll, "Digital Coding of Waveforms: Principles and Applications to Speech and Video", Prentice Hall, 1984.

27. N.S.Jayant, Private Communication.

28. S.M.Kay and R.Sudhakar, "A Zerocrossing Based Spectrum Analyzer", IEEE Trans.Acoust., Speech, Signal Processing, Vol.ASSP-34, No.1, pp.96-104, 1986.

29. S.Knoor, "Reliable Voiced/Unvoiced Decision", IEEE Trans. Acoust., Speech, Signal Processing, Vol.ASSP-27, pp.263-267, 1979.

30. S.Krishnan and K.K.Paliwal, "System Design Consideration in Adaptive Transform Coding of Speech", J.Acoust.Soc.Ind., Vol.15, pp.125-129, 1987.

31. J.C.R.Licklider, "Effects of Amplitude Distortion upon the Intelligibility of Speech", J.Acoust.Soc.Amer., Vol.18, p.429, 1946.

32. J.C.R.Licklider and I.Pollack, "Effects of Differentiation, Integration and Infinite Peak Clipping upon the Intelligibility of Speech", J.Acoust.Soc.Amer., Vol.20, p.42, 1948.

33. J.C.R.Licklider, "The Intelligibility of Amplitude-Dichotomized, Time-Quantized Speech Waves", J.Acoust.Soc.Amer., Vol.22, pp.820-823, 1950.

34. J.Makhoul, "A Fast Cosine Transform in One and Two Dimensions", IEEE Trans.Acoust., Speech, Signal Processing, Vol.ASSP-28, No.1, pp.27-34, 1980.

35. L.R.Morris, "The Role of Zerocrossings in Speech Recognition and Processing", Ph.D.Dissertation, Dep.Elec.Eng., Imperial College Sci.Technol., Univ.London, England, 1970.

36. L.R.Morris, "The Role of Zerocrossings in Speech Recognition and Processing", Proc.IEEE Conf.Speech Commun. Processing, pp.446-450, 1972.

37. M.J.Narasimha and A.M.Peterson, "On the Computation of the Discrete Cosine Transform", IEEE Trans.Commun., Vol.COM-26, No.6, pp.934-936, 1978.

38. R.J.Niederjohn, M.W.Krutz and B.M.Brown, "An Experimental Investigation of the Perceptual Effects of Altering the Zerocrossings of a Speech Signal", IEEE Trans.Acoust., Speech, Signal Processing, Vol.ASSP-35, No.5, pp.618-625 ,1987.

39. L.R.Rabiner and R.W.Schafer, "Digital Processing of Speech Signals", Printice-Hall, NJ, 1978.

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

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

42. A.A.G.Requicha, "The Zeros of Entire Functions Theory and Engineering Applications", IEEE Proc., Vol.68, pp.308-328, 1980.

43. A.Seeky, "A Computer Simulation Study of Real-Zero Interpolation", IEEE Trans.Audio-Electroacoust., Vol.AU-18, pp.43-54, 1970.

44. A.Segall, "Bit Allocation and Encoding for Vector Sources", IEEE Trans.Inf.Theory, Vol.IT-22, No.2, pp.162-169, 1976.

45. C.S.Sridhar, Ph.D.Thesis, Dep.Elec.Eng., IIT, Madras, 1975.

46. 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.

47. Y.Todakoro and T.Higuchi, Comments on "Discrete Fourier Transform via Walsh Transform", IEEE Trans.Acoust., Speech, Signal Processing, Vol.ASSP-27, pp.295-296, 1979.

48. J.M.Tribolet and R.E.Crochiere, "A Vocoder-Driven Adaptation Strategy for Low-Bit Rate Adaptive Transform Coding", Proc.Int.Conf. on Digital Signal Processing, Florence, Italy, pp.638-642, 1978.

49. J.M.Tribolet and R.E.Crochiere, "Frequency Domain Coding of Speech", IEEE Trans.Acoust., Speech, Signal Processing, Vol.ASSP-27, No.5, pp.512-530, 1979.

50. J.M.Tribolet and R.E.Crochiere, "A Modified Adaptive Transform Coding Scheme with Post-Processing Enhancement", Proc.IEEE Int.Conf., ICASSP-80, pp.336-339, 1980.

51. H.B.Voelcker, "Towards a Unified Theory of Modulation, Part I: Phase-Envelope Relationships", Proc.IEEE, Vol.54, pp.340-353, 1966.

52. ————————, "Towards a Unified Theory of Modulation, Part II: Zero Manipulation", Proc.IEEE, Vol.54, pp.737-755, 1966.

53. P.A.Wintz, "Transform Picture Coding", IEEE Proc., Vol.60, pp.809-820, 1972.

54. J.Wiren and H.L.Stubbs, "Electronic Binary Selection System for Phoneme Classification", J.Acoust.Soc.Amer., Vol.28, p.1082, 1956.

55. 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.

56. Yair Shoham, "Hierarchical Vector Quantization with Application to Speech Waveform Coding", Ph.D.Thesis, Dept.Elec.& Comput.Eng., Univ.of California, 1985.

57. C.K.Yuen, "Computing Robust Walsh-Fourier Transform by Error Product Minimization", IEEE Trans.Computers, Vol.C-24, pp.313-317, 1975.

58. R.Zelinski and P.Noll, "Adaptive Transform Coding of Speech Signals", IEEE Trans.Acoust., Speech, Signal Processing, Vol.ASSP-25, No.4, pp.299-309, 1977.

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

LIST OF PUBLICATIONS OF THE AUTHOR

1. "Use of Zerocrossing Information for the Extraction of Speech Parameters", JASI, Vol.15, pp.96-101, 1987.

2. "Speech Signal Reconstruction from Composite Zerocrossings and its Application in Sub-band Coding", JASI, Vol.16, pp.18-23, 1988.

3. "A Performance Improvement Study of Adaptive Transform Coding of Speech at Low Bit Rate", JASI, Vol.16, pp.256-259, 1988.

4. "Speech Sample Estimation from its Composite Zerocrossings", Proc.Symposium on Signals, Systems and Sonars, March 9-11, 1988, NPOL, Cochin.

5. "Parametric Representation of the Dynamical Instabilities and Deterministic Chaos in Speech Signals", Proc.Symposium on Signals, Systems and Sonars, March 9-11, 1988, NPOL, Cochin.

6. "Reconstruction of Speech by Sampling its Zerocrossings", Presented in the National Symposium on Computer Aided Engineering, Feb.27-28, 1987, S.V.U.C.E., Tirupati.

7. "DCT/DWHT Adaptive Transform Coding of Speech", Tata McGraw-Hill, Proc. NACONECS-89, Nov.2-4, 1989.