THE LOW BIT-RATE CODING OF SPEECH SIGNALS

by

Mun Chen

A thesis submitted for the degree of
Doctor of Philosophy in the Faculty
of Engineering, University of London.

Department of Electrical Engineering,
Imperial College of Science and Technology,
University of London.

June 1976

ABSTRACT

This thesis is concerned mainly with the development of digital coding techniques that make possible the transmission of speech of reasonable quality, at bit-rates of approximately 1 bit per Shannon sample, with a coder whose complexity is not prohibitively greater than that of simple waveform coding methods. In the thesis, various well known methods of digital speech coding are reviewed briefly. In the review, emphasis is placed on coding principles that are to be further explored in the development of new coding techniques. During the course of the research reported on in the thesis, extensive use was made of a mini-computer and its peripherals, and two valuable research tools were developed, namely a speech analysis and processing system; and a speech intelligibility testing system. Descriptions of these are included in the thesis.

The search for new coding techniques involved investigations in the areas of amplitude dithering, residual encoding, and phase dithering. In the investigations into amplitude dithering, two new methods of dithered quantization with preservation of zero-crossings were found to have an intelligibility improvement of approximately one bit per sample as compared with normal fixed-level PCM quantization. The investigation has shown that effective speech digitizers of extreme simplicity could be developed with these new quantization methods, though they would not be capable of the desired low (1 bit per Shannon sample) output bit-rates.

In the thesis, techniques are described that are possible as a means of obtaining speech encoding at output bit-rates close to 1 bit per Shannon sample. Specifically, it was found that such output bit-rates were possible by the use of a one-bit quantizer

coupled with a new lattice-form of predictor operating within the framework of a residual encoder. The usual direct-form of predictor, while simpler in construction, was found to be unacceptable at such bit-rates, but was found to be satisfactory at higher bit-rates.

The technique of phase dithering was introduced to overcome some difficulties faced with the use of a coarse quantizer in a residual encoder. Further improvement in the performance of a one-bit residual encoder was found to be possible with the phase dithering technique. This technique also makes possible the effective implementation of a simplified set of parameter adaptation algorithms for the residual encoder.

ACKNOWLEDGEMENTS

I wish to express my gratitude to Dr. L.F. Turner, whom I am priviliged to have had as supervisor. His keen enthusiasm in the subject has been a constant source of encouragement and has greatly influenced the path of this research. In addition to providing valuable technical guidance and criticisms, especially towards the manuscript of this thesis, he has also arranged for much needed financial assistance to my study. For these, I am most grateful.

I am also fortunate in having met Dr. D.J. Goodman of Bell Laboratories, USA, during his academic visit to Imperial College and having the benefit, through many discussions, of his broad knowledge in the field of speech coding. His advice and suggestions have been of great help towards this research and are gratefully acknowledged. I am also grateful to Mr. C.B. Wan of the Department of Computing and Control, Imperial College, who has given valuable advice on optimization theory and estimation theory.

Various forms of assistance extended to me from staff and student members of the Communication Section are too numerous to mention. In detail, however, I wish to thank, Mr. L. Stenning for stimulating discussions in matter of human sound perception, and Dr. M. Apperley and Mrs. J. Dunn for assistance with computing problems; and, in general, all other members of the Section for discussions and advice in one way or another. Additionally, thanks are due to Mrs. M. Mills for her forbearance in the typing and re-typing of this thesis.

The bursary provided by the Procurement Executive, Ministry of Defence, England, and the fee grant provided by the British

Council are gratefully acknowledged. I am also indebted to the University of Singapore for granting 3 years of unpaid leave plus extensions to enable me to carry out this research.

Finally, my deepest gratitude is due to my parents and my younger sister. Their unconditional love and confidence in me are sources of joy and their moral and financial support has made my study possible. To them, I dedicate this thesis.

CONTENTS

LIST OF FIGURES AND TABLES

CHAPTER I

Introduction - Digital Speech Communication

Digital techniques for speech transmission offer many advantages and these have been summarized by Jayant[(1)], who states:

"...digital representation offers ruggedness, efficient signal regeneration, easy encryption, the possibility of combining transmission and switching functions, and the advantage of a uniform format for different types of signals...."

It is felt that the need for a uniform format is likely to lead to the extensive use of digital speech representation, since, in the long term, speech-type information is likely to be transmitted over a general communication system in which digital signalling will almost certainly be employed throughout. In this system it is likely that speech signals will be handled, together with other signals, such as computer data, video signals, facsimile data, news dispatches, market information etc. The system will operate using common standardised switching, storage and transmission functions, and little or no regard will be paid to the origin of the signals other than in the matter of priority. Digital techniques, in which all signals are represented in pulse form, open the way to such a unified communication system.

In the first part of this introductory chapter, a brief survey of the better known methods of digital coding of speech signals is presented. This is followed, in Section I.2, by a more specific study of methods of low-bit-rate waveform coding and this, in turn, leads to a statement of the aims and motivation of the research carried out and presented in the thesis. A breakdown of the organization of the thesis is given in section I.3.

I.1. The Digital Coding of Speech - a Brief Survey

Broadly speaking, the digital coding of speech signals can be classified into two main categories: Waveform Coding and the Vocoder Methods. Waveform coders attempt to preserve the waveform of the original speech signal. If the waveform of the received and decoded signal is identical to the original signal from which the digitally coded version was derived, it could safely be assumed that the received signal would sound the same as the original. On the other hand, the vocoder methods are concerned only with that the received signal should sound like the original. The fact that the received signal may have a waveform quite different from the original does not matter. In the vocoder methods, features of speech important to the perception process (i.e. how it sounds) are extracted from the speech signal. These are coded digitally and transmitted to the receiver, where a signal that possesses the same features is synthesized. Provided that all perceptually significant speech features are extracted, and the synthesized signal is constructed accordingly, the speech at the receiver end, as it is perceived by the human ear, should resemble the original. It is generally accepted that the vocoder methods are more efficient than the waveform coders in the use of transmission bandwidth. This is because, in the vocoder system, speech is reduced to its bare essentials and anything contained in the original speech that does not affect human perception, is discarded as redundant. Although vocoders are generally more efficient as regards the use of bandwidth, they tend to be complex and a point in favour of waveform coders is that they are generally simpler in construction and thus less expensive. A more detailed survey of various well-known speech coders is given below.

A. Waveform Coders:-

1) Pulse Code Modulation (PCM)

Basically PCM is a straightforward analogue-to-digital (A/D) conversion process. The analogue speech signal is sampled and the samples are quantized into a range of quantization levels with each level represented by an individual digital code-word. At the receiver, a voltage corresponding to the quantization level represented by the received digital code is regenerated at the appropriate sampling instant. A sequence of such regenerated voltages thus forms the received signal. PCM is significant in that, historically, it was the first method used for converting analogue speech signals into digital form and that it is still the only widely used method of digital speech transmission. Strictly speaking, the first step in any digital processing scheme for the bandwidth reduction, the analysis, the recognition or the storage etc., of speech is a PCM coding.

It is easy to see that the greater the number of quantization levels, the higher is the fidelity of the reproduced signal. However, the greater the number of levels, the higher also is the number of bits required to code a section of signal. If PCM is to be used for purposes of transmission, then often steps have to be taken to economize on the transmission bandwidth required. The choice of quantization step-size and clipping level[2,3], companding methods (non-unform quantization)[4,5,6] and step-size adaptation methods [7,8,9,10] etc. are all methods that have been developed as a means of reducing the bandwidth requirement of PCM coding.

The technique of amplitude dithering[11,12,13,14] is also a method of achieving reduced transmission bandwidth with minimal sacrifice in performance. In this case, however, the word 'performance' is used to mean subjective intelligibility and quality, and not the waveform difference between the original and received signal. This topic will be dealt with in depth in Chapter 3. Critical surveys of various PCM methods can be found in Ref. (1,3).

2) Differential Coding (with fixed predictor)

The method of differential coding represents an important step towards the bandwidth reduction of digitized speech. Instead of directly quantizing the input signal, the difference between the input and a predicted signal is quantized. At the decoder, a received signal is reconstructed from this quantized difference. In this category, one finds the well-known methods of delta modulation, differential PCM and many of their variants. Delta modulation, is, broadly speaking, only a special class of differential PCM in which a one-bit quantizer is used. Despite its basic similarity to differential PCM there are, however, many detailed differences. For example differential PCM normally works at the Shannon sampling rate, whereas oversampling at many times the Shannon rate is often employed for delta modulation. Also, the methods of step-size adaptation are different for delta modulation and differential PCM.

In differential coding methods the predicted signal is usually taken to be the previous signal sample and this is used for comparison with the input. In practice the predicted signal sample can be produced from the quantized difference signal using analogue methods of integration, or can be produced digitally using a unit delay in a recursive network as shown in Fig. I.1. A leaky integrator (which means, in the digital case, the coefficient, a, in Fig. I.1. is less than 1) is preferred to a perfect integrator. The leaky integrator is originally preferred since it avoids saturation effects and it prevents the catastrophic consequence that would result from the accumulating integration of transmission errors. As the understanding of differential coding increases, it is gradually being realised that a leaky integrator actually gives better prediction of the input speech signal and therefore is more suitable for the purpose of redundancy reduction

It is sometimes mistakenly argued that it is because the difference signal is less redundant, or the total amplitude variation of the difference signal is much less than that of the input signal fewer bits are required to code it to a given degree of accuracy than are required to code the original input. The fact is that the number of bits required for a quantizer to code a certain signal to a certain degree of accuracy defined by the SNR of the output of the quantizer is dependent only on the probability distribution, PDF, of the signal and is independent of the magnitude of signal fluctuations or its auto-correlation function. Removal of redundancy from the input signal does not necessarily render the PDF of the difference signal more suitable for quantization. The differential coder can only reduce the number of bits necessary to code a speech signal to a specific accuracy if the quantizer is placed within the feedback loop involving the predictor. This can be illustrated clearly by considering a typical differential encoder and decoder as shown in Fig. I.2., the difference between the input signal, S, and the predicted signal, P, is E, the prediction error, and the error, N, introduced by the quantizer is the quantization error. The SNR of the received signal, $\hat{S}$, is given by

$$\mathrm{SNR} = \frac{\overline{S^2}}{\overline{(\hat{S} - S)^2}}$$

where $\overline{X}$ means the time average of X.

But, $\hat{S} = \hat{E} + P = E + N + P$

and $S = E + P$

Therefore,

$$\mathrm{SNR} = \frac{\overline{S^2}}{\overline{N^2}}$$

$$= \frac{\overline{S^2}}{\overline{E^2}} \cdot \frac{\overline{E^2}}{\overline{N^2}} \qquad \text{---- (I.1)}$$

The ratio $\overline{E^2}/\overline{N^2}$ is the SNR of the prediction error, E, after quantization. As discussed above, this would not necessarily be improved by the reduction of redundancy in E. The overall SNR, of the received signal, $\hat{S}$, however, is improved by the ratio $\overline{S^2}/\overline{E^2}$ if the predicted signal, P, reduces the power of E by removal of redundancies in S. A simple mental exercise should suffice to show that this improvement in SNR cannot be realised if the quantizer is placed outside the feedback loop.

Many variations of the basic differential coding scheme have been investigated by numerous researchers. A great deal of effort has been devoted in considering numerous variations of the predictor network.

Among the better known variations are the Double Integration Delta Modulator[15,16] and the differential PCM employing three sections of delay[17]. These systems are aimed at producing a better prediction of the input signal by using a predictor (with fixed coefficients) that matches the long term statistics of speech waves. However, on account of the fact that speech signals are inherently quasi-stationary, the improvements over single integrator differential coders are fairly marginal for most practical purposes. More significant improvements, however, result from the use of various companding methods. In the case of delta modulation, the most notable schemes of improvement are Adaptive Delta Modulation[18] . Digitally Controlled Delta Modulation[19], High Information Modulation[20] etc. In most of these companding methods adaption is carried out using information obtained from an examination of the output bit stream. For example, a series of pulses of the same polarity would mean that the delta modulator was overloaded and that an increase in step-size is desirable. On the other hand, the reverse would be true when a string of pulses of alternating polarity are encountered. The continuous Delta Modulator[21] has a unique and rather ingenious companding strategy. The envelope information of the input speech is extracted and added as a slow-varying signal to the speech signal to be transmitted and together they are used as input to the delta modulator. The decoded output signal is lowpass filtered to reproduce the envelope information and the step-size of the quantizer is adjusted accordingly. The same is done synchronously at the receiver to obtain the step-size information. This method of multiplexing independently determined step-size information into the output bit stream without actually using more bits enables the delta modulator to be used at bit rates that are so low (below 1.5 bits/Shannon sample) that other companding methods break down. The reasons for the inability of the other dompanding methods to perform at extremely low bit rates will become clear in Section 4, Chapter IV.

Nearly all the companding methods used with differential PCM (multi-bit quantizer) are based on the principle that if the outer levels in the range of quantizer levels are used more often then the quantizer must be overloading and vice versa. Studies in this area are mainly concerned with optimal companding strategies, that is, with by how much the step-size should be increased or decreased if a certain quantization level is to be obtained at the output. Some examples of these can be found in Refs.(9,22,23,24).

An excellent comprehensive survey of delta modulation methods is to be found in Steele[25] and other surveys of delta modulation and differential PCM, including comparative studies, can be found in Refs. (1,26,27,28).

3) Adaptive Predictive Coding:-

A significant step in connection with the improvement in performance of differential coders was the conception[29] of the method of adaptive predictive coding. This method takes into account the quasi-stationary nature of speech signals and updates the parameters of the predictor to suit the changing signal statistics. One of the main difficulties associated with adaptive predictive coding is concerned with the problem of choosing a suitable predictor network sufficiently versatile to cope with different speech utterances whilst at the same time preventing the network from becoming inhibitively complex. Another difficulty is in determining the optimal predictor coefficient values given the short term signal statistics. Prior to the development of integrated circuits, the prospect of having to carry out more than ten multiplications or having to use more than a hundred logic gates was intimidating. Although, to some extent, implementation problems still exist, the advent of estimation theory, recent developments in digital signal processing techniques and improvements in modern computer technology have made it possible for these difficulties to be overcome in the main.

In their pioneering work, Atal and Schroeder(29) held that the process of speech production could be modelled approximately by a linear all-pole network excited by a minimal energy source (i.e. white noise or train of impulses). This minimal energy source signal could thus be extracted from speech waves by passing it through the inverse of the all-pole network (i.e. an all-zero network). A predictor network was derived such that this minimal energy signal became the prediction error, E, in the differential coder of Fig. I.2. Hence, the SNR improvement (i.e. $\overline{S^2}/\overline{E^2}$), due to the differential arrangement was maximized due to the minimization of the energy of E. Another source of redundancy, the periodicity of voiced speech, was also exploited by Atal and Schroeder for reduction of the energy of the prediction error E. Thus, their predictor consisted of two sections as shown in Fig. I.3., where

$$P_1 = \alpha \; Z^{-m}$$ was for the extraction of periodic redundancies

and $$P_2 = \sum_{1}^{n} a_i Z^{-i}$$ was for the purpose of predicting the short term waveform fluctuations.

The value m is the pitch period of the voice excitation, α is the normalized correlation coefficient between speech samples of m sampling periods apart and n is the order of the predictor. The optimal values of the coefficients a_i's of the short term predictor were calculated by Atal and Schroeder from the auto-correlation function of a stored section of speech. A matrix inversion method was used to calculate the optimal coefficient values for the whole stored section. These coefficient values are updated for every consecutive section of the input speech and are transmitted to the receiver together with the pitch information, the coefficient α, the quantizer step-size and the quantized prediction error. This method achieves a remarkable SNR gain of approximately 20 dB as compared with PCM. A real-time implementation of an essentially similar principle to this is reported in Goldberg and Shaffer[30] and simplified hardware realization omitting the extraction of the periodic redundancies is described in Dunn[31].

It is important to note that the coefficients of the short-term predictor do not have to be determined by matrix inversion. Provided the statistics of the speech signal change sufficiently slowly, an iterative approach can be equally effective and also often simpler if the order of the predictor is high. The iterative method can either be a steepest descent (or optimal gradient) optimization of the predictor coefficients so that the prediction error can be reduced to a minimum; or a Kalman filtering method using the same criterion. The iterative method has been studied by Scagliola[32] who has also made allowance for zeros as well as poles in his speech signal specification. A non-iterative solution to the values of the predictor coefficients, if the speech signal is assumed to possess zeros as well as poles, is as yet unknown[33]. Gibson et al[34], using an all-pole model, examined the application of Kalman filtering for the estimation of the predictor coefficients and found that it gave slightly better performance than the optimal gradient method. An innovation by Gibson et al. is their method of determining the predictor coefficients from the quantized information alone. As this information is also available to the decoder, the same operation can be performed at the decoder to extract these coefficient values. This synchronous estimation obviates the need to send the predictor coefficients. As in adaptive differential PCM, the adaptation of the (multi-bit) quantizer step-size is also done synchronously from the output bit stream at both the encoder and decoder and the transmission of the step-size information is thereby avoided. This approach is termed "residual encoding" by Gibson et al.. Jayant(1) in his survey, quoted a similar scheme by Stroh(35), details of which, unfortunately, are unavailable to the author. Jayant also quoted from Ref. (35) the view that the method of residual encoding will not be effective if the quantizer is too coarse (i.e. less than 2 bits). This is a view by which the author does not subscript to. Gibson(34) used a three-level quantizer ($\doteq$ 1.58 bits/sample) with success and as discussed in detail in Chapters 4 and 5 new schemes using a one-bit quantizer have been developed by the author and these schemes have been used to demonstrate the feasibility of residual encoding with coarse quantization. Another interesting direction of development, still on the theme of residual encoding, is described in Cohn and Melsa(36). They observed that the quantizer output of a multi-bit residual encoder is highly redundant. A variable-input-length, fixed-output-length code is proposed for further compression of this quantizer output. A compression from 2.3 bits/sample to 1.5 bits/sample has been achieved.

A general study of the earlier adaptive predictive methods with quantitative performance figures can be found in Noll[27] and this is useful as reference data for researchers in this field.

B. Vocoder Methods:-

The vocoder method of speech transmission is a vast subject, and in a few pages it is possible only to review very briefly the important aspects of the field. Thus, the discussion which follows should not be considered as a comprehensive review and for more general information, the reader is recommended to consult works such as those of Flanagan[37], Holmes[38] and the collection of papers by Flanagan and Rabiner[39]. In the survey, emphasis is placed on those vocoder methods which involve the use of linear predictive coding (LPC). This is done since a clear physical insight into LPC is important to the development of the new coding methods that are described in Chapters 4 and 5.

Basically, a vocoder transmission process can be divided into two parts namely analysis and synthesis. The analysis is carried out at the transmitting end where the speech signal is reduced to some essential features. The receiver attempts to synthesize, from information about these features, a signal that sounds like the original speech. Research effort has been directed at complete vocoder schemes, as well as the advancement of the specific techniques of analysis and synthesis. Though a diversity of methods have been studied in the past, the principal features of speech that are universally recognized as important to its perception by the human being are the shape of power spectrum of the speech and the mode of excitation giving rise to the speech. In articulatory terms, the mode of excitation refers to whether the speech sound is generated by vibration of the vocal chord or by a rush of air through a constricted space. The vibration of the vocal chord provides a quasi-periodic excitation the spectrum of which is shaped by the vocal tract, formed by the mouth, throat and nasal cavities. Sounds produced in this way are termed "voiced" sounds, and those noise-like sounds produced by a rush of air are termed "unvoiced" sounds. The various subtleties of the pitch of the voiced excitation, and the waveform of the glotal pulse train etc. are important in helping human beings to identify a particular speaker or speaker characteristics. For example, a female speaker generally is higher pitched than her male counterpart, and some coarse speakers appear consistently to lose one or two pulses in the supposedly periodic pulse train of vocal chord vibration. These excitations, together with the shape of the vocal tract through which they have to pass before radiation out of the mouth, determine the spectrum shape of the sound that is heard. By varying the geometry of the vocal tract, a speaker exercises control over the spectrum shape of his speech and in this way achieves a further degree of distinction between various utterances. For example, the vowels /e/ as in"hate" and /æ/ as in "hat" are both voiced sounds produced with the tongue humped near the front of the mouth cavity but the hump is raised higher in the case of /e/ than in the case of /æ/.

Although the process of speech production is relatively well understood (see the works of Fant[40] and Flanagan[37]) relatively little is known about how different utterances are actually distinguished in the auditory system. Despite the remarkable discovery by von Bekesy[41] that the cochlea in the inner ear is capable of performing a frequency analysis, many questions remained unanswered. Von Bekesy's theory cannot, for example, explain how voiced sounds are differentiated from unvoiced sounds, since the frequency analysis performed by the cochlea is not sufficiently sensitive to distinguish between the periodically pitched power spectrum of a periodic time waveform and the continuous power spectrum of a non-periodic noisy signal. That the pitch information is extracted from the heavily lowpassed waveform at the narrow end (the helicotrema end) of the cochlea does not seem to be a likely explanation, since it is well known that a musical "fifth" (e.g. 500Hz+750Hz) retains perfect musical harmony despite the fact that the period of repetition is neither the period of the lowest frequency component nor that of the second lowest frequency component. The period of repetition is in fact twice that of the lowest frequency component. The phenomenon of biaural hearing (the ability to locate accurately the positions of a sound source, even if it is behind one's back) and Cocktail Party Effect (the ability to listen to a particular person in an extremely noisy envinronment) add further complications to the confused picture. Attempts to explain these and many other problems have been made on the basis of hypotheses relating to the nature of higher levels of brain processing. One of the most convincing and readable account of this subject is by Unwin[42]. Despite the considerable effort that has been devoted to the question of perception, identification etc. an interesting unsolved question of evolution exists. Why, in the process of natural selection, has the brain been developed to perform the differentiation of periodic and nonperiodic sounds? This particular ability, which enables human beings to appreciate harmony and dislike traffic noises, does not seem relevant to survival.

This is probably one of the main reasons why the resonant tube theory is still propounded despite the experimental evidence offered by von Bekesy to the contrary. In the resonant tube theory, it is argued that the acoustic wave entering through the outer ear sets up a standing wave in the cochlea tube and it is suggested that the cochlea is able to adjust its length until a resonance is achieved with the incoming signal. A sense of satisfaction is obtained if this resonance is achieved. Alternatively, if the cochlea adjustment repeatedly fails to reach this state of resonance, a sense of annoyance results. Thus, it is considered that the ability to distinguish between periodic and nonperiodic sounds is only an accidental facility that arises as a by-product of the mechanism of cochlea detection. The main weaknesses of the resonant tube theory are that it lacks experimental evidence, and to explain how the power spectral envelope of the acoustic input is sensed it has to resort to the hypothesis that increditly complex higher-level brain processings is involved. An example of the latest developments in this direction is the Walsh transform hypothesis by Stenning[43].

At present, the only reliable factors that the designers of vocoders can rely on are (i) if the power spectral envelope and (ii) the voicing information are preserved; then the resynthesized speech will probably sound satisfactory. This, however, is of limited value, as slight variations in the power spectrum shape (e.g. a formant shift) or variations in detailed shape of the glottal waveform[44] do affect the perceptual impression of a speech sound. If an attempt were made to transmit all these details then this would prove to be prohibitively wasteful as regards transmission channel bandwidth, and hence a successful vocoder system has to extract the essentials and omit the remainder if the output bit rate is to be reduced to an acceptable value. In order to reduce the bit rate and maintain a subjectively acceptable speech quality the designer has to use his ingenuity and experience, and his interpretation of published psycho-acoustic data. Some well known vocoder methods are discussed below.

1) Channel Vocoder:-

The first vocoding system was that invented by Dudley [45]. This system is a channel vocoder, which uses a direct approach to the problem of preserving the speech power spectrum and the mode of excitation. The power spectrum of input speech signals is sampled at discrete frequency points. As the statistics of speech change slowly, these discrete samples of its power spectrum are coded and sent at a much slower rate than would be required in order to transmit the time waveform itself. As regards the speech excitation, a decision is made on whether the speech signal is periodic or nonperiodic and, if it is periodic, the pitch of the periodic waveform is measured. All other information about the excitation waveform is discarded, and this thus results in a considerable saving of bandwidth. The justification for this approach is to be found in the psycho-acoustic finding that the phase of a signal has little effect on sound perception (if one neglects the phenomenon of biaural hearing). Thus, all noisy (non-periodic) signals of identical power spectra sound the same, and the same is true for periodic signals of identical power spectra and pitch. In the early channel vocoders, a bank of band-pass filters were used for discrete frequency sampling of the power spectrum [45,46]. A schemmatic diagram of a channel vocoder, including both the analyzer and a typical synthesizer, is shown in Fig. I.4.

Recent developments have been directed at the measurement of the spectral information by digital methods such as by the use of the Fast Fourier Transform (FFT)[47] or by using a bank of digital filters[48]. The principal limitations of channel vocoders are two-folded. Firstly, a discrete sampling of the power spectrum is not a particularly efficient way of preserving the perceptually relevant spectral details (this should become clear from the discussion of Formant vocoders), with the result that the channel vocoder suffers from an unnaturalness in speech reproduction if too few channels are used. On the other hand, the use of a large number of channels requires a great deal of bandwidth for the transmission of this spectral information. Secondly, the voicing decision and the extraction of pitch from the time waveform is a task that is difficult to perform accurately. Many simple methods are simply not accurate enough and this also contributes towards the unnaturalness of reproduced speech.

2) Voice excited vocoder:-

The voice excited vocoder avoids the necessity of a voicing decision and pitch extraction by sacrificing some transmission bandwidth. A schematic diagram of the voice excited vocoder is shown in Fig. I.5. A baseband band-pass filter is used to extract a low-frequency narrowband section from original speech and this can be transmitted by a waveform coding method. At the receiver this baseband signal is processed by a non-linear distortion circuit which flattens and broadens the power spectrum of the baseband signal without destroying its periodicity (or non-periodicity). This flattened and broadened signal is used as the excitation for the synthesizer.

A typical upper cut-off frequency of the baseband filter is about 950 Hz and the lower cut-off frequency is usually not less than 250 Hz[49]. As shown in Fig. I.5, the power spectrum information is extracted and transmitted as in a channel vocoder. However, the application of the voice excited method of preserving the excitation information is not restricted to use in a channel vocoder. Successful voice excited formant vocoders have also been developed[50].

3) Cepstrum vocoder:-

The method of cepstrum vocoding[51,52] provides a very accurate technique for estimating the pitch period and other excitation information. This method involves the use of the concept of homomorphic filtering which has been applied both to picture and speech coding[53,54,55]. For speech coding the purpose of homomorphic filtering is to separate the excitation from the effect of the vocal tract spectrum shaping. Very briefly, the speech waveform, s(t), as radiated from the lips can be considered as the convolution of the excitation time function e(t) and the vocal tract impulse response v(t).

i.e. $$s(t) = e(t) \otimes v(t)$$

Or, in the frequency domain,

$$S(\omega) = E(\omega) \cdot V(\omega)$$

where ω is the angular frequency.

Taking logarithm of both sides, the expression

$$\ln S(\omega) = \ln E(\omega) + \ln V(\omega)$$

is obtained, and on taking the inverse Fourier transform, we have

$$F^{-1}(\ln S(\omega)) = F^{-1}(\ln E(\omega)) + F^{-1}(\ln V(\omega))$$

Thus, the inverse Fourier transform of the logarithm of the frequency spectrum of the speech signal is the sum of two distinct parts, of which $F^{-1}(\ln E(\omega))$ is contributed by the excitation and $F^{-1}(\ln V(\omega))$

is contributed by the impulse response of the vocal tract. Now, the inverse Fourier transform of a function in the frequency domain gives a function in the time domain. Hence both $F^{-1}(\ln E(\omega))$ and $F^{-1}(\ln V(\omega))$ preserve some of the properties of the time waveforms $e(t)$ and $v(t)$ respectively, in that $F^{-1}(\ln E(\omega))$ is a wave of long duration whereas the envelope of $F^{-1}(\ln V(\omega))$ decreases with time (because the vocal tract cannot be unstable). To distinguish $F^{-1}(\ln E(\omega))$ and $F^{-1}(\ln V(\omega))$ from the original time function $e(t)$ and $v(t)$, they are termed the excitation function and the impulse response function in the quefrency (not time) domain respectively. The representation of a signal in the quefrency domain is called the cepstrum of the signal. A typical cepstrum of a voiced speech signal is shown in Fig. I.6. The part contributed by the impulse response of the vocal tract (i.e. $F^{-1}(\ln V(\omega))$) is confined to an area around the origin and the part due to the excitation source (i.e. $F^{-1}(\ln E(\omega))$) appears as a periodic pulse train from which the pitch period can be easily determined.

For the purpose of vocoding, the frequency spectrum data generated in the first Fourier Transform (usually by an FFT algorithm) of the speech signal (i.e. the function $S(\omega)$ from $s(t)$) could be transmitted over the channel. Alternatively, another Fourier transform could be carried out on the impulse response part (i.e. $F^{-1}(\ln V(\omega)$) of the cepstrum (resulting in $\ln V(\omega)$) and the anti-logarithm of the result would yield the frequency domain description of the vocal tract transfer function ($V(\omega)$). Transmission of the vocal tract frequency response is often found to give better results than the frequency spectrum data obtained by a direct FFT in the speech waveform. This is because of the finite length windowing necessary for the FFT procedure.

If the excitation function is included in the input data for FFT, the truncated excitation function is not of flat power spectrum (unless the width of the window is very large or exactly an integer multiple of the pitch period[56]), and this results in a distortion of the signal spectrum. Another method analogues to transmitting the vocal tract frequency response (actually used in Ref. (55)) is to convert the frequency domain description of the vocal tract into its impulse response by a further FFT. This facilitates the final synthesis by digital methods since the receiver output is simply the convolution of this impulse response and the excitation function.

4) Formant Vocoders:-

The previously mentioned vocoder methods all suffer from the fact that in order to reproduce speech of acceptable naturalness, the spectral envelope of the original speech has to be sampled very finely with the result that a large bandwidth would be required to transmit this amount of spectral data. As a matter of fact, it is not necessary to transmit all the data. Psycho-acoustic experiments have shown that the positions of the sharp peaks in the power spectrum of a speech signal are of particular importance to its perception. These peaks are called formants and correspond to the poles in the transfer function of the vocal tract.

With the formant vocoder, instead of attempting to preserve every detail of the power spectrum of the original speech signal, the positions of the formant peaks, plus, in some cases, the intensities of the formants, are transmitted. Also, the situation can be further simplified since it is possible to a large extent to predict (and thus artificially recreated at the receiver) the intensity information from a knowledge of the formant frequencies[57]. Slight error in the intensities does not result in a drastic degradation in the quality of the reproduced speech[37]. For most practical purposes a minimum of three formants is required for satisfactory representation of the speech spectrum. The most direct method of identifying the formant frequencies is to use a large array of channel filters (approximately 64) and as in the earlier formant vocoders[58,59] pick the frequencies at which the filter output is the highest. In modern formant vocoders, the tendency is to use a digital computer for extraction of the formant data. Formant information can be obtained by a bank of digital filters or by the method of discrete Fourier transform, followed by a peak peaking procedure for automatic identification of the formants[60,61]. The method of ceptrum filtering can also be applied, though at extra cost, to alleviate the effect of spectral distortion caused by finite length windowing of the excitation source signal and to produce an accurate pitch information. Alternatively, the method of linear inverse filtering, to be discussed in the next section, can also be used for accurate estimation of the formant frequencies[62].

5) Linear Predictive Coding:-

Like the formant vocoders, the method of linear predictive coding also attempts to preserve the peaks of the speech spectral envelope. This is not done, however, by the straight forward approach of fine sampling of the power spectrum followed by a peak picking procedure, but by an indirect method:- Consider, as is usually the case for speech signals, a certain spectral envelope consisting of many peaks. On account of the high harmonic content of the profile of the peaks, it is necessary, according to Shannon's sampling theorem, to sample at a high rate in order to preserve details of the perceptually significant peaks. In other words, a large number of samples has to be taken. The need for a large number of samples can be overcome in the case in which one is interested only in the peaks. In such a case, an all-pole function approximation of the spectrum can be envisaged which matches the original spectrum in both the position and height of the peaks but allows for differences in the profile of the valleys. Such an all-pole function is of the form

$$S(Z) = \frac{1}{1 - \sum_{1}^{n} a_i Z^{-i}} \qquad \text{----} \quad (I.2)$$

where n is the order of the all-pole function. The inverse $1/S(Z)$ of this all-pole function is an all zero function:-

$$1/S(Z) = 1 - \sum_{1}^{n} a_i Z^{-i} \qquad \text{----} \quad (I.3),$$

and the inverse Fourier transform of this inversed transfer function (Eq.(I.3.)) is its impulse response which is a time series of finite length, the samples of which give the coefficient values a_i's in Eq. (I.3). These coefficients, in turn, specify completely the all-pole approximation of the original spectrum. Since, as it is well known (see Atal and Schroeder[29]), a pair of complex conjugate roots of the denominator of Eq. (I.2) specify a pole frequency and its height, an all pole function of order 2X is sufficient to describe X number of peaks in the spectrum. In other words, only 2X samples of the impulse response of the inverse spectrum need to be taken, which represents a considerable saving over direct sampling of the speech spectrum.

To determine the samples of the impulse response of the Eq. (I.3) the rather involved procedure implied in the above conceptual discussion need not be followed. In practice the method of linear predictive coding calculates directly the coefficients of the all-zero transfer function in Eq. (I.3) from the time waveform of the speech input. This procedure can be explained briefly as follows.

Supposing that the speech signal consists of only poles (that is, its spectrum can be completely specified by the all pole function of Eq. (I.2), then the passing of the speech signal through a linear filter whose transfer function is exactly the inverse of the speech spectrum (which means an all zero filter as Eq. (I.3)) would flatten exactly the power spectrum of the input signal. In other words, all correlations between signal samples are removed by the all zero filter and the resulting filter output is of minimal energy. Thus the coefficients a_i can be calculated so that an all zero filter using these coefficient values reduces the power of the input speech signal to a minimum. Hence, the problem of determining the samples of the impulse response of the inverse filter is essentially that of inverse Wiener filtering in parameter estimation theory. From a knowledge of the auto-covariances or auto-correlations of the input speech signal, values of the coefficients can be found by a matrix inversion. Atal and Hanauer[(63)] used a knowledge of the auto-covariances and Markel[(62)] the auto-correlation to determine the coefficient values. They are referred to respectively as non-stationary and stationary formulations[(64)]. Atal and Hanauer[(63)] had also pointed out that the effects of zeros in the speech spectrum can be accounted for by using an inverse filter model of an order larger than that necessary to describe the formant peaks, because a zero can be approximated by a large number of poles. The method of Itakura and Saito[(65)] also uses an all-zero inverse filter but it is significantly different from the above two in that it employes a special lattice filter structure which is reproduced in Fig. IV.3.a. With this structure, the redundancies (or correlations) of the speech waves are removed stage by stage. Each filter stage is a four-terminal device. The coefficient of each stage is calculated solely from the inputs to the stage and the coefficient is designed so that the outputs from the stage are reduced to minimal power. It has been shown(65) that if the coefficients of all the stages are optimized this way, the final output from the last stage is of minimal energy, and this thus functions as Wiener inverse filtering without the need for a matrix inversion. The coefficients so calculated are not the discrete samples of the impulse response relating to the inverse spectrum, but they serve the same purpose in that they specify the peaks of the original speech spectrum. The coefficients are called by Itakura and Saito, partial correlation coefficients (PARCORs). Given an inverse filter of the same order, a one-to-one relation exists(67) between the PARCOR coefficients and the direct-form coefficients of the inverse filter of Eq. (I.3). It has been shown (66) that the PARCOR coefficients actually correspond to the reflection coefficients of a vocal tract model made up of cascaded uniform tubes of the same length and different diameter.

The methods of linear predictive coding are intrinsically suitable for vocoder applications, as the minimal energy signal from the inverse filtering is essentially flat in spectrum. Thus, if the input speech were a voiced sound, the inverse filtered output would appear as a train of sharp impulses from which the pitch information could be determined easily and accurately. Likewise, if the speech input is unvoiced the output is noisy and devoid of sharp regular peaks, thus simplifying the task of deciding whether the sound is voiced or unvoiced. At the receiver, a simulated excitation source can be generated from the voicing, pitch and intensity information and the speech can be synthesized by passing the excitation source through the inverse of the all-zero inverse filter. In the case where coefficients are measured for the direct-form of the inverse filter, a resynthesizing filter of the form as in Eq. (I.2) can be used. Real-time and hardware implementations of linear predictive vocoders using the direct-form of inverse filter structure can be found in Refs. (71 and 72). If the lattice-form of inverse filter is used for speech analysis, the coefficients measured are PARCOR coefficients. To synthesize the speech signal from PARCOR coefficients, the synthesizing filter structure given in Itakura et al.[65] can be used. Alternatively, an equivalent set of direct form coefficients can be calculated from the PARCOR coefficients using the conversion relationship given by Markel and Grey[67] and speech can be synthesized using a filter having a direct-form structure.

The Linear Predictive Coding methods discussed so far assume an all-pole input speech signal. If zeros are to be included in the vocal tract model, the Wiener inverse filtering approach, with matrix inversion, would result in indeterminate solutions[33]. Scagliola[32] has proposed an iterative method (using gradient optimization), for determining the parameters of a vocal tract model incorporating zeros as well as poles. The iterative method has the additional advantage that it is much simpler in implementation than the method of matrix inversion. A possible drawback of the pole-zero model is that whereas the all pole model becomes more accurate as the order of the filter is increased, there are no systematic rules for the choice of the order of the denominator or the numerator in the model involving both poles and zeros. However, it has been shown experimentally[32] that, with reasonably ad hoc choices of order, the pole-zero model gives generally a better spectral representation than the all-pole model of the same complexity.

Another iterative method of determining the inverse filter coefficients has been proposed by Gibson et al.[68]. Using an all-pole vocal tract model, they have compared the optimal gradient (called stochastic approximation by Gibson et al.) method, as used by Scagliola[32], with Kalman filtering using sequential estimation algorithms and they have found that the Kalman filtering method produces a better quality resynthesized speech, because of its faster convergence rate. The Kalman method does however involve approximately the same number of calculations as the matrix inversion method though it may be simpler to implement in hardware form.

The direct transmission of the inverse filter coefficients is not exactly the most efficient way of utilizing the transmission bandwidth. If it is supposed that accurate recognition of 3 formants is sufficient for a certain speech transmission purpose, then in order to specify these 3 formants, a vocal tract model of 6 coefficients is theoretically adequate. However, to determine the position (and height) of these 3 formants, a 6th order inverse filter is certainly not adequate. A pole that has not been taken into account and, even worse, a zero in the input speech can cause significant shifts in the formant data so determined. To avoid this a model of an order much higher than 6 has to be adopted, and the use of this higher-order filter means more coefficients have to be transmitted and this results in an unnecessary waste of bandwidth. For further bandwidth compression, an additional formant vocoding operation can be applied. Accurate formant data can be generated with a vocal tract model of sufficiently large order and only those vital formants need be transmitted. Methods of formant analysis by linear predictive coding have been discussed in Refs. (62, 63, 69, 70). The simplest is that of Markel[62] in which an FFT is performed on the impulse response of the inverse filter. The inverse of the resulting spectrum is the vocal tract transfer function and the formants can then be determined by a peak picking procedure.

On the other hand, by an approach similar to that used in voice excited vocoders, the necessary order of the vocal tract model can be much reduced at the expense of higher transmission bandwidth. In addition to his hardware implementation of an adaptive predictive waveform coder, Dunn[31] had also experimented with a linear predictive coding method in which, by quantizing and transmitting the residual signal from the inverse filter and synthesizing from this quantized residue, a second order inverse filter was found to be adequate. This can be explained intuitively in terms of the fact that a signal of flat power spectrum can be subjected to any degree of amplitude quantization without changing its power spectrum shape[73] or its periodicity (or non-periodicity). This means that its power spectrum remains flat; and if the input signal is periodic, the pitch period will be preserved exactly. On the other hand, if the input signal is noise-like the quantized version will remain noisy. The flatter the input signal power spectrum, the less is the extent of the distortion of the power spectrum shape introduced by the quantizer and it appears that a second order inverse filter flattens the spectrum of the input speech to such an extent that the resulting spectral distortion introduced by the quantizer is acceptable to human ears.

I.2. Motivation and Scope of Research - Low-Bit-Rate Waveform Coding

The main motivation for the research reported in this thesis was the desire to develop a relatively uncomplicated low-bit-rate system for digital coding of speech signal. The aim was not to strive for extremely low-bit-rate transmission, or to achieve exceptionally high quality of speech reproduction, but to strike a possible compromise between these and the cost of implementation of the system. The complexity of the system should be such that its application in a reasonably sized office can be envisaged and its cost should be comparable with that of existing data modems . A transmission bit rate of approximately 7.2 Kbits/sec is an attractive target since it would then be possible to send the output digital bit stream using 3 commerically available 2.4 Kbit/sec modems in parallel. The quality requirement for the reproduced speech should not be too stringent, but the reproduced speech is required to be clearly and effortlessly intelligible.

With the bandwidth requirement in mind it appeared at the outset that vocoding methods were the obvious means for achieving the objective laid down above. However, they were rejected on the grounds of the high complexity involved. The majority of vocoding methods require the separation of the excitation function and the vocal tract transfer function. This action enables speech information to be coded efficiently, but from the brief survey in the last section it should be clear that the essential features associated with these functions have to be extracted very carefully, and this requires involved processing methods if an acceptably natural speech reproduction is to be achieved. It appears that vocoder methods are only really successful if they are sufficiently sophisticated. This point can be illustrated by comparing the methods of linear predictive coding and adaptive predictive coding. The former method is a vocoder method and the latter a waveform coding method, and both are based on the principle of linear prediction. With adaptive predictive coding, Atal and Schroeder(29) found that an 8th order prediction filter gave impressive results ; and other implementations using 2nd to 4th order filters(30,31) have been shown to work satisfactorily. In linear predictive coding, 10th (or higher) order inverse filters are commonly used(73,71,72) in order to reproduce the formant data to an acceptable accuracy. It is conceivable that by allowing for higher bit rate than commonly used in vocoding methods, it might be possible to devise a scheme such that the complexity of the vocoder could be reduced with small sacrifice of its performance. The voice excited vocoders is an example of what might be achieved in this respect. However, it was felt, at the outset of the work that the possibilities in this direction were rather limited. It was thus decided to examine waveform coding and to attempt to reduce the necessary output bit rate associated with this method.

Before proceeding in this direction, it would seem to be sensible to recapitulate on some existing low-bit-rate waveform coding methods and examine their aptitudes and failings. The term'low-bit-rate' is taken to mean an average output rate of 2 bits per Shannon sample, or less. Various adaptive delta modulation and differential PCM methods fall into the top of this category. They give satisfactory performance at around 2 bits per Shannon sample but not at much lower bit rates. The output bit rate of adaptive differential PCM methods is limited by the very nature of the quantizer step-size adaptation method which requires a quantizer of at least three levels. A two-level quantizer, though, is used in the adaptive delta modulation methods. With these methods satisfactory results can only be obtained by over-sampling of the input speech signal (i.e. sampling at a frequency greater than Shannon sampling rate). Their step-size adaptation algorithms become ineffective when the sampling rate is lower than about 1.5 times the Shannon rate. The method of continuous delta modulation(21) overcomes this difficulty by means of an independently estimated quantizer step-size which is coded in the sub-audio band of the regenerated signal. By this ingenious method very coarse speech utterances can be reproduced with a transmission bit rate very close to 1 bit per Shannon sample. A system of higher quality, though requiring a greater bandwidth, is the adaptive ~~differential~~ PCM coder devised by Wilkinson(74). A 4-level (2-bits) quantizer is used. The polarity bit is sent each sample but only one amplitude bit is transmitted for every consecutive sample. This can be done since the amplitude information is highly redundant. With this method only $1\frac{1}{2}$ bits per sample are used and the quality of the reproduced speech approaches that of a 2 bits adaptive differential system. Similar conclusions were also reached by Turner(75) in an independent study of the statistics of the output bit stream of differential PCM coded speech signals.

The methods mentioned above are extremely simple to implement; but, due to the primitive nature of the prediction network the quality of speech reproduction at very low bit-rates (e.g. less than 1.5 bits/sample) leaves much to be desired. However, the technique may be useful in the military application on account of their ruggedness and simplicity. The only known low-bit-rate waveform coding systems that could give speech reproduction approaching a commercially acceptable quality are those employing the adaptive prediction principle, first studied by Atal and Schroeder(29). These systems, however, are of a complexity many order of magnitude higher than the simple schemes with non-adaptive predictors. The real-time implementation of the scheme of Atal and Schroeder requires the use of a special-purpose computer of speed about 10 times that of a common mini-computer (200 nsec memory cycle time against about 2,000 nsec) and even then it is only possible to implement a 4-coefficient adaptive filter which is half the number used by Atal and Schroeder. Even the much simplified hardware adaptive predictive coder of Dunn[31] using only a two-stage predictor network is disproportionately complicated in construction for a performance slightly better than the best of the fixed predictor methods*.

None of the coders so far mentioned in this section has the capability, or even the potential, of operating at bit rates below 1 bit per Shannon sample. It is generally true that waveform coders, which have to construct the signal sample by sample, need at least one fresh bit of information per sample for the reason that as the signal is sampled at the Shannon rate, every new sample has to contain an unexpected element. Nevertheless, a perfectly periodic signal can be sampled at less than the Shannon rate since a knowledge of one single period would enable the regeneration of a complete periodically repeated sequence+. This principle is applied in the waveform interruption/reiteration coder of Frei et al[76]. The basic coding device of Frei et al. is a simple adaptive delta modulator working at a sampling rate of about 20 Ksamples/sec. For a voiced periodic speech signal only one period out of every 2,3 or 4 is sent. At the receiver, the missing or "interrupted" periods are filled in by "reiteration" of the transmitted period. The same interruption/reiteration procedure is followed in unvoiced sections, but the length

---

* In terms of SNR performance, the Dunn's coder achieves about 10 dB for 9.6 Kbits/sec (about 1.5 bits/Nyquist sample) and Jayant[1] quoted about 8 dB for ADPCM at 10 Kbits/sec using similar signal bandwidth.

\+ This is not a violation of information theory as other than the first period, all subsequent periods do not carry any fresh information.

of transmitted information is fixed rather than pitch dependent. Evidently, this method requires an accurate pitch estimation and a large buffer store, not only to store the information of the one transmitted section for the purpose of regeneration, but also to produce, from the variable rate output of the coder a constant output data rate for synchronous transmission. An average output bit rate of 4.8 K to 10 K bits/sec, depending on the interruption/reiteration ratio, is claimed. Clever though this method may be, the reproduced speech is clearly not going to sound particularly natural. Firstly, the method for regenerating the unvoiced sections would need considerable refinement. In its existing form, it produces a periodic output from a non-periodic noisy input. Secondly, voiced speech signals are not perfectly periodic. Not only do the waveforms differ slightly from a pitch interval to another, but the adjacent pitch periods are also not exactly identical and waveform discontinuities could result when fitting together sections of speech whose lengths are not exactly of a pitch period.

This was the situation regarding low-bit-rate waveform coders at the beginning of the research reported in this thesis. None of the methods described could meet the specifications set out at the beginning of this section. Nonetheless, the performance of these methods were sufficiently promising to allow optimism that the targets were practically attainable. In the light of and with the achievements of existing coders in mind, it was possible, at that stage, to restate, specifically, a set of realistic goals, namely:-

1) Bit Rate Requirements:- The interruption/reiteration tecnique is unsatisfactory and a simple method of overcoming its complications does not seem immediately at hand. Thus it is decided

that effort should be concentrated on trying to achieve a bit rate of 1 bit per Shannon sample and not attempting to break this barrier.

2) Quality Requirement:- At a bit rate close to 1 bit per Shannon sample, the quality of the reproduced speech should be at least as good as that of the reproduced speech from the simplified adaptive predictive coder of Dunn, i.e. should have an SNR of about 10 dB.

3) Complexity:- A system of complexity in between the simplest existing adaptive predictive coder and the most complex of the fixed predictor methods should be aimed for. As a user terminal device, compactness of the coder would also be desirable. Hence, large buffer storage should be avoided as far as possible. This means that a system that is intrinsically of uniform output data rate would be preferred to one requiring buffer regulation.

I.3. Organisation of Thesis

In trying to achieve the goals set out in the previous section, the research work was conducted in three relatively distinct phases. Firstly, the quantizer, the essential element in all digital coders, was reappraised to see if there exists a better method of quantization than the conventional one using fixed decision threshold. An investigation of new methods of dithered quantization was conducted and the performance of these new methods were compared with that of a fixed threshold quantizer. Small performance gains were measured for some of the methods investigated. In the second phase, attention is shifted from the quantizer to the coder network. Using a conventional quantizer, the adaptive predictive coding principle was examined from a new approach. The results were found to be immediately encouraging in that the coders developed met all desired specifications, though only marginally on the complexity issue. In the third phase, effort was concentrated on development and refinement. Within the general framework of the new coders, methods were sought that would further enhance their performance, with special emphasis on reducing complexity. An obvious next step would seem to be to try a combination of the dithered quantization method and the new coding schemes. Disappointingly, however, the method of dithered quantization does not appear to be suitable for use with the new coding schemes. In addition, on closer examination, it appears, in any case, that the performance of the new coders is not likely to be improved by dithered quantization. However, a new dithering technique, that of phase dithering, is found to significantly improve the performance of a new coder. Moreover, with suitable modifications, which invoke some sacrifice of performance, a very much simplified coder is created. This simplified coder meets the bit rate and performance criteria with a complexity only about an order of magnitude greater than that of an adaptive delta modulator. A detailed account of these three phrases of research is given in Chapters III, IV and V.

A useful by-product of the period of research is the creation and development of a speech analysis and processing system based on the use of a minicomputer and its peripherals. It includes essential functions such as data handling and graphic output by the X-Y plotter. A description of this system is given in Chapter II. It is written with the purpose of serving as a reference guide for future users of the various routines and programs in the system. Readers may skip this chapter without losing continuity of the thesis.

Another aspect of the research, indirectly related to the development of the coders, is the problem of subjective evaluation of the coder performance. Chapter III also describes a subjective intellibility testing system employed during the research. The system, in its own right, contains some innovations since it is one of the earliest attempts at a fully automated computer administered subjective speech intelligibility measurement scheme*. From the design of the listening tests to the presentation of statistically analysed results, the system requires only the minimum of operator intervention. It was unfortunate that the system could not be applied after Phase 1 of the research because an essential peripheral device, the interactive graphic unit, was removed as a result of a re-organization of computing facilities.

In the final chapter (Chapter VI) an attempt is made at an objective assessment of the results of this period of work. Some suggestions are made relating to further research.

---

* As far as the author is aware, the only other such system ever described in published literature is that by Agarwa[77].

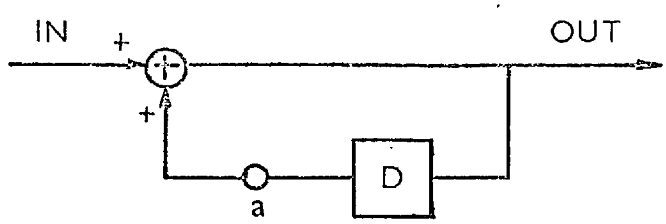


Fig. I.1. A digital integrator

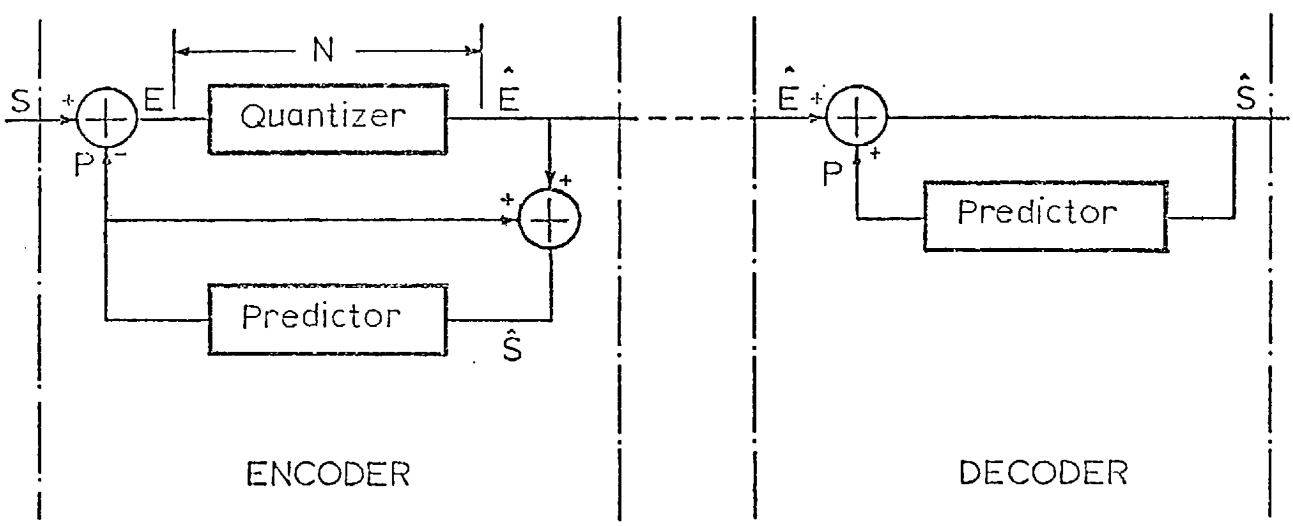


Fig. I.2. A typical differential encoder and decoder

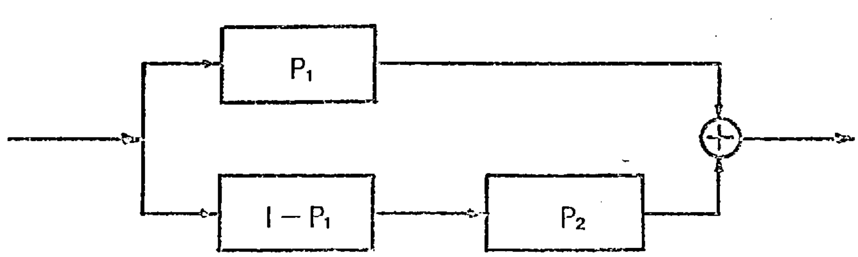


Fig. I.3. Two-stage predictor employed by Atal and Schroeder

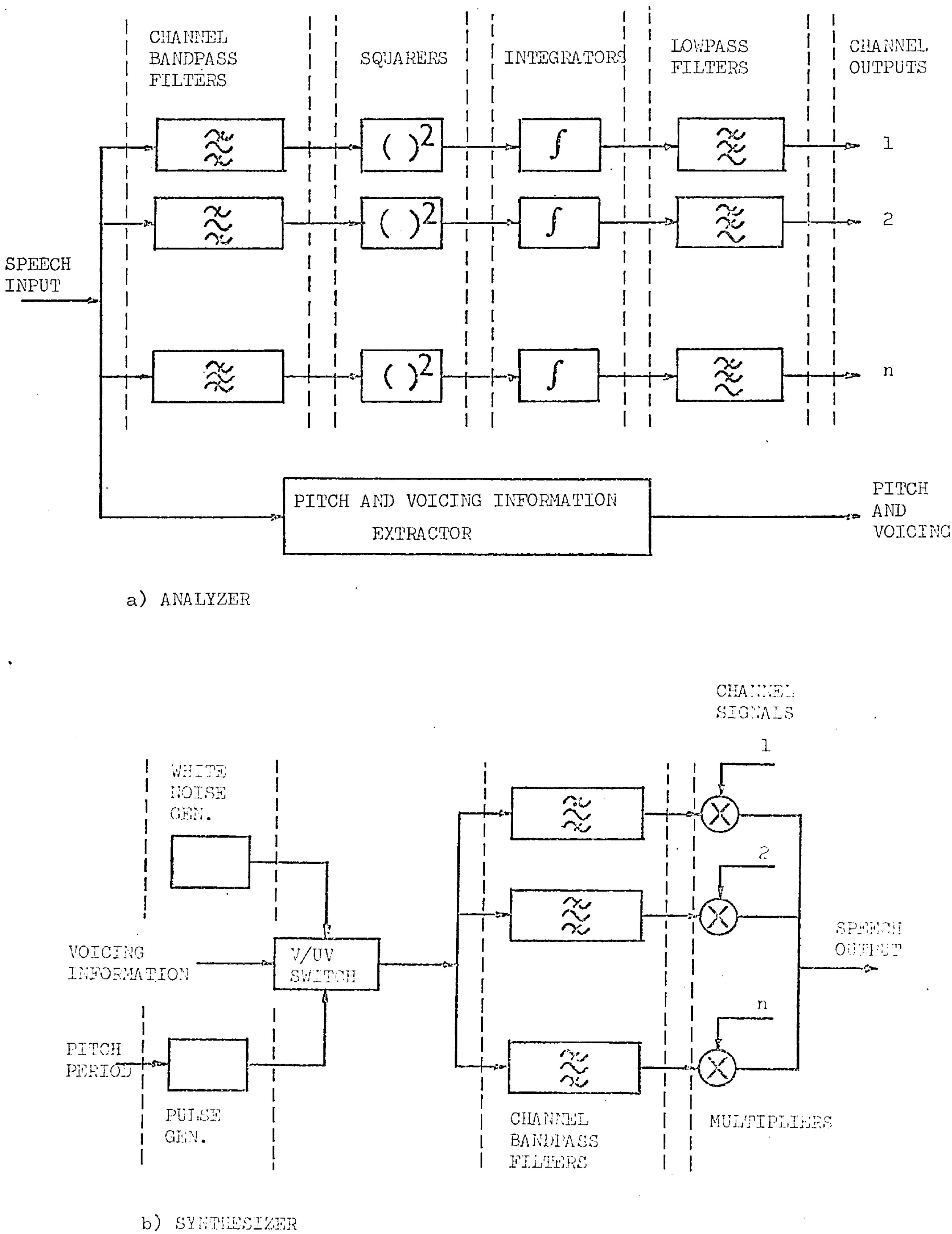


Fig. I.4. Block diagram of a channel vocoder

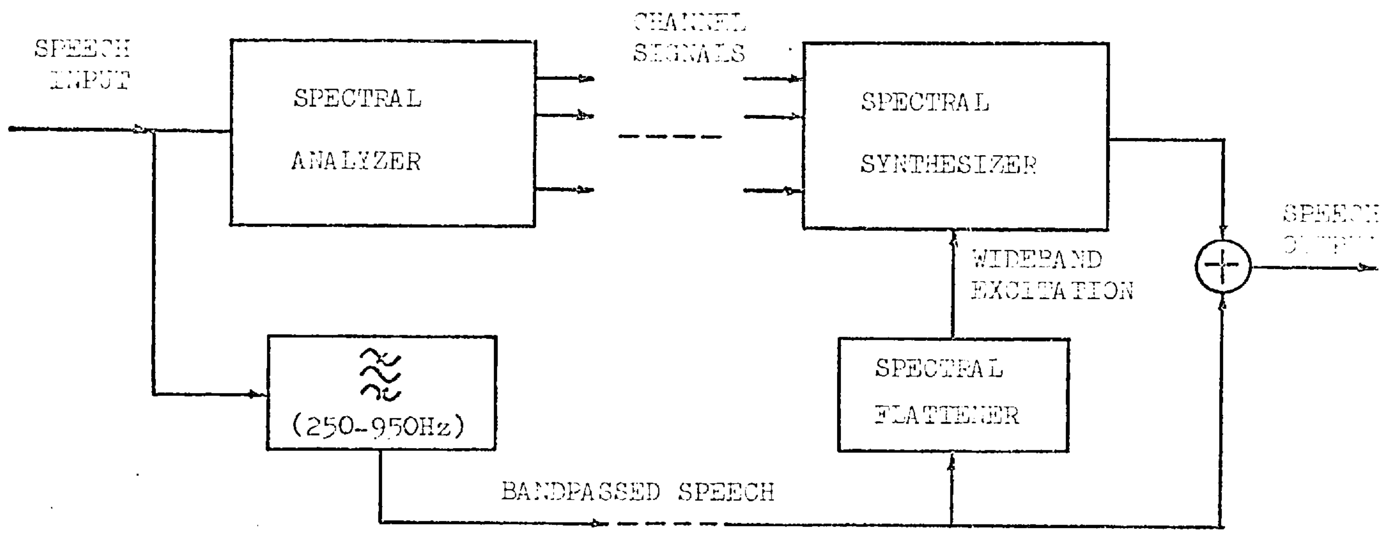
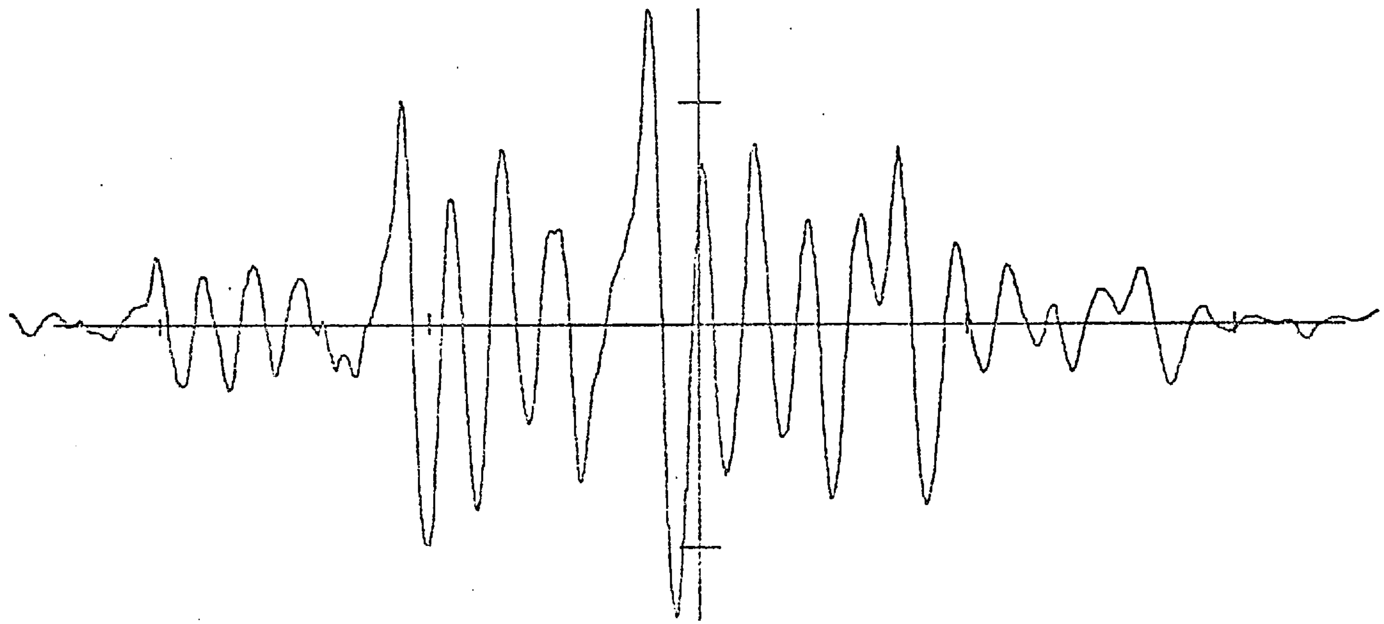
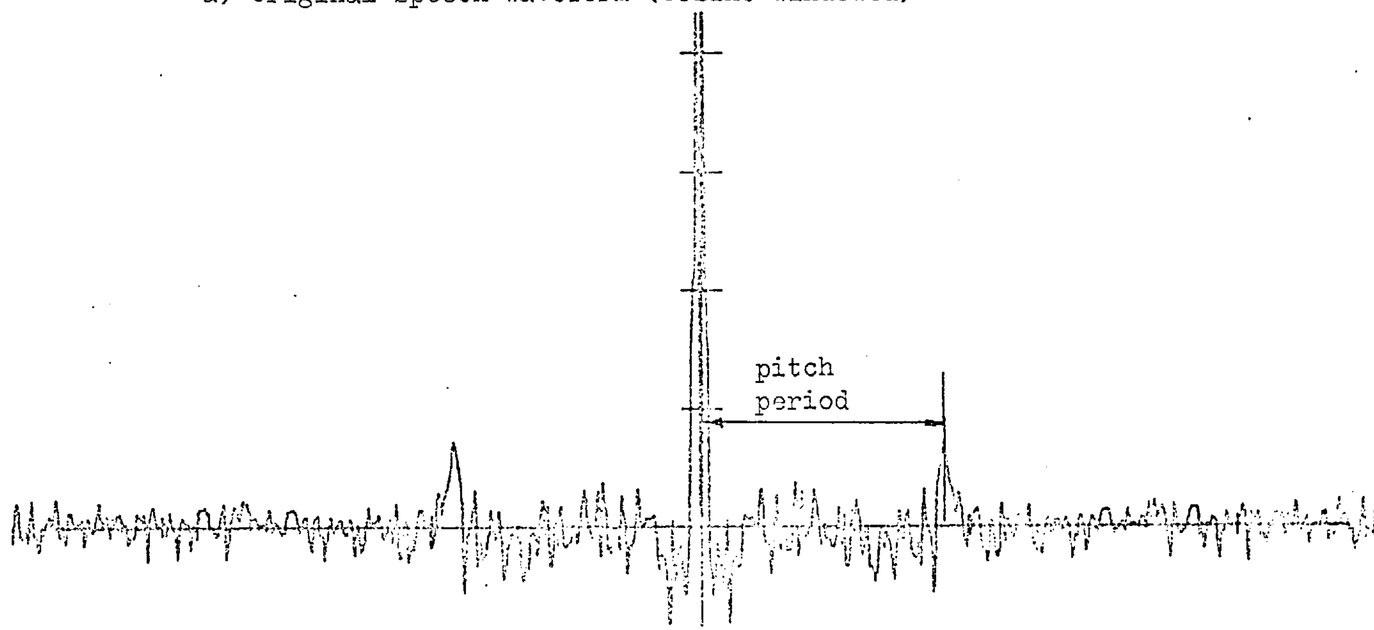


Fig. I.5. A voice excited channel vocoder



a) Original speech waveform (Cosine-windowed)



b) Cepstrum of the waveform above

<u>Fig. I.6.</u>

CHAPTER II

A Mini-Computer Based Speech Analysis and Processing System

The speech analysis and processing system to be described in this chapter was developed primarily for the purpose of performing the research described in this thesis. Initially, specialized programs were written for each particular research purpose but the inconvenience involved in writing new programs for slightly different tasks led eventually to a major re-think of the programming philosophy. Common operations and essential functions frequently required in speech research were identified and a system of useful and flexible routines was gradually built up. The system is by no means complete, but it is designed so that it is capable of expansion by simply adding new routines. In developing the system special emphasis was placed on the fact that, as far as possible, any processing or analysis requirement needing a detail knowledge of the intricacy of the computer operating system should be built into the basic routines. Using these basic routines, extremely complex manipulations of speech signals can be handled by any person having a knowledge of the FORTRAN programming technique. Hence, in its present state, this system is not only a convenient tool for the author's own research but should also prove to be useful for others researching the subject.

This system was built around the Electrical Engineering Departmental PDP 15 computer, with DOS-15 (Disc Operating System) system software. The peripheral devices include a fixed head disc store, two DEC tape transports, an A/D convertor, a D/A convertor, a line printer, and a teletype unit. Supplimentary analogue devices, which can be 'hooked on' to the A/D and D/A convertors are an X-Y plotter a variable frequency clock generator, a high quality audio tape recorder, an FM tape recorder and two variable cut-off frequency analogue low-pass filters. Within the possibilities of these available hardware facilities, the basic functions considered necessary for speech analysis and processing, and which are built into the present system, can be divided into the following two categories.

(1) Speech data acquisition, handling, storage and output.

(2) Graphic output through the X-Y plotter.

Section 1 and 2 in this chapter deal respectively with these two categories. Some specialized programs which could be of service to future research are described in Section 3. The discussion in Section 4 points to parts of the present system that could benefit from modifications and suggestions are also made relating to some possible additions to the system.

II.1. Speech Data Handling

The A/D and D/A convertors are the only link between the digital world of the computer and the analogue form in which speech signals originally exist and ultimately take. To transfer speech signals into or out of the computer, sampling routines for the A/D and D/A convertors have first to be designed. Relatively uncomplicated speech processing can be preformed synchronously with the data input, if the input speech signal is digitized through the A/D convertor, processed and, if required, the result sent out through the D/A convertor, all within one sampling period. This method obviates the need for extensive data storage. An example of this is the speech quantization and clipping program described in Section 3. The application of synchronous processing is not limited to real-time processing. Complicated processing requirements that are beyond the real-time capabilities of the computer can also, in principle, be handled in a similar manner, by using the FM tape recoder to slow down the rate of data input and output. However, the synchronous processing method is inconvenient for the following reasons:

(1) The synchronous processing method often takes a longer time to operate on a section of data than a method that is not synchronized to the sampling rate, as the sampling frequency has to be sufficiently slow to allow for the successful completion of the longest possible processing operation that may be required on a single sample.

(2) Consistency in the input data is impossible to achieve. Supposing that the SNR of two coding methods are to be compared, in two separate runs using the same (analogue) source material. The two sections of speech sampled in these two runs can never be identical because of the differences in the starting point, slight changes in the sampling frequency and in the amplifier gain. As a result, the validity of the SNR comparisons may be slightly suspect.

(3) A laborious setting-up procedure has to be followed each time the equipment is shut down or altered. For example, the input levels have to be adjusted so as to make optimum use of the full dynamic range of the A/D convertor; D.C. drifts in the booster operational amplifiers have to be compensated correctly, sampling frequency alignment has to be carried out, etc.

Thus, with these points in mind, the general speech data handling routines (other than the sampling and output routines) were designed for sampling-rate-independent operation. Maximum usage was made of bulk storage devices, with the digital DEC tape being used as the permanent store and a specially assigned area on the disc used as a temporary sketch store, and the memory cores of the computer were made to serve the function of data transfer buffer. With this arrangement, speech utterances can be stored permanently on digital tapes and, whenever required, read directly into the computer for processing. The following routines are available under the file name, DATRAN. They are all in the form of subroutines and can be called from a FORTRAN main program. All subroutine arguments are integers.

A. INITAD :-

This subroutine initializes the A/D-convertor handler to accept external sampling pulse interrupts, and it should be called once in the program before any analogue data input or output.

B. DATIO(IOPT,IDKFBK,NBLK) :-

This subroutine handles the analogue data input and output. In the input mode, samples of the input data are taken from the A/D convertor at a sampling rate determined by an external clock pulse generator, trimmed and packed, and used to fill up a buffer in the core memory. This buffer area is always assigned when the file DATRAN is loaded. Programmer using this routine does not have to bother about this buffer area. The trimming of the A/D converted 12-bits per sample signal down to one of 9-bit accuracy is done semi-exponentially, using 7 bits of mantessa and 2 bits of exponent to the base 2. This means that the resolution of low-amplitude signals is equivalent to 10-bits per sample of linear quantization.

Every two consecutive 9-bit samples of the data are packed in the core memory into an 18-bit storage word. When the first core memory buffer is full, its content is transferred to the disc, while a second buffer takes over the storage of the incoming data. This operation is reversed when the second buffer is full, provided, of course, that, by then, the first buffer has been properly emptied onto the disc. The change-over of buffers cannot take place if the unloading buffer is not completely emptied and a "SAMPLING RATE TOO FAST" error message will be printed on the teletype. If all goes well, and the transfer of the required blocks of data is completed, a "TRANSFER COMPLETED" message will be printed. In a manner similar to the above, in the output mode data are read from the disc, unpacked and decoded, and sent out through the D/A convertor at a rate controlled by the external clock pulse that is received by the A/D convertor. The maximum clock rate that can be handled with the existing hardware is about 11 KHz on input and about 10.5 KHz on output. For applications where the sampling rate requirement is higher than this, the analogue signal can be 'slowed down' with the FM tape recorder, and a sampling frequency within this hardware capability can be used on the 'slowed-down' signal. The arguments to this subroutine that have to be specified by the calling program are:

IOPT :- Mode of operation.
+ve for input mode,
-ve for output mode.

IDKFBK :- First (relative) disc block to begin transfer.

The available disc area starts at (relative) block No. 0 and ends at block No. 255. The total size of the available disc area is 256 blocks. Each block of data contains 256 storage words and hence, stores 512 samples of data.

NBLK :- Number of blocks to be transferred.

An automatic safety check is made in the routine to prevent data being transferred from (or to) any area of the disc other than the 256 blocks specifically reserved for data storage.

C. RDMSD(IDIV,IBUF,IFSTBK,NBLK) and WTMSD(IDIV,IBUF,IFSTBK,NBLK):-

These subroutines are for communication between the mass storage devices (the disc and the DEC tape) and the computer core memory. Using RDMSD, specific block, or blocks, of data can be read from a mass storage device, unpacked, decoded and loaded into a user specified area in the core. The process is reversed when WTMSD is used and data in the core is written on to the mass storage device. These two subroutines share the same arguments. They are listed as follows.

IDIV :- Mass storage device designation.

7 for disc,

1 for tape.

IBUF :- Name of user specified core array (integer) to accept (RDMSD) or provide (WTMSD) data from/to the mass storage device.

The size of this core array has to be equal to, or greater than 2*NBLK*256.

IFSTBK :- The block on the mass storage device where transfer is to begin.

NBLK :- Number of blocks to be transferred.

D. TKTRAN(ITDRN,IDKFBK,IDTFBK,NBLK):-

This subroutine effects a direct data transfer between the two mass storage devices, the disc and the tape. A block of data is read from the donor-device into an internally assigned buffer core array and written directly from this buffer on to the acceptor-device. This is repeated until the total data transfer required is completed. The arguments of this subroutine are:

<u>ITDRN</u> :- Transfer direction

+ve for transfer from tape (DT) to disc (DK),

-ve for transfer from disc (DK) to tape (DT).

<u>IDKFBK</u> :- The block on disc where transfer is to begin.

<u>IDTFBK</u> :- The block on tape where transfer is to begin.

<u>NBLK</u>:- Number of blocks to be transferred.

A programming example using these four subroutines is shown below. This program reads into block Nos. 0 - 26 of disc a complete sentence of speech stored in block Nos. 32 - 58 (27 blocks) of digital tape, processes the speech data by a subroutine called PROCES(IIN,IOUT), plays the processed data through the D/A convertor for subjective evaluation and finally stores the processed data in block Nos. 60 - 86 of the digital tape. An external clock pulse has to be supplied to the A/D convertor for the purpose of clocking the D/A output. This external clock can be left on all the time the program is running.

```
C
C       PROGRAM DIREXM
C
        DIMENSION IIN(512),IOUT(512)
C
        CALL INITAD
        CALL TKTRAN(1,0,32,27)
        DO 100 I=1,27
        IA=I-1
        CALL RDMSD(7,IIN,IA,1)
        DO 200 J=1,512
200     CALL PROCES(IIN(J),IOUT(J))
        IB=IA+128
100     CALL WTMSD(7,IOUT,IB,1)
        CALL DATIO(-1,128,27)
        CALL TKTRAN(-1,128,60,27)
C
        END
```

II.2. Graphic Output by X-Y plotter

The essence of graph plotting using an X-Y plotter is to supply a timed sequence of analogue signals to the X and Y axis inputs of the plotter so as to control the pen movements in the X-Y plane, and also to provide a third signal to control the up-down motion of the plotting pen. The program developed for this plotting routine contains certain novel features and it is, therefore, worth recording at this point. A design objective of the plotting routine was that the computer should not be held-up while a graphic output is being produced at the plotter. In other words, computations should be able to continue in the computer while command signals are being sent to the X-Y plotter at a slow but constant clocking rate. This constant clock pulse train is provided by a 50 Hz real-time clock in the computer. A clock pulse from the real-time clock interrupts the main program, performs some calculations and sends the appropriate set of command signals to the X-Y plotter. Following this, control returns to the main program which continues from the point of interruption. On the next clock pulse, the main program is again interrupted, and calculations are again performed to decide what the next set of command signals to the plotter should be. These new calculations could follow a procedure different from the previous procedure. Suppose that the plotter is to draw straight lines joining three points.Between the first point and the second point, the calculations to be performed have to interpolate towards the second point so that a sequence of intermediate points are generated, one at each clock interrupt, so as to drive the plotter pen gradually through these intermediate points and thereby plotting a straight line between the first point and the second point. As soon as the second point is reached, however, the procedure to be followed is entirely different from the last interrupt cycle. It has to establish the new direction of interpolation between the second point and the third point and has to proceed to generate the intermediate points lying on the line between the second point and the third point. Thus the routine into which a clock interrupt should be directed, just before the second point is reached, is different from the routine entered on the next clock interrupt. Hence a programmed sequence of interrupt routines has to be arranged so that successive clock interrupts are directed to the appropriate routines each effecting a certain action on the X-Y plotter and the sum total of these actions forming the desired plot. It is possible to set up a list of routine names such that the first clock interrupt would be directed to the routine whose name is at the top of the list and the next clock interrupt would automatically go to the next routine on the list, and so on. However, this read-from-the-list method is essentially only suitable for sequential operation (i.e. the execution pointer goes down the list step by step) as there is no provision for looping or jumping to shorten the length of the list, in those cases where there are repetitive operations. Therefore, if the plot is more complicated than the three-point curve example discussed above, a very long list has to be set up.

The normal method for tackling this problem is to use a slave central-processor (another computer) that can be programmed to execute, one at each clock pulse, a planned sequence of tasks. An example of this, is the PDP interactive graphic display unit, VT 15. It has its own processor (graphic processor) that reads from an area (the display file) of the main computer memory at a rate decided by the display clock. The instructions in the display files, which are executed by the graphic processor, consist of display commands (which control the deflections and intensity of the electron beam) and some basic jump instructions (only recognizable by the graphic processor). The display file is set up by the main program before display commences. Once the display is started, the graphic processor goes through the display file on its own, leaving the processor of the computer free to carry on with the execution of the main program. Programming of the display file has to be done in the machine (or Assembler) language.

The programming technique evolved for this plotting routine, can be likened to the creation of a virtual software processor within the computer. Whenever a clock interrupt occurs, this software processor "steals" some time from the running of the main program in order to execute a portion of a "Procedure File". The plotting "Procedure File" directs the appropriate plotting action at each successive clock interrupt and is written like a normal FORTRAN subroutine. This "Procedure File" is made up of all legal Fortran operations (including the use of subroutines) plus some special basic plotting routines. The set of special basic plotting routines includes all fundamental plotting commands required with an X-Y plotter namely :- 'Pen-up', 'Pen-down', 'Move-pen' (fast movement between two points without bothering what intermediate path the pen may take) and 'Interpolate' (straight line movement between two points). Depending on the nature of the basic plotting routine, and the distance a certain pen movement has to travel, it takes one or more clock cycles to complete the operation of one basic plotting routine. Clock interrupts always enter into the same clock handler.

While inside the clock handler, the processor is made to execute a "Procedure Switch", which directs the program counter of the computer to an address called the "Execution Pointer". At the start of a plot, the "Execution Pointer" is moved to the top of the "Procedure File". As the first clock interrupt comes in, execution of the main program is halted and the processor is directed to execute from the top of the "Procedure File". When a basic plotting routine in the "Procedure File" is entered into, a certain plotting function is performed. On exit, however, this basic plotting routine, unlike the usual subroutines, does not return so as to carry on with the execution of the program from which the subroutine was called. (i.e. It does not return to the "Procedure File"). Instead, it returns to the clock interrupt handler, at a point just below the "Procedure Switch" and, eventually, control is returned, via the clock handler, to the main program which had been interrupted. This completes one clock interrupt cycle. If the execution of the basic plotting routine is not completed by this clock cycle, before leaving the basic plotting routine, the "Execution Pointer" is set to a new point within this basic plotting routine so that when the next clock interrupt occurs the clock handler will again direct the execution, via the "Procedure Switch", straight to this new entry point within this basic plotting routine. If the complete execution requires only one clock cycle, or the job to be performed by this basic plotting routine is completed by the present clock cycle, then, before exit, this basic plotting routine changes the "Procedure Switch" such that the "Execution Pointer" now points to the line in the "Procedure File" immediately below the statement where this basic plotting routine was called. Hence, with successive clock interrupts, the "Execution Pointer" gradually moves down the "Procedure File" and a plot is produced accordingly. The execution of the main program is temporarily halted whenever a clock interrupt occurs and is resumed whenever a cycle of a basic plotting routine is completed. At the end of a "Procedure File", the clock has to be stopped and the plotting action terminated. A special, single clock cycle, basic plotting routine, ENDPLT, that does not produce any output to the X-Y plotter, is used for this purpose. A plotting "Procedure File" must always end with a "CALL ENDPLT" statement.

Using the available set of basic plotting routines :- PENCON,MOVE,INTPOL and ENDPLT (available under the file name PLTCON), experienced programmers should find no difficulty in designing "Procedure Files" for their own particular plotting requirements. For the purpose of waveform tracing a versatile plotting subroutine, PLOTTR (X,Y,INDATA,XINCRM,XEXTRM,YEXTRM,XCAL,YCAL,INEW,ISZPSN, ICMPLT), has been designed using the above mentioned "Procedure File" structure. This subroutine uses subroutines from the following files:- IXPLOT, XYPLOT, AXIS, PRCDR. It accepts data input arrays X and Y, performs the necessary scaling, draws X and Y axes, including calibration marks, and plots the X and Y data in one continuous solid line. Various options can be specified through its arguments. An X-Y plot option joins up successive points formed by the corresponding number pairs in the X and Y arrays and incremental X plot option automatically increments X by constant amounts for each successive element in the Y array. The size and the position of the plot can also be specified such that the final graph occupies a full, one half or one quarter of the available plotter table area. For the purpose of waveform comparison, two successive plots can be superimposed on the same axis, by using the "New" of " Old" graph option.

The arguments of this subroutine are:

X :- X array.

If the incremental-X plot option is used, only the starting value of X in X(1) is required.

Y :- Y array.

INDATA :- Array size.

XINCRM :- X increment.

+ve for incremental-X plot option,

zero or -ve for X-Y plot option.

XEXTRM :- For program controlled scaling of X

XEXTRM(1) - positive X extreme,

XEXTRM(2) - negative X extreme.

YEXTRM :- For program controlled scaling of Y.

YEXTRM(1) - positive Y extreme,

YEXTRM(2) - negative Y extreme.

XCAL : - On exit gives units of X per X-axis division.

YCAL :- On exit gives units of Y per Y-axis division.

INEW :- New/Old graph option.

+ve for new graph,

zero or -ve for super-imposition on the last drawn axes. ISZPSN will be ignored for this option.

ISZPSN :- Size and position option (array of 2).

Indicated below as ISZPSN(1),ISZPSN(2)

-ve, +ve	+ve, +ve
-ve, -ve	+ve, -ve

0 , +ve
0 , -ve

0 , 0

ICMPLT :- Completion indicator.

This is set by PLOTTR to 0 at the beginning of plot and to 1 at completion.

This plotter normally performs an auto-scaling such that the plot spans the available X and Y ranges. For a "New" graph, calibration marks on the X and Y axes are spaced such that each division on the axes corresponds to an integer power of 10. When a plot is to be superimposed, calibration marks on the last drawn axes have to be adopted. Then, scaling is done such that the maximum use is made of the available X and Y ranges, while each division on the X and Y axes is limited to one figure of significance. Supposing that the axis calibrations of the two superimposed curves have to be identical, say, for the purpose of waveform comparison, this can be done through a program controlled scaling, by manipulations of the values of the arguments, XEXTRM and YEXTRM. They are simply treated as X and Y data by the scaling routine, but are not plotted. Hence, if the respective extremes (i.e. maximum and minimum) of the two arrays of X and Y data, including the XEXTRM and YEXTRM, are made identical for two successive plots, the resulting axis calibrations would also be identical. On the other hand, if auto-scaling is desired, both elements of XEXTRM and YEXTRM should be set to zero.

This plotter, when called will first bring the plotting pen to the bottom right corner of the plotter table and output "START?" on the teletype. Plotting commences on any keyboard input. It will then type "RESTART?". If the restart of the same plot is desired, a "Y" is replied. On receiving any other reply, the plotter returns control to the calling program.

A programming example of PLOTTER is given below. This program traces the waveform of the speech signal stored in block Nos. 25

and 26 on the disc. Two half-size graphs are produced, with the waveform of the signal in block No. 25 occupying the top half and that in block No. 26 occupying the bottom half of the plotting paper. Auto scaling is used. The plotted graphs are shown in Fig. II.1 and the teletype print out in Fig. II.2. The teletype message shows that the calibration per axis division is 10^2 on the X-axis and 10^4 on the Y-axis for the first graph, and 10^2 on the X-axis and 10^3 on the Y-axis for the second graph.

```
C
C       PROGRAM PLTEXM
C
        DIMENSION X(1),XEXTRM(2),YEXTRM(2),ISP(2),Y1(512),Y2(512)
        COMMON IY1(512),IY2(512)
        EQUIVALENCE (IY1(1),Y2(1))
C
        X(1)=-32.
        XEXTRM(1)=0.
        XEXTRM(2)=0.
        YEXTRM(1)=0.
        YEXTRM(2)=0.
        ISP(1)=0
C
        CALL RDMSD(7,IY1,25,2)
        DO 10 I=1,512
10      Y1(I)=FLOAT(IY1(I))
        ISP(2)=+1
        CALL PLOTTR(X,Y1,512,+1.,XEXTRM,YEXTRM,XCAL,YCAL,+1,ISP,ICMPLT)
        WRITE(5,1000)XCAL,YCAL
1000    FORMAT(1X,"XCAL=",1PE8.1,2X,"YCAL=",1PE8.1)
C
        DO 20 I=1,512
20      Y2(I)=FLOAT(IY2(I))
C
100     IF(ICMPLT.EQ.0)GO TO 100
C
        ISP(2)=-1
        CALL PLOTTR(X,Y2,512,+1.,XEXTRM,YEXTRM,XCAL,YCAL,+1,ISP,ICMPLT)
        WRITE(5,1000)XCAL,YCAL
C
200     IF(ICMPLT.EQ.0)GO TO 200
C
        STOP
C
        END
```

II.3. Some Useful Programs

During the course of the research many complete programs have been developed. Most of these were designed for very specific purposes and would be of no interest to other researchers. However, two particular programs that may be of value to others are described briefly immediately below.

A. A variable quantization and clipping level PCM simulation program* (PCMCLP):-

In subjective intelligibility measurements, it is often useful to have a reference standard against which the performance of the new coding methods can be compared. PCM, being the most commonly used method is ideal as a comparison standard. Statements of subjective intelligibility often take the form of "....% correct score with monosyllabric words" which can mean very little since the use of a different word list, or the same word list spoken by another speaker, can result in significantly different intelligibility test scores. A more consistent subjective intelligibility measure should be a statement of the kind "at ....% level of confidence, the subjective intelligibility is better than .... bits per sample linear PCM on the same test material with... dB of clipping". Then anyone interested in the subjective quality of the new coding method can simulate a linear PCM system, with the specified bit rate and clipping level, and thereby obtain an indication of the quality (of the new coding method).

---

* This program has been used in a subjective study of perceptual effects of quantization, clipping level and signal bandwidth on speech intelligibility[81].

The program, PCMCLP, can conveniently be used to generate a linear coded PCM test standard. It performs real-time clock-synchronous processing on the input speech signal. The input signal enters through the A/D convertor of the computer and the processed output is simultaneously available at the output of the D/A convertor. An external clock pulse, whose frequency is equal to the required sampling frequency has to be supplied to trigger the A/D convertor. Both the quantization (bits per sample) and clipping level can easily be specified through the computer console switch register keyboard. The procedure for doing so is briefly explained below.

The console switch register keyboard consists of 13 toggle switches which are numbered 12 - 0 from left to right. In the program, the switches from 12 to 1 are for the purpose of selection of quantization and clipping levels. If they are all in the 1 state, the coding system selected is a full 12-bits per sample linear PCM coding with no clipping. If the switch No. 12 (the left hand most switch) is changed to the 0 state, it means 11-bits per sample quantization with 1 bit (i.e. 6 dB) of clipping. If the switch No. 11 is also changed to the 0 state, the clipping level is increased to 12 dB. The program counts, from the left-hand side, the total uninterrupted number of switches in the 0 state, and takes that as the clipping level and, following that, the total uninterrupted number of switches in the 1 state and uses that as the number of bits per sample. Thus, if the states of the switches, Nos. 12 to 1 are, from left to right, 0,0,1,1,1,1,1,0,0,0,0,0, the coding method selected as a 5-bits linear PCM with 12 dB of clipping. If switch No. 8 is changed to the 0 state resulting in a sequence of console-switch states, from left to right, of 0,0,1,1,1,0,1,0,0,0,0,0, the coding method selected becomes 3-bits linear PCM with 12 dB of clipping. The isolated 1 state of switch No. 7 is disregarded. To change from one coding method to another, the desired new coding method is selected on the console switches Nos. 13 to 1, then the state of switch No. 0 is changed. The change-over of coding system only takes place after the state of the switch No. 0 is changed.

B. A speech data recording and playback program (DTEDIT):-

This program was built using the DATRAN subroutines and is very useful for many general purpose speech data recording, playback, and transfer operations. It has three basic modes of operation, namely:

1) RECORD:- for transfer of input speech from the A/D input to the computer disc store, the transfer rate being synchronized by an external clock pulse.

2) PLAYBACK:- for transfer of speech data from the disc to the D/A output. The rate of sample output is again controlled by the external clock pulse.

3) TKTRAN:- for transfer of speech data between the disc and the digital tape.

It interacts with the operator in a question and answer manner. When loaded, it first prints on the teletype the mode that the program is in. Then the question "CHANGE FUNCTION?" is printed. If this is the required mode, an "N" is typed as the answer. If the mode is not correct than a "Y" is typed. If a "Y" is typed then another mode will be selected by the program and the above procedure is repeated. By answering "Y", repeatedly, the desired mode will be reached ultimately. Once a mode is entered on receipt of the answer "N", the program will ask further auxiliary questions to establish the parameters of the transfer. For example, the starting block and the number of blocks to be transferred etc. will be requested. The operator can always extricate himself from a mode entered into accidentally by returning a negative number for the number of blocks.

A useful application of the program is in the editing of speech data to be stored permanently on the digital tapes. Because of storage space limitations, it is very wasteful to store the silent intervals between test sentences, or test words as well as the utterances. With this program, a long length of test material can be transferred from an analogue tape recorder via the A/D convertor onto the disc and the material recorded on the disc can then be played back through the D/A convertor to determine the blocks containing the wanted test material. Only those blocks containing the required material are then transferred to a selected area on the digital tape for permanent storage.

II.4. Discussions

Using the basic building blocks discussed in the earlier parts of this chapter, powerful speech processing and analysis programs can be constructed. These routines form the basis of the simulations and studies of the various new coders described in Chapters IV and V. In the course of the work forming the basis of Chapter IV and V, spoken sentences were sampled at various frequencies and stored permanently on digital tapes, so that they were immediately available as source material for use in evaluating the performances of various coding systems. In preparing the digital tapes, the level of the analogue speech signal at the input to the A/D convertor was carefully adjusted such that, for each complete spoken sentence, the sample of highest amplitude was just on the point of overloading the A/D convertor. After this adjustment, the power level of each recorded sentence was of approximately the same value and this thus obviated the need for amplitude normalization of the speech data within the computer.

The speech processing and analysis system has been used extensively by the author and a number of his colleagues and no major short-comings have been discovered. However, as experience has been gained, it has been found that there are a number of small alterations that would be beneficial and these are summarised as follows:-

1. Speech trimming:- In the system, speech is sampled and the A/D converted 12-bit samples are trimmed to 9-bit words before storage, so as to save storage space. An exponential trimming with 7 bits of mantessa and 2 bits of exponent to the base 2 is used, and this corresponds to a piece-wise-linear μ-law logarithmic quantization with $\mu = 32$. In speech coding, a value of 100 or more is generally accepted as being the most suitable value for μ. Thus, it is desirable that the trimming algorithm should be modified so that the mantessa part is reduced to 6 bits and exponent part increased to 3 bits; which corresponds to $\mu = 256$. It should be pointed out, however, that a slight drawback of using $\mu = 256$ is that resolution at low signal amplitude is equivalent to 13 bits linear PCM; which is 1 bit greater than the resolution available from the A/D convertor. Some information storage capability is therefore wasted.

2. X-Y plotting speed:- The plotting routine has been designed so as to maximize the plotting speed without degrading the clarity of the resulting plot. This, however, appears to place a mechanical strain on the X-Y plotter. Although the plotting speed can be

reduced by simply changing the value of a constant term in the PLTCON routines, it appears to be more desirable to alter the plotting routines in an alternative manner so that the plotting speed can be specified by the user as one of the subroutine arguments (say, 4, meaning full speed and 3, three quarter speed etc.).

The analysis and processing system is capable of extension simply by the addition of new routines which could fit into the general data structure. Welcomed additions would be routines for spectrogram generation and printing, a routine for simulating a recursive sharp-cut off low-pass filter and a routine for rapid computations of auto-correlation functions. Spectrographic data from the speech signal can be generated readily by the use of a FFT routine[78]. Other useful information regarding spectrogram generation and printing can be found in Ref.(80). The low-pass filtering routine can be designed such that the filter coefficient values are automatically calculated with the specification of the cut-off frequency. Comprehensive accounts of digital filter designs can be found in Ref. (79). For the fast computation of auto-correlation functions, the algorithm by Lopresti et al.[82], based on the fast Walsh transform technique, can well be used.

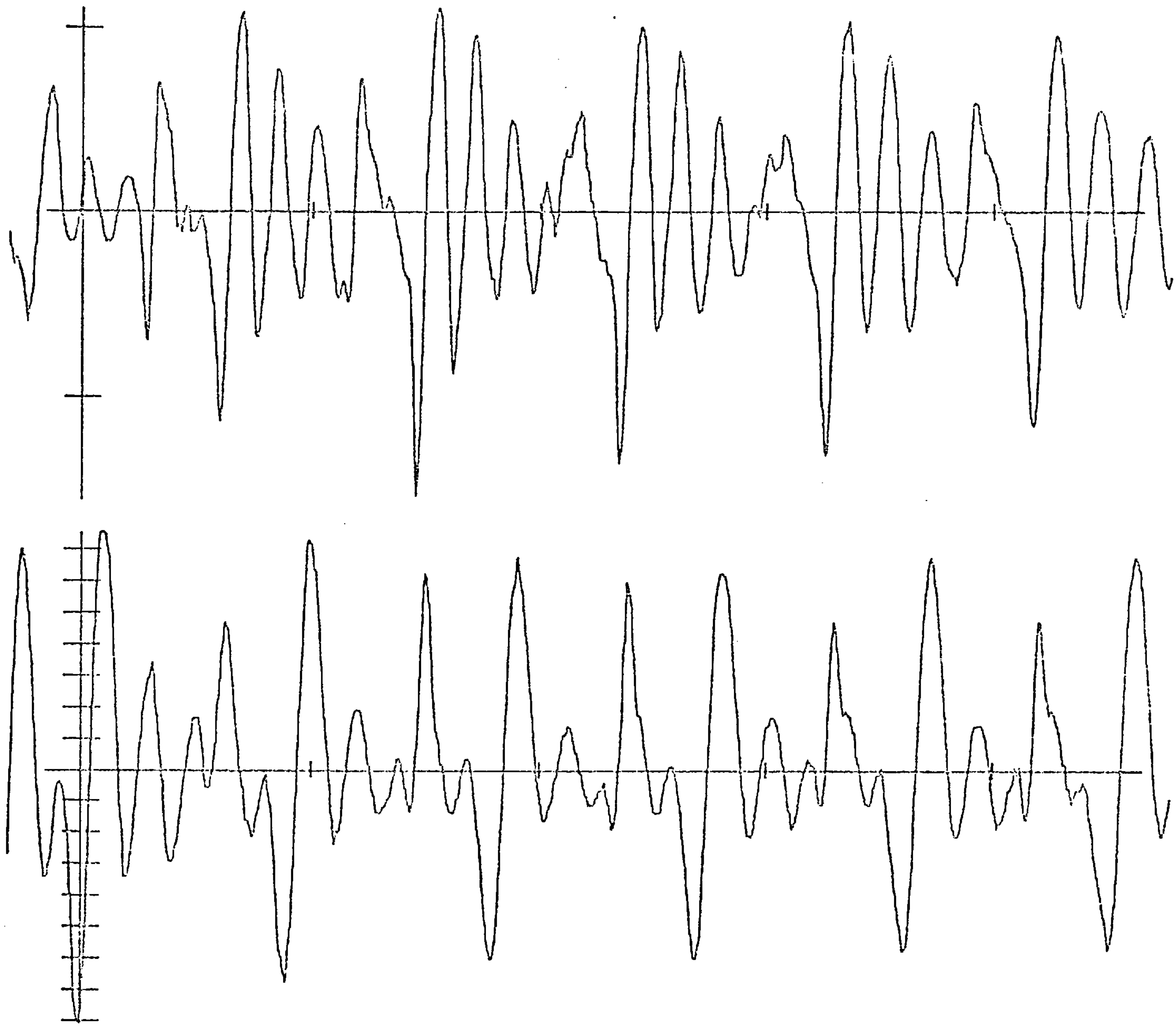


Fig. II.1. Plotter output examples

```
>←PLTEXM,PLOTTR,PRCDR,AXIS,IXPLOT,XYPLOT,PLTCON,DATRAN
↑S↑S
START?
Y

RESTART?


XCAL= 1.0E+02   YCAL= 1.0E+04
START?


RESTART?


XCAL= 1.0E+02   YCAL= 1.0E+03
```

Fig. II.2. Teletype output corresponding to Fig. II.1

CHAPTER III

Amplitude Dithering for Speech Quantization

With normal fixed-level methods of quantization, such as those used in the conventional PCM encoding of speech signals, the quantization error tends to be signal-dependent or signal-correlated. At low bit-rates, where the quantization levels are necessarily coarse, the effect of such signal-dependent quantization error is especially annoying and fatiguing. In investigations into digital picture coding, Roberts[83] found that the technique of pseudorandomly dithered quantization tended to smooth out the abrupt changes in grey scale resulting from coarse quantization. It was found that, given the same number of bits of quantization per sample, the variance of the quantization error with the dithering method was the same as that with the fixed-level method, but the received picture appeared perceptually to be better in quality by at least 1 bit per sample.

The question naturally arises as to whether pseudorandomly dithered quantization will yield a similar improvement when applied to the encoding of speech signals. Experimental investigations have been conducted by Wood and Turner[13] and Rabiner and Johnson[12] in this respect. The conclusions reached in two independent studies were almost identical. It was found that the subjective preference is always improved but the subjective intelligibility of the coded speech is not necessarily improved. Rabiner and Johnson investigated a wide range of bit-rates*, and found that the dithered quantization

* The term "bit-rate", as used in this chapter is synonymous with the number of bits per sample as a fixed sampling frequency is assumed.

results in a speech reproduction that was more intelligible at high bit-rates but less intelligible at low bit-rates, with the cross-over point occurring at about 4 bits per sample.

The improvements in subjective preference can probably be attributed to the "whitening" of the quantization noise due to the method of dithered quantization. It has been shown in Jayant and Rabiner[11] that this dithering of the quantization levels produces a quantization error that is independent of the input signal and the abrupt steps of quantization error that is normally associated with low-bit-rate fixed-level PCM tend to be "smeared" by this very action. However, the "whitened" quantization noise has the adverse effect of masking the consonant sounds to a greater extent than the normal fixed-level type of quantization[12]. This is quite understandable since a large number of consonant sounds are unvoiced, i.e. are rather noise-like and lack sharp formant peaks in the power spectrum shape. Thus, it is likely that the addition of a signal-independent white noise which is of not inconsiderable amplitude at low bit-rates could confuse the auditory perception system and lead to difficulties in the identification of consonants.

It is a pity that the method of pseudorandom dithering, although it improves the subjective quality of coarsely quantized speech, should also reduce its intelligibility. Coarsely quantized speech without dithering is unacceptable mainly because of its appalling harshness. Low-bit-rate coding of speech by the method of fixed level quantization actually retains a high level of intelligibility. Licklider's experiments[85,86] on amplitude dechotomized, time quantized speech waves[85] has demonstrated that the intelligibility of speech is not significantly destroyed, even with infinite clipping, which is equivalent to one-bit per sample quantization. Thus if a low-bit-rate quantization method could be found that reduces the harshness of the quantized speech and at the same time achieves at least as good a degree of intelligibility as that achieved by the normal non-dithered method of quantization, it would find application in those areas in which low-bandwidth, low-cost and simplicity are at a premium, and absolute clarity is not essential.

Wood and Turner[13] were aware of the results of Licklider[84,85] and Morris[86], concerning the importance of zero crossings in speech intelligibility and they conjectured that the reduction in the intelligibility of dither-quantized speech was due to the irregular effect that the dither had on the zero crossings of the speech signal. The position of zero crossings can be either advanced or retarded by dithering. Also it is possible for an existing crossing to be eliminated or for a new crossing to be introduced. This led Wood and Turner to suggest that there might exist a compromise between low-bit-rate PCM using fixed-level quantization, with its irritating harshness, and that using dithered quantization, with its reduced intelligibility. They suggested that a method of dithered quantization in which zero crossings are preserved might lead to a satisfactory compromise.

There are many ways of preserving the zero crossings of a dither-quantized speech signal. Some require extra bandwidth, some increase the variance of the quantization error and some reduce this variance. Although there are a number of methods of dithering so as to preserve the zero-crossings, they all share one common feature. Because of the signal-dependent alterations to the dithering noise sequence that are necessary in order to preserve the zero crossings of the signal, the resulting quantization error, unlike the quantization noise associated with normal dithering, is signal dependent but not as dependent as that resulting from normal fixed-level quantization. The varying degrees of signal dependence of the quantization error make it very difficult to predict the performance associated with the various methods of zero-crossing preservation.

In this chapter, 5 methods of dithered quantization with zero crossing preservation are described, and the chapter deals with an investigation, by subjective intelligibility testing, of the comparison of the performance of the five new methods of quantization with the performance of the normal fix-level and dithered methods of quantization. In the comparison it was found that two of the methods of dithering with zero-crossing preservation were more intelligible than the normal fixed-level quantization method[14].

III.1. Dithered Quantization

In this section the normal method of dithered quantization is first reviewed and a brief indication is given of how the quantization error that results is statistically independent of the input signal. The effect that the dithering has on the zero crossings of the signal is also demonstrated. Following this, five new dithering schemes are described. The methods were devised so as to suppress the irregular effects of the random dithering on the zero crossings of the speech signal. Each of the schemes results in a quantization error slightly different from the other. Rough estimates of the quantization error variances associated with the five methods are given in the Appendix.

A. Normal pseudorandomly dithered quantization:-

The normal method of pseudorandomly dithered quantization is illustrated in Fig. III.1. A pseudorandom dithering noise is added to the input signal, and, for each example, the sum of the signal and the dithering noise is quantized. The same pseudorandom noise is then subtracted from the quantized sum, and the resulting signal is the dither-quantized version of the input. If the dithering noise has a uniform probability distribution function with a zero mean and a range equal to the step size of the quantizer (as illustrated in Fig. III.2.), the variance of the quantization error remains the same as that without dithering and the quantization error is independent of the input signal. These properties of the dithered quantization error can be demonstrated easily by considering Fig. III.3.

Fig. III.3.a. shows a situation that arises with normal fixed-level quantization, where the amplitude of the input signal, X, falls within the quantization slot $K\Delta$ and $(K+1)\Delta$, and thus is quantized to $(X)_Q = (K+1/2)\Delta$ (see also the quantization law in Fig. III.5.). The quantization error, E_O, is thus given by,

$$\begin{aligned} E_O &= (X)_Q - X \\ &= (K + 1/2)\Delta - X \end{aligned}$$

Note that this quantization error is strongly dependent on the actual position of the signal, X, within this quantization slot. Assuming a constant probability density of X within the slot, $k\Delta$ to $(k+1)\Delta$, then if X does fall into this slot, the quantization error is equal-probable to take any value from $-\Delta/2$ to $\Delta/2$, and not any higher or lower. This error distribution is true no matter which quantization slot the signal, X, falls. Thus, assuming that the quantizer is not overloaded, the error variance is given by

$$\overline{E_0^2} = \int_{-\Delta/2}^{\Delta/2} E_0^2 \cdot p(E_0)dE_0$$

$$= \int_{-\Delta/2}^{\Delta/2} E_0^2 \cdot \frac{1}{\Delta} dE_0$$

$$= \frac{\Delta^2}{12}$$

Fig. III.b. shows a situation where the uniformly distributed dithering noise, N, is added to the signal, X, before quantization. The total shaded area represents the range of values that (X+N) may take given the value of X. A part of this area (slant striped) falls within the quantization slot, $K\Delta$ to $(K+1)\Delta$, and another part of this area (horizontally striped) falls just outside this quantization slot. Assuming that $(X+N_1)$ is inside the slant striped area, i.e.

$$X - \Delta/2 \leqslant X + N_1 \leqslant (K+1)\Delta \qquad \text{----} \qquad (\text{III.1})$$

it will be quantized to $(K+1/2)\Delta$, i.e.

$$(X + N_1)_Q = (K + 1/2)\Delta \ .$$

Thus, after subtraction by the dithering noise, the dither quantized version of X is

$$(X)_D = (X + N_1)_Q - N_1$$

$$= (K + 1/2)\Delta - N_1$$

From Eq. (III.1) ,

$$-\Delta/2 \leqslant N_1 \leqslant (K+1)\Delta - X.$$

Therefore

$$(K + 1)\Delta \geqslant (X)_D \geqslant X - \Delta/2 \quad ,$$

which means that $(X)_D$ again lies in the slant striped area. Similarly, if $(X+N_2)$ is inside the horizontally striped area the final dither-quantized output is also in the horizontally striped area. Thus, for any given X, the dither-quantized signal, $(X)_D$, can take any value between $(X - \Delta/2)$ and $(X + \Delta/2)$ and its probability is uniformly distributed within this range if the dithering noise has a uniform PDF. Hence, the quantization error, E_D, given by

$$E_D = (X)_D - X$$

has a uniform PDF between $-\Delta/2$ and $\Delta/2$, for any given X, and is independent of the position of X. The variance of this quantization error is given by

$$\langle E_D^2 \rangle = \int_{-\Delta/2}^{\Delta/2} E_D^2 \cdot p(E_D)\, dE_D$$

$$= \frac{\Delta^2}{12} \quad ,$$

which is equal to the undithered case.

This normal method of dithered quantization has an irregular effect on the zero crossings of the input signal which is clearly demonstrated in Fig. III.4. It shows a section of a speech waveform that has been reconstructed after quantization both by the fixed-level and the dithering method. It can be seen that the dithering method can result in the zero crossings not only being significantly advanced or retarded, but also deleted or inserted. This arises because the random dithering noise can cause some signal samples to acquire opposite polarities from the original signal samples. It is these reversals of polarity that are the sole cause of the disruptions in zero crossings and they are thus referred to later in this chapter as "false zero crossings". To preserve the original zero crossings of a signal, all that is needed is the suppression of the false zero crossings.

B. Dithered-quantization methods with preserved zero crossings:-

Consider a quantization law as shown in Fig. III.5. False zero crossings can arise only from the dithering of signal samples that have amplitudes in the range $-\frac{1}{2}\Delta$ to $+\frac{1}{2}\Delta$, since the dithering noise has a maximum magnitude of $\frac{1}{2}\Delta$. If a sample is positive and has a magnitude that is somewhere in the range 0 to $\frac{1}{2}\Delta$ and the psuedo-random noise is of the opposite polarity, and has a magnitude greater than that of the signal, the sum of the signal and the noise will be negative and will be quantized to a level of $-\frac{1}{2}\Delta$. Subsequent subtraction of the negative pseudorandom noise, as its magnitude is less than $\frac{1}{2}\Delta$, can never bring the processed sample back to positive, with the result that a false zero crossing is introduced. On the other hand, if a change in polarity does not arise as a result of the addition of the dithering noise, then after quantization has taken place, subsequent subtraction of the dither, which is of magnitude less than $\frac{1}{2}\Delta$, can never result in a false zero crossing.

Thus, provided the dithering noise which is subsequently to be subtracted is limited so that its amplitude is always less than $\frac{1}{2}\Delta$, then false zero crossings can be eliminated if changes in polarity due to the addition of the dithering noise are suppressed at the encoder (the dithering and quantizing part of the system).

The encoder action can be interpreted by considering that it modifies the dithering noise sequence at critical points such that the above mentioned reversal of polarity due to the addition of the dithering noise N, could never arise. There are many ways in which this suppression can be achieved and the following are illustrative examples:

(1) reduce the pseudorandom noise to zero at those instants when a false zero crossing is about to be introduced by the encoder,

(2) reverse the polarity of the pseudorandom noise at those instants when a false zero crossing is about to be introduced by the encoder,

(3) if $(X+N)$ is of opposite polarity to the signal, X, change the dithering noise to N', where $N' = -2X-N$, so that $X+N' = -(X+N)$, which is of the same polarity as X.

In fact, as long as the polarity of the sum of the signal and the modified dithering noise is the same as the polarity of the signal, any modification to the dithering noise at any time instant, including that when the dithering noise would not have caused a false zero crossing at the encoder, will serve the purpose of false-zero-crossing suppression. Clearly, every attempt should be made to preserve the general nature of the pseudorandomly dithered quantization and hence it is desirable to modify the pseudorandom noise as little as possible. Ideally, though not always possible, this means changing the pseudorandom noise only when false zero-crossings have to be prevented.

It will be noted that, with each of the above three examples of modification of the encoder dithering-noise sequence, the output from the encoder is the same, in that, whenever a signal sample has an amplitude in the range of 0 to $+\frac{1}{2}\Delta$, it is always quantized to $+\frac{1}{2}\Delta$, and whenever the amplitude is in the range of 0 to $-\frac{1}{2}\Delta$, it is always quantized to $-\frac{1}{2}\Delta$. The action of the encoder always appears to be one in which the dithering is "frozen" whenever a signal sample has an amplitude in the range of $-\frac{1}{2}\Delta$ to $+\frac{1}{2}\Delta$. In the investigation considered in this thesis, only encoding methods that have this "freezing" property are studied. And, in particular, in the investigation, only two procedures may be followed when modifying the dithering noise they are:-

(1) modification of the dithering noise by reducing it to zero,

and (2) modification of the noise by reversing its polarity.

Although the output from the encoder is not affected by the method of suppression adopted, the overall output obtained from the decoder (that part of the system where the dithering noise is subtracted) depends on the dithering noise sequence subtracted by the decoder from the quantized encoder output. Hence, the overall system performance (associated with the properties of the overall quantization error) will clearly depend on the noise sequence used at the decoder. Ideally, the method of pseudorandomly dithered quantization requires that a synchronized dithering noise sequence, identical to that added by the encoder, be subtracted at the decoder. In this way, the variance of the quantization error is reduced to a level that is essentially the same as that of normal undithered PCM. As the two proposed procedures of modification of the encoder noise sequence do not result in a modified noise of an amplitude greater than $\frac{1}{2}\Delta$, the use of an identical sequence at the decoder does not re-introduce false zero crossings corrected at the encoder and hence is acceptable from the point of view of false-zero-crossing suppression. One complication is that, if the decoder is to remove a noise sequence identical to that added by the encoder, then it must have detailed knowledge of when the dithering noise is modified at the encoder, and of the way in which it is modified. If the decoder does not have this knowledge, differences, between the pseurorandom sequence added at the encoder and that subsequently subtracted at the decoder, will have to be tolerated.

As mentioned previously, five techniques for eliminating false zero crossings were examined. These techniques, which are described below, can be divided into two fundamentally different categories. In the first category, which is referred to below as "category-1 zero-crossing preservation", false zero crossings are suppressed at the encoder, by modification of the dithering noise sequence, but the decoder is not informed of the suppression. In the second category, "category-2 zero-crossing preservation", false zero crossings are suppressed at the encoder and the decoder is informed of the suppression. Category-1 preservation is particularly important from the point of view of bandwidth compression, since it does not require that any information, other than that relating to the values of the quantized samples, be transmitted. With category-2 suppressions, however, it is necessary to transmit additional information relating to the instants at which false zeros are suppressed.

Although it might appear at first sight that category-2 operation would require that a considerable amount of extra information be transmitted, and that synchronization problems might occur, this is not in fact the case. Although some additional information has to be transmitted, it is possible to inform the decoder of the suppressed false zero crossings, and to do so using very little extra information. One technique whereby this might be done is illustrated in Fig. III.6, in which a conventional Gray code is used for PCM transmission. In the scheme illustrated, the two outer levels of quantization are discarded, and the bit patterns associated with them are allocated to the first positive and negative quantization levels, and are used whenever a false zero crossing is suppressed. This means that, with an n-bit-per-sample quantization scheme, the number of available quantization levels is reduced from 2^n to 2^n-2.

1) Category 1 zero-crossing preservation (false zeros suppressed; no information relating to suppressions transmitted to receiver):-

Three schemes were investigated under this category. They are described below in terms of encoder-decoder pairs. The decoder hypothesizes on the manner in which the dithering sequence is modified by the decoder, and, as far as possible, the decoder modifies its own dithering sequence to match the modified sequence in the encoder. A good hypothesis would be valuable in enabling the decoder to achieve a better matching. Places where these dithering noise sequences at the encoder and decoder differ are pointed out. As the encoder output is not affected by the differences in the schemes, the actual differences in construction between the schemes exist only in the decoder and the different encoder actions are only imagined.

<u>Scheme a</u>: In this scheme, the encoder was structured so that, whenever a false zero crossing would have been introduced by addition of the pseudorandom noise sequence, the dithering was frozen; that is, the

dithering noise was reduced to zero. The associated decoder was structured so that it completely disregarded any possible false zero crossings, and functioned exactly as the decoder in a normal pseudorandomly dithered quantization scheme.

Clearly, with this scheme, the actual noise sequence added at the encoder differed from that subtracted at the decoder when false zero crossings were suppressed at the encoder. Under all other conditions, the two noise sequences were identical.

Scheme b: In this scheme, the encoder was structured so that, whenever the signal amplitude was in the range $-\Delta$ to $+\Delta$, and the noise was of opposite polarity to the signal amplitude, the dithering was frozen; that is, the pseudorandom noise to be added or subtracted was reduced to zero. The decoder was structured so that it examined the received quantized signal and changed the subtracted noise sequence in a manner depending on the received quantized signal level. If the received quantized signal level was found to be at either the first positive or first negative quantization level, and the noise was of opposite polarity, the decoder reduced the pseudorandom-noise sample to zero. Under all other circumstances, the decoder acted exactly as a decoder in a normal pseudorandomly dithered quantization scheme.

It can be seen that, with this scheme, the pseudorandom noise was frozen in the decoder whenever it was frozen in the encoder, and, in addition,

(1) when the amplitude of the signal was in the range $+\Delta$ to $+3\Delta/2$, it was reduced by a negative dither in the encoder to a value below $+\Delta$, and it was then quantized to the first positive quantization level.

(2) when the amplitude of the signal was in the range $-3\Delta/2$ to $-\Delta$, it was increased by a positive dither at the encoder to a value above $-\Delta$, and it was then quantized to the first negative quantization level.

These are clearly the two circumstances when the noise added at the encoder differed from that subtracted at the decoder.

Scheme c: In this scheme, the encoder-decoder pair was structured so that, whenever the input sample was in the range $-\Delta$ to $+\Delta$ and the sum of the input sample and the dithering noise was also in the range $-\Delta$ to $+\Delta$, the encoder reduced the pseudorandom noise to zero, irrespective of its polarity, and the decoder reduced the noise to zero whenever the received quantized amplitude was at either the first positive or the first negative quantization level. It will be noted that, with this scheme, the noise sequence added at the encoder differed from that subtracted at the decoder under exactly the same circumstances as those under which differences occurred with scheme b.

2) Category 2 zero-crossing preservation (false zero suppressed and information relating to suppressions transmitted to receiver)

Under this category, the two schemes described below were investigated. They are again described in encoder-decoder pairs.

Scheme d: In this scheme, both the encoder and the decoder were structured so that the pseudorandom noise was reduced to zero whenever a false zero crossing would have been introduced by the addition of the pseudorandom noise.

Scheme e: In this scheme, both the encoder and the decoder were structured so that the pseudorandom noise was reversed in polarity whenever a false zero crossing would have been introduced by the addition of the psuedorandom noise.

III.2. Subjective Intelligibility Tests

In order to evaluate the effectiveness of the new coding schemes as compared with the normal fixed-level and dithering methods of quantization, subjective tests were conducted and speech utterances coded by the various methods were used in the test. The coding of the speech utterances and administration of the subjective tests were all performed under the centralized control of a PDP-15 mini-computer. The details of the coder simulation, the design of the subjective tests and the operation of the testing system are described in this section.

A. Simulation of the quantization schemes:-

In all, seven quantization methods were simulated. The methods, which are listed in Table III.1, included normal fixed-level quantization, normal pseudorandomly dithered quantization and the five new methods of quantization. When testing each of the quantization schemes, it was possible to specify any one of three different bit-rates (3,4, and 5 bits per sample).

The simulation of various quantization schemes (encoder-decoder pairs) was relatively straight forward with the input speech being processed on a sample-by-sample basis. The general scheme of the simulation system is similar to that shown in Fig. III.1, though the figure does not include the various modifications, to the dithering noise, that were necessary in order to preserve the zero crossings. To obtain normal fixed-level quantization, the dither was simply frozen at all times. The pseudo-random noise was generated by the multiplicative congruntial method[87] and a histogram of the amplitude distribution of the generated random noise is shown in Fig. III.7. As expected, the distribution is approximately flat and fits the dithering noise specifications fairly well.

B. Design of subjective tests:-

The word lists of House et al.[88] were used as the test material for the reasons that subjective tests based on these word lists are particularly suitable for an automated administration and scoring, and that it does not require a trained test crew. In addition, in view of the observations by Rabina and Johnson [12] that consonant sounds are more susceptible to pseudorandom dither, attention was focussed on sounds of this type and word lists constructed on the basis of the recognition (or rather the confusion) of consonant sounds, were chosen for use in the test.

The test material consists of six lists each containing 50 American/ English monosyllabic words. The lists are so arranged that there are 50 ensembles of six easily confused words (one from each list). The six words in each ensemble differ only in the consonant (or absence of consonant) immediately preceding or following the vowel. The total of 300 words in the lists includes all possible consonant sounds and should be presented so that the listener (subject) is given the ensemble of six words and has to identify, out of these six possible choices, one closest to a speech utterance heard. This technique is called "identification from a closed response-set" as distinct from an "open response-set" where the subject hears a spoken word and has to identify the word heard from (almost) infinite number of possibilities. Identification from a closed response-set greatly simplifies the task of automated scoring since with this type of test there is rarely any question of an error arising as a result of incorrect spelling on behalf of the subject due to the subject's unfamiliarity with the word lists or the speaker.

A number of attempts were made at various test designs and initial trial runs were carried out. From the experience gained during these initial experiments, some criteria were laid down for a test design that was both convenient to administer and generally applicable for other subjective-intelligibility-measurement requirements. The criteria decided upon were :

(1) The test results should be capable of being fitted into an analysis of variance framework.

(2) In order to reduce listening fatigue, each subject should not, at any one session*, be given a test containing more than 150 words.

(3) The test should be capable of being extended indefinitely (by using additional subjects) so that the number of test scores could be accumulated until a predesignated degree of statistical significance is reached in the test scores+.

---

* In addition, experience has shown that it is preferable to use fresh subjects rather than to use subjects that have been used previously since learning effects have to be taken into account when using people in the latter category.

\+ With many sophisticated test designs, the total number of tests to be performed is implicit in the design structure. If the test results fail to be statistically significant, the remedy is either to repeat an identical set of tests, or to devise a new test discarding all the old results. Both of these are impractical or or perhaps impossible for the purpose of subjective intelligibility testing.

It was found that if all the criteria were to be met then either a reduced set of test words or unbalanced sets of test words had to be used. A reduced set of test words is undesirable firstly because it limits the variety of sounds that can be used; and secondly because the learning effect of the subjects increases and is difficult to account for. On the other hand, the use of unbalanced sets of test words means that the words used to test each system are different, and also, different words are presented to each subject. This raises some doubts as to the validity of the test results and makes interpretation difficult. Although there are difficulties, the problems associated with the use of unbalanced sets of test words can be overcome if the nature of unbalance is randomized. In such a situation, the effect of the unbalance in the words used for testing the various coding systems manifests itself in the form of random fluctuations in the test scores and is thus accounted for in the total error variance of an "analysis of variance" performed on the test data.

With the above facts in mind, in the final design, the test words were simply selected at random, and with equal probability, from a pool of available test words. Other controlling factors, that could be identified and isolated, were all balanced. For the particular investigation into the relative intelligbility associated with the various quantization schemes, this meant that each subject, in a single test session, encountered all possible combinations of quantization schemes, bit-rates and display positions on the interactive graphic screen*; and conversely, each quantization scheme was tested with all possible combinations of subjects, bit-rates and display positions.

The order of presentation to a subject of the various combinations of the controlling factors was also randomized. This was done in the

---

* See sub-section III.2.C.

following manner. Suppose that the total number of combinations of the various sources of variations is N and a subject is to identify K.N tests words in a session, where for a balanced test K is an integer. A table of the N combinations is first drawn up, and a sequence of combinations is formed by consecutive random selections from the N possibilities. As soon as a particular combination is selected an entry is kept in its corresponding position in the table and if the number of entries in the table corresponding to that combination has already reached K, then the combination is discarded and hence is not added to the sequence. The process is repeated until the table is completely filled.

C. An automated subjective intelligibility testing system:-

The execution of the subjective test was completely automated using a PDP 15 computer together with its associated A/D convertor, D/A convertor and interactive graphics facilities.

The test material, consisting of the six House word-lists, was pre-arranged and stored on digital tapes. The preparation of these digital tapes was done as follows. Initially, a high-quality analogue tape recorder was used in recording the 6 word lists onto an analogue magnetic tape. The word lists were recorded as they were spoken by an American-speaking male person. The words were then low-pass filtered and converted into digital form using an A/D convertor sampling at 6.5 KHz and stored using 9 bits (linear) per sample on six reels of digital tape*. Each reel of tape thus contained 50 words with each word being drawn from a separate set of six easily confused words.

---

* The idea of exponential trimming had not been considered at this stage of research.

As only two tape transports were available, it was not possible to use all the test material simultaneously. To overcome this difficulty and to preserve randomness in the presentation of test material, two of the six tapes were selected at random. The selected material, which thus consisted of 100 easily confused words, was then presented as test material to two subjects, one after the other. After the tests involving each two subjects had been completed, two new subjects were called, two new tapes were selected at random, and the tests were repeated.

As regards the presentation of material to a particular subject, the scheme of events was as follows. A word was selected at random by the computer from the 100 words available on the two tapes, and the ensemble of six easily confused words, of which the selected word was a member, was displayed on the interactive graphics unit. The six words were positioned in random order at fixed points along a horizontal line. At the same time as the word was being selected, and the ensemble of six words was being displayed, one of the seven quantization schemes to be compared was selected at random and one of the 3 bit-rates used was also selected at random. On selection of one of the 21 possible combinations of quantization scheme and bit-rate, the word that had been selected at random for auditory presentation to the subject was processed by the simulation routine corresponding to the chosen quantization scheme and bit-rate, and the processed signal was then passed through the computer's D/A convertor, low-pass filtered and amplified, and then fed to the subject via a pair of high-quality earphones.

The sequence of events perceived by the subject, and the actions to be taken by him, were as follows. First, a list of six words appeared before him on the screen of the display unit, and then, after about 0.5 sec. the spoken word was heard. The six words remained on the screen for 5 sec. during which time the subject was required to select which of the six words he considered the spoken word to be. After the 5 sec. period had elapsed, the set of words was removed from the screen, and a new set appeared in readiness for the next test. In recording his decision, the subject used a light pen that automatically informed the computer of the subject's selection. The computer was programmed so that the set of words displayed, the actual spoken word, the word selected by the subject and the score obtained by the subject was recorded automatically, both in line-printer print-outs and punched paper tapes. The subject was awarded a score of one for a correct identification, and a score of zero for an incorrect identification or for failure to make a decision in the allotted 5 sec. time interval. As the test proceeded, the paper-tape records of the test results were accumulated and then, from time to time, these results were fed as input to an "analysis of variance" program, this continued until sufficient subjects were tested and a pre-decided level of statistical significance thereby reached.

The test program required a certain setting-up procedure when it was first loaded. Information about the number of quantization systems and the number of different bit-rates to be tested had to be fed in. This information initialized the program to a detailed test procedure, in accordance with the structure discussed in the previous sub-section. Among other things, it enabled the program to decide on the number of tests to be performed in a session. This number was simply the total number of combinations of the quantization schemes, bit-rates and display positions, multiplied by the largest integer keeping the product less than 150. It was by this facility that the program is rendered universally applicable to the subjective comparison of other speech processing systems. All that is necessary to achieve this is the substitution of the appropriate simulation routines.

III.3. Statistical Analysis of Intelligibility Test Results

A total of 36 subjects were tested, with each subject being asked to identify 126 words. The tests were organized so that each of the 21 possible combinations of quantization schemes and bit-rate were tested 216 times; i.e. 216 words were used in testing each of the 21 combinations.

The results of the tests, in which a scoring system as described in Section III.2.C. was adopted, are shown in Table III.2, and the results of an "analysis of variance" on the test data are presented in Table III.3.

In carrying out the analysis of variance, a factorial design was used in which the controlling factors were: the number of bits per sample (B), the quantization scheme (Q), the subject (S), and the position (P) at which the correct word was displayed on the screen . Also, a mixed effect model was used, with the factor "subject" being considered as the random-effect factor and other factors as fixed-effect factors. The F ratios were computed using the method presented in Guenter[89], and the results of the analysis show that all the factors were significant at the 99.5% confidence level, with the bit-rate being the most significant factor. All interactions were either not significant or below the 75% confidence level.

In order to evaluate the relative effectiveness (quality) of the seven quantization schemes, Duncan's multiple-range test[90] was applied. The results of applying this test are shown in Table III.4, in which the quantization schemes are arranged from left to right, in order of merit. The values of shortest significant ranges were derived using the formula:

$$R_k = \sqrt{((MS_e)n)D_{(\alpha;k,df_e)}}$$

where R_k = shortest significant range of k sums,

MS_e = mean-square error; in this case, since S is taken as the random effect, MS_{QS} is taken to be MS_e.

n = number of tests for each quantization method. (= 648)

$D_{(\alpha;k,df_e)}$ = the value from the table of significant studentized ranges for Duncan's multiple-range test (df_e = degrees of freedom of MS_e, i.e. of MS_{QS})

In Table III.4, the heavy horizontal bars under the columns of quantization schemes indicates the significance of the difference between methods. Where the difference between quantization schemes is not significant, the columns are underscored by the same heavy bar.

The results given in Table III.4 show that, although schemes 1, 4 and 5 scored higher than scheme 0 (normal PCM encoding), the difference between the three schemes and scheme 0 is not statistically significant. Also, it is not possible from the statistical evidence to make any firm quantitative statement as to the relative performance of schemes 0, 1, 2, 4 and 5. The results of Table III.4 do show, however, that quantization schemes 3 and 6 are superior to quantization scheme 0. Although it is difficult from the statistical evidence to make exact quantitative statements as to the superiority of schemes 3 and 6, the scores shown in Table III.2 indicate that the two schemes have a superiority of at least 1-bit/sample compared with normal PCM encoding.

III.4. Discussions

The investigations reported on in this chapter, have shown that if a pseudorandomly dithered quantization is performed in such a way that the zero-crossings pattern of the speech is preserved, then, at low bit-rates, it is possible for the intelligibility of the quantized speech to be improved compared with that of the correspondingly low-bit-rate PCM quantized speech; and the improvement can be of the order of 1 bit per sample. The two zero-crossing preservation schemes that show this improvement are the schemes numbered 3 and 6 that were described in the chapter. These schemes are virtually as easy to implement as the normal dithered-quantization method. Scheme 3 is particularly important, from the point of view of bandwidth compression, since it is of the type in which the receiver is not informed of the zero-crossing preservations. This means that no extra bandwidth is necessary, other than that required for the transmission of information about the quantized sample values, and therefore, the quoted advantage over PCM encoding of 1 bit per sample is a true advantage. The advantage quoted for scheme 6 has to be reduced slightly on account of the small additional amount of information that has to be transmitted to notify the receiver of the zero-crossing preservations.

It seems likely that scheme 3 or scheme 6, when coupled with a suitable quantization step-size adaptation strategy could form the basis of a simple, reliable and economical low-bit-rate speech digitizer. The quality of speech reproduction from such a low-bit-rate digitizer (2 - 4 bits/sample) would, of course, not be of commercial-telephony quality. However, its intelligibility would remain tolerably high and because of the dithering, the subjective acceptability would also be notably better. Thus, there are areas of application, such as those where the quality requirements are not too stringent, where the technique would be valuable on account of its savings in transmission bandwidth and low equipment cost.

An important question that arises from the series of subjective tests, is the question of why, intelligibility wise, schemes 3 and 6 should be significantly better than schemes 2, 4 and 5 and, why any of the methods of dithering should be better than fixed-level PCM encoding. An analysis of the quantization-error variance of the dithering schemes (see Appendix) shows that, with schemes 2, 3 and 4, the error variance is slightly increased compared with normal PCM encoding, whereas, with scheme 5, the error variance is approximately the same as with PCM encoding and with scheme 6, the error variance is slightly less than with PCM encoding. This might explain the superior behaviour of scheme 6 but it throws no light on the question of the superiority of scheme 3. A possible answer might be found from a consideration of the way in which the dithering noise sequences in the encoder and the decoder are modified from a true pseudorandom sequence and the nature of the differences that exist between the encoder and decoder sequences for category 1 schemes. Another factor that might have a bearing in the issue is the way in which the resulting quantization errors are dependent on the input signal. Although investigations giving insight into the above aspects may be helpful in explaining the relative performance of the various quantization schemes, it is felt that the crux of the problem lies in our present very incomplete state of knowledge of the human auditory perception system, especially regarding what makes one type of quantization error more harmful to intelligibility than another. It is almost certain that a greater understanding of the human speech processing system would lead to the creation of amplitude dithering schemes of better performance than those investigated, but this is not a subject to be considered further in this thesis. An alternative approach to the question of the low-bit-rate encoding of speech will be adopted. In the following chapter, attention will be diverted to the possibility of bandwidth compression by the use of redundancy extracting coding networks.

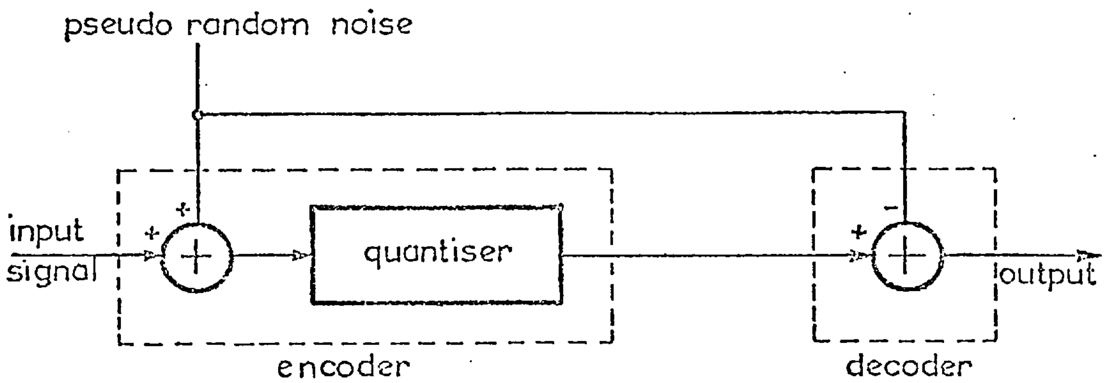


Fig. III.1. A normal amplitude dithered quantization scheme

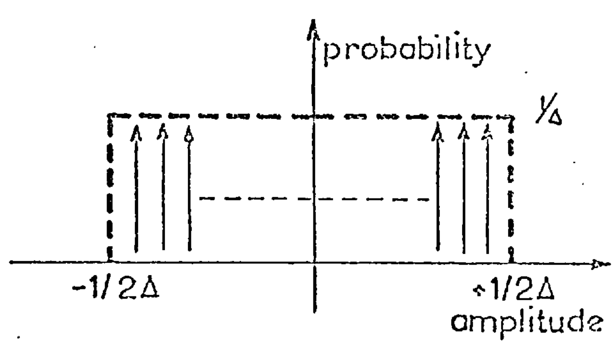
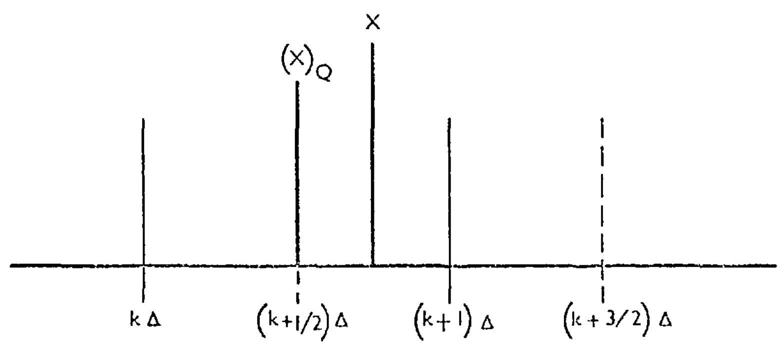
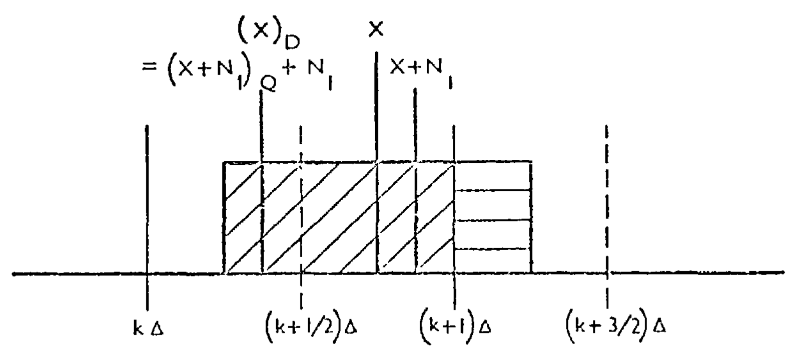


Fig. III.2. Probability distribution of pseudo-random noise



a) Normal fixed-level quantization



b) Normal dithered quantization

Fig. III.3.

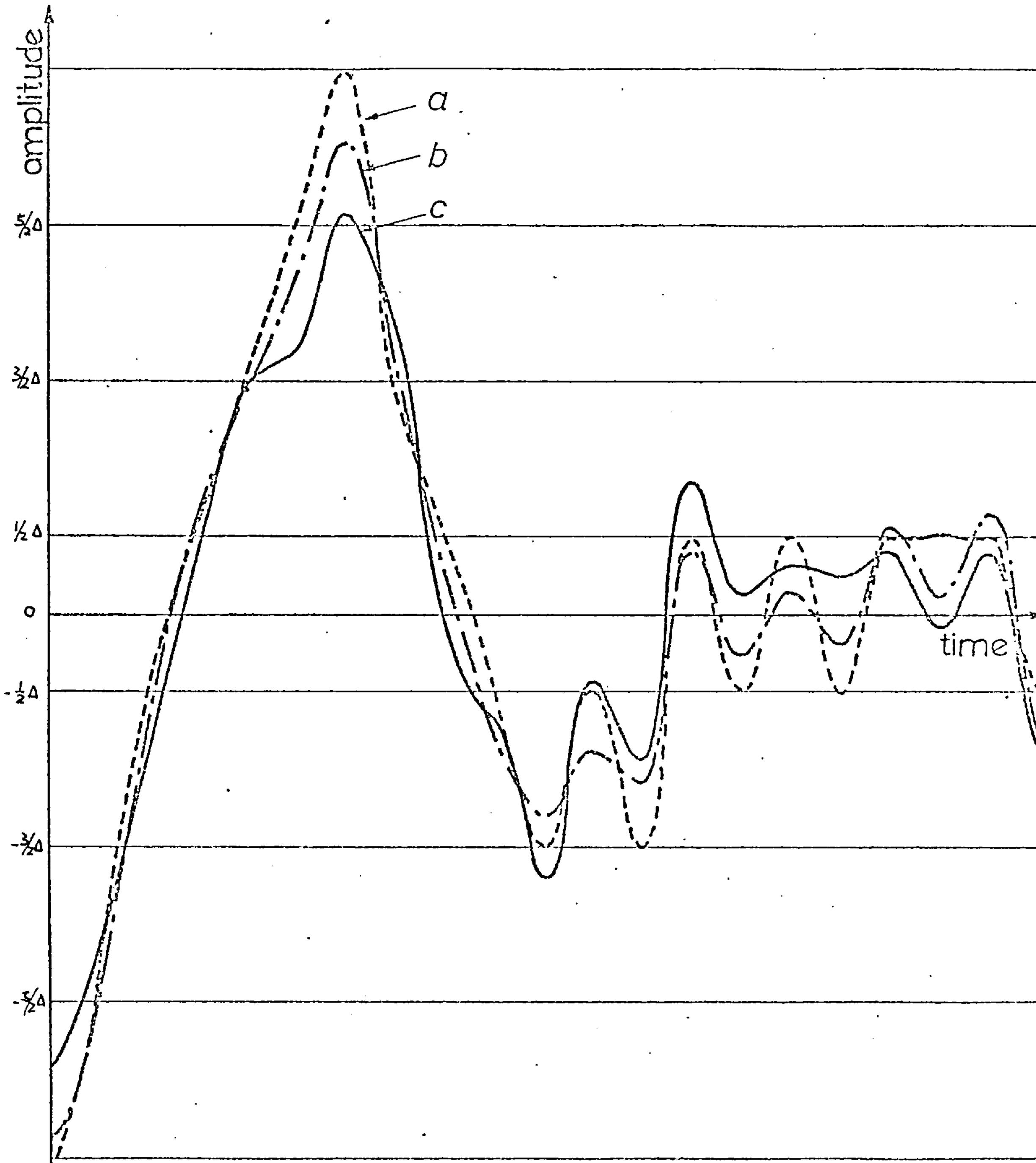


Fig. III.4. Illustration of effect of pseudo-random dithering on zero crossings

a) 3 Bits/sample PCM quantization

b) 8 Bits/sample speech waveform

c) 3 Bits/sample pseudo-randomly dithered quantization

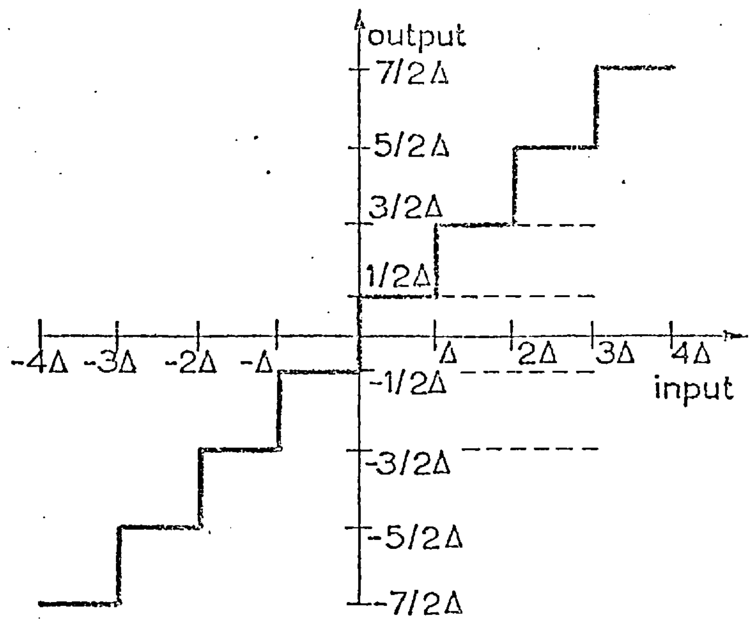


Fig. III.5. Quantization without dead zone

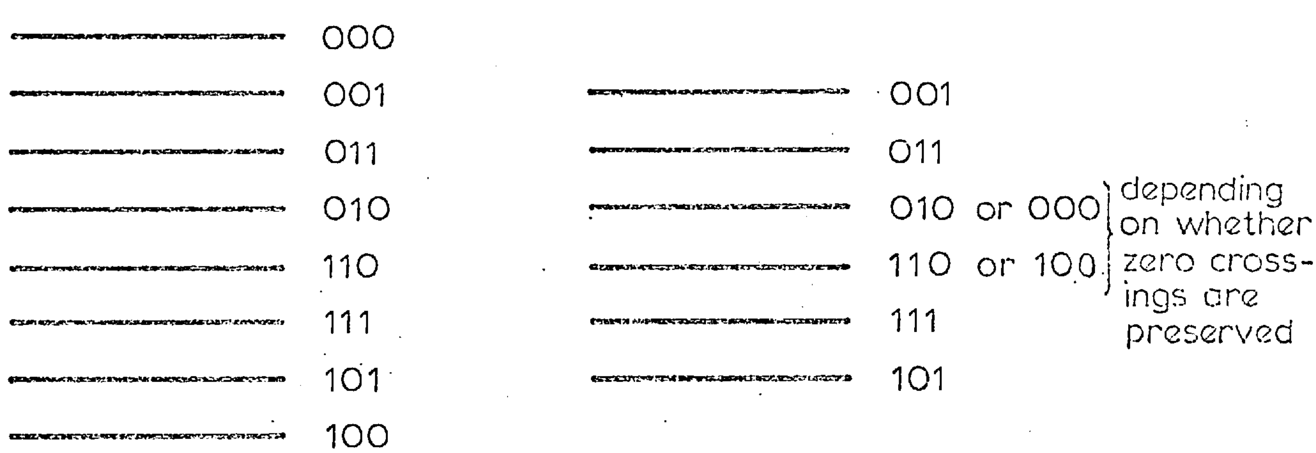


Fig. III.6. Scheme for informing the receiver of preserved zero crossings

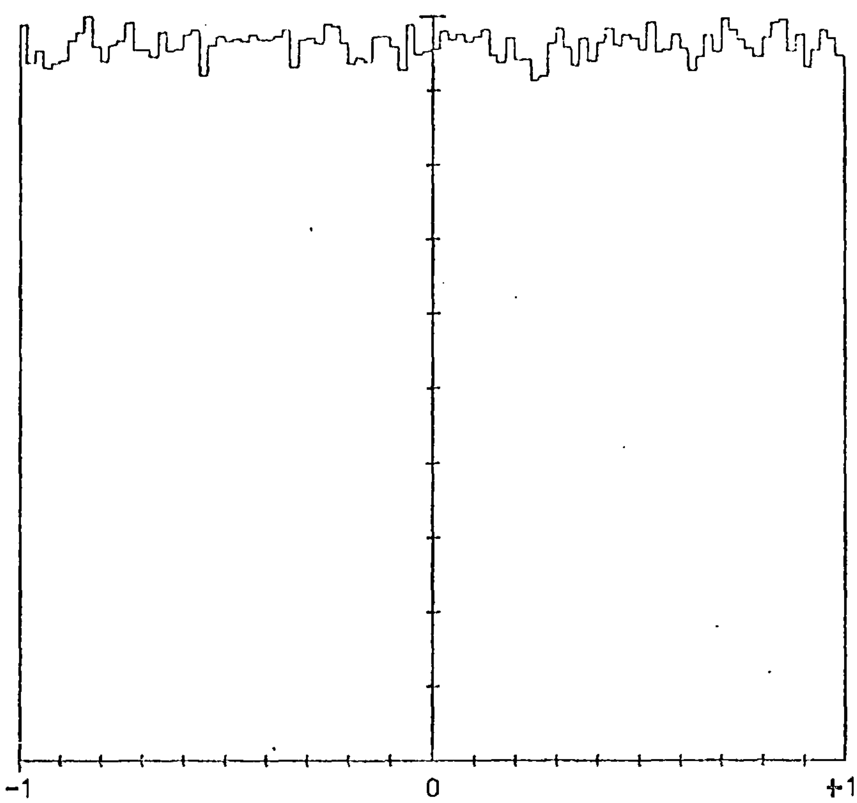


Fig. III.7. Histogram of the pseudo-random generator output

Code	Method	Preservation of Zero Crossings	Identical Dithering Noise in Encoder and Decoder
0	Straight PCM	yes	Not Applicable
1	Normal Dithered PCM	no	yes
2	Category 1, Scheme a:- No freezing of dither at decoder	yes	no
3	Category 1, Scheme b:- Selective freezing of dither at first positive or first negative quantization levels	yes	no
4	Category 1, Scheme c:- Freezing all dither at first positive or first negative quantization levels	yes	no
5	Category 2, Scheme d:- Dither frozen at false zero crossings	yes	yes
6	Category 2, Scheme e:- Dither inverted at false zero crossings	yes	yes

Table III.1. Quantization schemes tested

Bit Rate (Bits/sample)	Quantization Method							Sum
	0	1	2	3	4	5	6	
3	167	165	158	179	156	172	180	1177
4	174	176	173	188	186	177	188	1262
5	178	191	182	189	184	180	190	1294
Sum	519	532	513	556	526	529	558	

Table III.2. Test scores of quantization schemes

Source of Variation	Degrees of Freedom	Mean Squares	F Ratio	Significance Level
Q	6	.4669	MS_Q/MS_{QS} = 3.9159	$>$ 0.995
B	2	2.4182	MS_B/MS_{BS} = 17.2200	$\gg$ 0.995
P	5	1.1037	MS_P/MS_{PS} = 8.7734	$\gg$ 0.995
S	35	.5319	MS_S/MS_{QBPS} = 3.6865	$\gg$ 0.995
QB	12	.1404	MS_{QB}/MS_{QBS} = 0.9807	N.S.*
QP	30	.1152	MS_{QP}/MS_{QPS} = 0.8480	N.S.
QS	210	.1192	MS_{QS}/MS_{QBPS} = 0.8265	N.S.
BP	10	.0844	MS_{BP}/MS_{BPS} = 0.5725	N.S.
BS	70	.1413	MS_{BS}/MS_{QBPS} = 0.9796	N.S.
PS	175	.1264	MS_{PS}/MS_{QBPS} = 0.8759	N.S.
QBP	60	.1452	MS_{QBP}/MS_{QBPS} = 1.0060	$\ll$ 0.75
QBS	420	.1432	MS_{QBS}/MS_{QBPS} = 0.9925	N.S.
QPS	1050	.1358	MS_{QPS}/MS_{QBPS} = 0.8759	N.S.
BPS	350	.1473	MS_{BPS}/MS_{QBPS} = 1.02121	$<$ 0.75
QBPS	2100	.1443		

*N.S. - Not Significant

Table III.3. Analysis of variance of subjective-test data

Q	6	3	1	5	4	0	2	Shortest Significant Ranges
Sums (S_Q)	558	556	532	529	526	519	513	α = 0.05
$S_6 - S_Q$		2	26	29	32	39	45	R_2 = 24.4
$S_3 - S_Q$			24	27	30	37	43	R_3 = 25.6
$S_1 - S_Q$				3	6	13	19	R_4 = 26.5
$S_5 - S_Q$					3	10	16	R_5 = 27.2
$S_4 - S_Q$						7	3	R_6 = 27.7
$S_0 - S_Q$							6	R_7 = 28.1
	6	3	1	5	4	0	2	

<u>Table III.4.</u>

Ducans' multiple-range test:-

Quantization schemes underscored by the same heavy bar

are <u>not</u> significantly different.

CHAPTER IV

Residual Encoding with a One-Bit Quantizer

The method of adaptive predictive differential encoding* is known to be very attractive for the efficient digitization of speech signal. Atal and Schroeder[29], using a differential encoding scheme that exploits the redundancies in speech waves, have achieved a remarkable SNR advantage over PCM. In their scheme, the coefficients of the short-term adaptive predictor are calculated by the inversion of an auto-correlation matrix and the predictor coefficients and quantizer step size are transmitted separately to the receiver. If the number of predictor stages is large, an iterative adaptation procedure[32] is often less complex, and can be computationally more efficient in calculating the values of the predictor coefficients. In a recent development[34], it has been shown that it is not necessary to transmit separately information about the quantizer step size and the predictor coefficients, since, by using the so-called "residual encoding" scheme, these parameters can be determined synchronously from the transmitted information bit stream. A disadvantage of the residual encoding scheme proposed by Gibson[34] is that the algorithm for the quantizer step-size adjustment requires the use of a multi-bit (greater than 2 levels) quantizer and this thus imposes a lower limit on the minimum number of bits that are required per sample. One method to overcome this and reduce the number of bits appearing at the quantizer output of a residual encoder

---

* See also Chapter I, Section I.1.A.3.

would be to use a variable-input-length to fixed-output-length code similar to that used by Cohn and Melsa(36). This technique does, however, suffer in that it requires the use of buffer storage, with its attendant disadvantages.

In this chapter, the possibility of using a one-bit (two-level) quantizer as a means of reducing the bit rate required in a residual encoding scheme is examined and an algorithm is developed so that the step size of the one-bit quantizer can be adapted from the output bit stream. Additionally, the structure of the predictor part of the residual encoder is considered. In the works mentioned previously in this chapter, the predictor network used is a variant of the so-called "direct-form" of vocoder filter structure used by Atal and Hanauer(63) and by Markel(62). In this chapter, a new form (a lattice-type structure) which is suitable for use in residual encoding, is derived. The new predictor structure is developed from the lattice-type vocoder filter structure* of Itakura(65,91). The performance of the direct-form and the newly developed lattice-form of predictor structure are compared when using the same one-bit quantizer, and it is shown that the lattice-form of predictor has a number of advantages as compared with the direct-form of predictor and that it is particularly suitable for use in low bit rate (less than two bits per Shannon sample) applications.

This chapter is structured in the following way:

In Section IV.1, the question of the predictor structure is first considered and a lattice-type of predictor is developed that is suitable for use in a differential (and residual) encoder. Then, in Section IV.2, the problem of adjusting the step size in a one-bit

---

* See also Chapter I, Section I.1.B.5.

quantizer is examined and a strategy is developed which makes it possible to synchronously adjust the quantizer step size from information contained in the output bit stream. In Section IV.3, the question of parameter optimization is considered and a set of parameter adaptation algorithms is derived. If a special case of a residual one-bit encoder is considered, where the predictor order is 1 and predictor coefficient is fixed, it is found that the step size adaptation algorithm is reduced to that of an adaptive delta modulator. This relation between one-bit residual encoding and adaptive delta modulation is described in Section IV.4. A discussion of the relative theoretical merit of the residual encoders using the two types of predictor structures is given in Section IV.5. In the actual implementation of these new coders it was discovered that some modifications, to the parameter adaptation algorithms derived in Section IV.3, could result in improvements in their performance. These are described in Section IV.6. In Section IV.7, experimental results are presented relating to the performance of the direct-form and lattice-form of predictors when using a one-bit quantizer and employing real speech as input signal.

IV.1 Predictor Structure

It has been shown in Chapter I, Section A.2 that with a typical differential encoding and decoding scheme as in Fig. I.2, the SNR of the received signal, $\hat{S}$, can be maximized by minimizing the prediction error, E, i.e. by minimizing the output from the comparator comparing the input, S, and the predicted signal, P. The well-known vocoder filter structures of Atal and Hanuaer(63) and Itakura and Saito(65) are very suitable as a means of removing the redundancies from speech signals, and thereby reducing the energy contained in the signal. Predictor networks derived from these vocoder filter structures provided an effective way of p edicting speech so that the prediction error is reduced and the output SNR thereby improved.

In deriving the lattice-form of predictor, it is helpful to begin by reconsidering the conventional direct-form of predictor.

A. The direct-form predictor:-

The vocoder filter used by Atal and Hanuaer(63) and by Markel(62) is a direct inverse of the all-pole vocal track model (Eq.(I.2)) and is thus an nth-order all-zero filter of the form,

$$H(Z) = 1 - \sum_{1}^{n} a_i Z^{-i} \quad . \qquad \text{---- (IV.1)}$$

If the input signal is of stationary statistics, then a set of coefficient values a_i, $i = 1,\text{-----},n$ can be calculated so as to minimize the energy of the filter output. This minimal energy output from the vocoder filter is exactly what is required as input to the quantizer in a differential encoding scheme. Therefore, the required error can be expressed as

$$E = S(1 - \sum_{1}^{n} a_i Z^{-i}) \quad ,$$

and it follows that in order to produce this prediction error E by subtraction of P from S, the predicted signal P has to be

$$P = S - E \qquad \text{---- (IV.2)}$$

$$= S \sum_{1}^{n} a_i Z^{-i}$$

This predicted signal can be expressed in time-domain form as

$$P = \sum_{1}^{n} a_i S_i$$

where $S_i = D^i S$ and D is a unit delay operator. In a differential encoding application, the received signal $\hat{S}$ is used instead of S as input to the predictor, thus

$$E = S - \sum_{1}^{n} a_i \hat{S}_i \qquad ---- \quad (IV.3)$$

and

$$P = \sum_{1}^{n} a_i \hat{S}_i$$

This is, of course, the well-known direct-form predictor.

In the differential encoder, the received signal $\hat{S}$ is constructed from the quantized prediction error $\hat{E}$. On replacing S and E in Eq.(IV.2) by $\hat{S}$ and $\hat{E}$ and rearranging, it follows that

$$\hat{S} = P + \hat{E}$$

and that the complete encoder circuit with the direct-form predictor can be drawn as in Fig. IV.1.

B. The lattice-form predictor:-

The lattice-form of the vocoder filter derived by Itakura and Saito[65] is reproduced in Fig. IV.2, together with its associated inverse filter. It is important to note that in the operation of this filter the redundancy in the input signal is removed successively by each of the cascaded stages of the filter. The coefficient b_i, at each stage i, is optimized so as to minimize the output E_{i+1} from that stage alone and in this way the final output E_{n+1}

is automatically of minimum energy[65].

Let us now attempt to construct a differential encoder from this vocoder filter so that the final output E_{n+1} is the input to the quantizer, that is, is the prediction error E. From Fig. IV.2a, it can be seen that

$$E_{n+1} = S - \sum_{1}^{n} b_i F_i$$

Hence, the predictor output P is given by

$$P = S - E_{n+1} \qquad \text{---- (IV.4)}$$

$$= \sum_{1}^{n} b_i F_i$$

In differential encoding, the predicted signal P should be derived only from information that is available to the receiver. Thus the values of F_i have to be replaced by the received versions $\hat{F}_i$ and hence

$$P = \sum_{1}^{n} b_i \hat{F}_i \qquad \text{---- (IV.5)}$$

The energy in the intermediate prediction error E_{i+1} at the output of each filter stage has to be minimized individually. This intermediate prediction error can be written down from Fig. IV.2.a, on replacing F_i by $\hat{F}_i$:

$$E_{i+1} = E_i - b_i \hat{F}_i \quad , \quad i = 1,2,\text{----},n, \qquad \text{---- (IV.6)}$$

where $E_1 = S$.

The received $\hat{F}_i$'s can be generated from the quantized version $\hat{E}_{n+1}$ of the final prediction error E_{n+1}, by using $\hat{E}_{n+1}$ as the input to the inverse vocoder filter shown in Fig. IV.2.b. Given the initial states of $\hat{F}_i$'s, new $\hat{F}_i$'s are found (see Fig. IV.2.b) by the following recursive relationships,

$$\hat{E}_i = \hat{E}_{i+1} + b_i \hat{F}_i$$

$$\hat{F}_{i+1} = D(\hat{F}_i - b_i \hat{E}_i)$$

$$\text{for } i = n, n-1, ----, 1. \qquad ---- \quad (IV.7)$$

and, $$\hat{F}_1 = D(\hat{E}_1) \qquad ---- \quad (IV.8)$$

The complete differential encoder using a lattice-form of predictor is shown in Fig. IV.3.

IV.2. A Strategy for Adjusting the Step-Size of a One-Bit Quantizer

It is convenient to represent a two-level quantizer of variable step size by a symmetrical clipper of unit amplitude followed by an amplifier of variable gain σ, as shown in Fig. IV.4. In this case,

$$\hat{E} = \sigma Q \quad . \qquad ---- \quad (IV.9)$$

Ideally, the gain σ should be such that the power of the quantization error N is a minimum. But, it is not possible to optimally adapt the step size of a two-level quantizer from a knowledge of the quantizer output $\hat{E}$ alone, because $\hat{E}$ contains no amplitude information. If, however, a two-level quantizer is to be used in a differential encoding application, then it can be argued that instead of attempting to reduce the power of the quantization error N, the step size can be adjusted so as to reduce the power of the prediction error E. Let us consider a linearized differential encoder as shown in Fig. IV.5, in which the quantizer is replaced by a linear model of proportional gain g whose value is such that the power of the difference between the actual quantizer output $\hat{E}$ and the linearized model output gE, i.e. $\overline{(\hat{E} - gE)^2}$, is a minimum. In this linearized differential encoder, the prediction error is given by

$$E = \frac{(1 - G)S}{1 - G(1 - g)} ,$$

where G is the transfer function of the predictor and (1 - G) is the all-zero filter H(Z) of Eq. (IV.1.)

If the input signal S is an all pole input then it can be seen that, provided $g \neq 1$, then for every pole in S cancelled by (1 - G), a new pole is produced by $\left\{1 - G(1 - g)\right\}$. This thus means that E remains correlated and this in turn implies that the power of E is only a minimum when g is equal to unity. A value of $g = 1$ means that the power of the quantization error N is a minimum. This is because

$$\begin{aligned}\overline{N^2} &= \overline{(\hat{E} - E)^2} \\ &= \overline{(\hat{E} - gE)^2} ,\end{aligned}$$

which is minimum by definition. In the case when there are zeros in S, a sub-optimal solution will result since, in the process of reducing $\overline{E^2}$, g may be so adjusted as to partially compensate for the energy contributed by the zeros. An adjustment of g away from unity causes an increase in the energy of E due to the $\left\{1 - G(1 - g)\right\}$ term, while, at the same time, possibly leading to a reduction in the contribution due to the zeros of S. The optimal value of g, for minimum energy of E, would be between the requirements of these two conflicting factors. In the case of speech signals, the energy contributed by its zeros is usually insignificant as compared with that contributed by its poles and hence the optimal value of g will be fairly close to unity for minimum $\overline{E^2}$. Under these conditions $\overline{N^2}$ will also be close to its minimum.

IV.3. Parameter Optimization

Before dealing with the question of parameter optimization it is important to note that as the prediction coefficients, and the quantizer step size, are to be adjusted identically at both the transmitter and the receiver, it is only necessary to consider the optimization at the transmitter (the encoder). The predictor coefficients and the quantizer step size are the parameters that have to be optimised automatically to take account of changes in the statistics of the input signal, and it is this problem which is now considered.

A. The direct-form Encoder:-

In the previous section, we saw that both the quantizer step size σ and the predictor coefficients a_i should be adapted so as to reduce the power of the prediction error E to a minimum. That is, the variance $\langle E^2 \rangle$ of E can be taken as the error criterion. That is,

$$\Gamma = \langle E^2 \rangle$$

One approach to the solution of the optimization problem is through the use of an optimal gradient method. From Eq. (IV.3), we have that

$$E = S - \sum_{1}^{n} a_i \hat{S}_i$$

Assuming that each coefficient a_i is independent of its previous values, then the S_i's are independent of the a_i's and we have

$$\frac{\partial E}{\partial a_i} = -\hat{S}_i \qquad , \qquad i = 1,2,----,n.$$

and, therefore,

$$\begin{aligned} \frac{\partial \langle E^2 \rangle}{\partial a_i} &= -2 \langle E\hat{S}_i \rangle \doteq -2 \langle \hat{E}\hat{S}_i \rangle \\ &= -2\hat{E}\hat{S}_i \end{aligned} \qquad ---- \quad (IV.10)$$

In the above expression, the quantized prediction error $\hat{E}$ is used to approximate the actual prediction error E, since E is unknown to the decoder and cannot, therefore, be used for parameter adaptation. Also, the produce $\hat{E}\hat{S}_i$ is used instead of the ensemble average $\langle \hat{E}\hat{S}_i \rangle$ because, at each sampling instant, $\hat{E}\hat{S}_i$ is the only information available about this ensemble average and is thus the closest estimate that can be made of its value.

To find the gradient of the error criterion with respect to the step size σ, we have from Fig. IV.1, on replacing $\hat{E}$ by σQ,

$$\hat{S}_1 = D(\sigma Q + \sum_{1}^{n} a_i \hat{S}_i)$$

where D is a unit delay operator,

and Q is the clipper output of unit amplitude.

And, therefore,

$$E = S - a_1(D(\sigma Q + \sum_1^n a_i\hat{S}_i)) - \sum_2^n a_i\hat{S}_i$$

From this expression, we see that E is independent of σ, if, as in the case of the predictor coefficients, the step size at the last iteration $D\sigma$ is assumed to be independent of the new step size σ. This difficulty can be avoided if the error criterion is changed to $\langle (D^{-1}E)^2 \rangle$, where D^{-1} is a unit advance operator. The minimization of $\langle (D^{-1}E)^2 \rangle$, with respect to σ, serves the purpose of reducing the time average of E^2 to a minimum, just as effectively as the minimization of $\langle E^2 \rangle$. The reason for this is easy to see. If the ensemble average at each sampling instant, be it $\langle E^2 \rangle$ or $\langle (D^{-1}E)^2 \rangle$, is reduced to a minimum, then the time average is a minimum as well.

It can be seen from the above that

$$D^{-1}E = D^{-1}S - (D^{-1}a_1)(\sigma Q + \sum_1^n a_i\hat{S}_i) - D^{-1}\sum_2^n a_i\hat{S}_i$$

and it follows that

$$\frac{\partial \langle (D^{-1}E)^2 \rangle}{\partial \sigma} = -2\langle (D^{-1}E)(D^{-1}a_1)Q \rangle \qquad \text{----} \quad \text{(IV.11)}$$

$$\doteq -2(D^{-1}\hat{E})(D^{-1}a_1)Q$$

Optimization of $\langle (D^{-1}E)^2 \rangle$ as an ensemble average must not be confused with the optimization of $\langle E^2 \rangle$ which is a different ensemble average. In other words, two error criteria are to be optimized simultaneously, namely:

(1) $\Gamma_a(a_1,a_2,----,a_n) = \langle E^2 \rangle$

(2) $\Gamma_\sigma(\sigma) = \langle (D^{-1}E)^2 \rangle$

There are obvious interactions between the two optimization processes. The interactions, although obvious, are extremely difficult to isolate, and rather than attempt this, it was decided to treat the minimization of these two error criterions as two independent optimization problems and to adjust carefully their relative rates of adaptation so as to obtain the best practical result.

With the problem thus formulated, it is a simple matter to establish the optimal distance along the vectors of steepest descent for the two optimizations. This could be done, for example, by using the differential method described in Aoki[92]. On doing this, the iterative algorithm for optimizing the coefficient values is found to be:-

$$a_i^{(k+1)} = a_i^{(k)} + r_a \frac{\hat{E}^{(k)}\hat{S}_i^{(k)}}{\mu_a^{(k)}} \quad , \quad i = 1,2,----,n. \quad ---- \text{(IV.12)}$$

where $X^{(k)}$ is the value of X at the kth iteration.

and $\mu_a = \sum_1^n \left\langle \left(\hat{S}_i^{(k)}\right)^2 \right\rangle$

On account of the fact that various approximations have to be made in arriving at Eq. (IV.10), the gradient estimate is rather noisy and a reduction factor r_a ($r_a \ll 1$) has been added in the above expression in order to slow down the rate of adaptation.

Similarly, the algorithm for the minimization of $\Gamma_\sigma = \langle (D^{-1}E)^2 \rangle$ with respect to σ is found to be:-

$$\sigma^{(k+1)} = \sigma^{(k)} + r_\sigma \frac{\hat{E}^{(k+1)} a_1^{(k+1)} Q^{(k)}}{\mu_\sigma^{(k+1)}}$$

where $\mu_\sigma^{(k)} = \left\langle (a_1^{(k)} Q^{(k-1)})^2 \right\rangle$

and r_σ is a reduction factor.

On subtituting σQ for $\hat{E}$ and rearranging, it is found that

$$\sigma^{(k+1)} = \frac{\sigma^{(k)}}{1 - r_\sigma \dfrac{Q^{(k+1)} a_1^{(k+1)} Q^{(k)}}{\mu_\sigma^{(k+1)}}}$$

As no knowledge of $Q^{(k+1)}$ is available at the kth iteration, the above can be more conveniently written as

$$\sigma^{(k)} = \frac{\sigma^{(k-1)}}{1 - r_\sigma \dfrac{Q^{(k)} a_1^{(k)} Q^{(k-1)}}{\mu_\sigma^{(k)}}} \qquad \text{----- (IV.13)}$$

which is obtained by substituting k-1 for k.

Good estimates of the ensemble averages μ_a and μ_σ, are provided by their time averages since these time averages are relatively independent of the adaptation process and vary mainly according to the statistics of the input speech, which change relatively slowly. In experimental work reported on in the following sections, exponential time averaging with a time constant of about 2msec was used for calculation of μ_a and μ_σ.

B. The Lattice-form encoder:-

As mentioned previously, with this form of prediction filter structure the prediction error power from each filter stage has to be reduced to a minimum with respect to the coefficient associated with that stage. If the filter is of order n, the adaptation of the predictor coefficients consists of n parallel single-dimensional optimization problems. From Eq. (IV.6), we have

$$E_{i+1} = E_i - b_i \hat{F}_i \quad , \qquad i = 1,2,----,n.$$

and hence

$$\frac{\partial \langle E_{i+1}^2 \rangle}{\partial \ b_i} = -2 \langle E_{i+1} \hat{F}_i \rangle$$

$$\doteq - 2\hat{E}_{i+1} \hat{F}_i$$

For step size adaptation, the error criterion $\langle (D^{-1}E)^2 \rangle$ is again used. In this case $E = E_{n+1}$, which is the final prediction error from the nth-order lattice filter. From Eqs. (IV.4) and IV.5), we have

$$D^{-1}E_{n+1} = D^{-1}S - D^{-1} \sum_{1}^{n} b_i \hat{F}_i$$

and from Eqs. (IV.7),(IV.8) and (IV.9),we have that

$$D^{-1}\hat{F}_1 = \hat{E}_1 = \sigma Q + \sum_{1}^{n} b_j \hat{F}_j$$

and that

$$D^{-1}\hat{F}_i = \hat{F}_{i-1} - b_{i-1} (\sigma Q + \sum_{i-1}^{n} b_j \hat{F}_j) \quad , \quad i = 2,3,----,n.$$

Therefore,

$$\frac{\partial (D^{-1}E_{n+1})}{\partial \quad \sigma} = - \left\{ (D^{-1}b_1)Q - \sum_{2}^{n} (D^{-1}b_i) b_{i-1} Q \right\}$$

and it follows that

$$\frac{\partial\langle(D^{-1}E_{n+1})^2\rangle}{\partial\sigma} = -2\langle(D^{-1}E_{n+1})\theta\rangle$$

$$\doteq -2(D^{-1}\hat{E}_{n+1})\theta$$

$$\text{where } \theta = -\frac{\partial(D^{-1}E_{n+1})}{\partial\sigma}$$

For an nth-order predictor, the following n+1 error criteria have to be optimized simultaneously:-

$$(1) \text{ ---- } (i) \text{ ---- } (n) \quad \Gamma_{b_i}(b_i) = \langle(E_{i+1})^2\rangle$$

$$(n+1) \qquad \Gamma_\sigma(\sigma) = \langle(D^{-1}E_{n+1})^2\rangle$$

If these are again considered to be independent then the following adaptation algorithms are obtained:

$$\sigma^{(k)} = \frac{\sigma^{(k-1)}}{1 - r_\sigma \dfrac{Q^{(k)}\theta^{(k-1)}}{\mu_\sigma^{(k-1)}}} \quad \text{----} \quad (IV.14)$$

$$\text{where } \theta^{(k)} = b_1^{(k+1)}Q^{(k)} - \sum_2^n b_i^{(k+1)}b_{i-1}^{(k)}Q^{(k)}$$

$$\text{and } \mu_\sigma^{(k)} = \langle(\theta^{(k)})^2\rangle$$

And

$$b_i^{(k+1)} = b_i^{(k)} + r_{b_i}\frac{\hat{E}_{i+1}^{(k)}\hat{F}_i^{(k)}}{\mu_{b_i}^{(k)}}, \quad i = 1,2,\text{-----},n. \quad (IV.15)$$

$$\text{where } \mu_{b_i}^{(k)} = \langle(\hat{F}_i^{(k)})^2\rangle, \quad i = 1,2,\text{-----},n.$$

In these expressions r_σ and r_{b_i} are reduction factors.

In addition to the interactions between the quantizer step size σ and the predictor coefficients b_i, it is easily seen that interactions also exist between the coefficient optimization processes. The coefficient b_i of stage i can reach its optimum, only after the coefficients b_1 ----b_{i-1} of the previous stages have all reached their optimum values. In the experimental work reported on in later sections a graded set of reduction factors $r_{b_i} = \frac{r_b}{i}$; $i = 1,2,\text{----},n$ were used and were found to work quite well. The use of graded reduction factors means that the rate of adaptation of the predictor coefficients gets progressively slower towards the final stage of the filter.

It can be seen that given a stationary input signal S, the $\hat{F}_i$'s, except $\hat{F}_1$, are non-stationary during the adaptation towards the optimal point. The power of $\hat{F}_i$ becomes gradually smaller as the coefficients of stages 1,2,-----,i-1 converge towards their respective optimal values. Thus, for i greater than 1, the normal time window of stationarity than can be assumed for speech signals is no longer valid for $\hat{F}_i$. In replacing the ensemble averages μ_{b_i} by their exponential time averages, a set of graded time constants were also used. With the time constant of $\overline{(\hat{F}_1)^2}$ as τ_1, the time constant τ_i of $\overline{(\hat{F}_i)^2}$ was set to be

$$\tau_i = \frac{\tau_1}{i} \quad , \qquad i = 2,3,\text{-----},n.$$

IV.4. Adaptive Delta Modulation

The encoders described can be reduced to a delta-modulator, by using a predictor of the first order and having its coefficient fixed at unity, i.e. $n = 1$ and a_1 or $b_1 = 1$.

From Eq. (IV.13) or (IV.14), the adaptation algorithm becomes:-

$$\sigma^{(k)} = \frac{\sigma^{(k-1)}}{1 - r_\sigma \dfrac{Q^{(k)}Q^{(k-1)}}{\langle (Q^{(k-1)})^2 \rangle}}$$

As the absolute value of Q is constant,

$$\langle (Q^{(k-1)})^2 \rangle = (Q^{(k-1)})^2$$

Therefore,

$$\sigma^k = \frac{\sigma^{(k-1)}}{1 - r_\sigma} \quad \text{if} \quad Q^{(k)} = Q^{(k-1)}$$

$$\sigma^k = \frac{\sigma^{(k-1)}}{1 + r_\sigma} \quad \text{if} \quad Q^{(k)} = - Q^{(k-1)}$$

with the constrain that $0 < r_\sigma < 1$

Thus, at the kth sampling instant, if the present output from the symmetrical clipper has the same polarity as the previous output the value of the step size multiplier, σ, is increased by a factor $1/(1-r_\sigma)$. If they are of different polarity, σ is decreased by a factor $1/(1+r_\sigma)$. This companding strategy is very similar to that employed by a number of adaptive delta modulators using instantaneous companding [18,20]. There is another class of adaptive delta modulators using syllabic companding [19,93], that also derive the companding information from the change (or no-change) of polarity in the output bit stream. In such schemes, a time average of the overload (i.e. no change in the polarity of adjacent output samples) and underload (i.e. change in polarity) information is used to control the quantizer step size. A very similar scheme would have been derived if, as illustrated overleaf, a time average had been used to estimate the ensemble average in the gradient expression (Eq. (IV.11)).

From Eq. (IV.11), and using $a_1 = 1$,

$$\frac{\partial\langle(D^{-1}E)^2\rangle}{\partial\ \sigma} = -\ 2\ \langle(D^{-1}R).Q\rangle$$

$$\doteq -\ 2\ \overline{(D^{-1}E).Q}$$

$$\doteq -\ 2\ \overline{(D^{-1}\sigma)\text{sign}((D^{-1}Q).Q)}\ \cdot\ |Q|^2$$

And assuming that σ changes very little, the above is

$$\doteq -\ 2\ \sigma\ \overline{\text{sign}\ ((D^{-1}Q).Q)}\ \cdot\ |Q|^2\ ;$$

and the step height adaptation algorithm becomes

$$\sigma^{(k+1)} = \sigma^{(k)}\ (1 + r_\sigma\ \overline{\text{sign}\ (Q^{(k+1)}Q^{(k)}}\ )$$

Thus the companding laws of these well known delta modulators can be interpreted as gradient optimization algorithms to minimize the power of the prediction error. The use of unity as the coefficient value for the first order predictor is quite justified. At the high sampling rate that such delta modulators operate, the low-pass filtered speech spectrum does appear to possess a strong pole at zero frequency. This might also explain why the companding laws break down at sampling frequencies approaching the Nyquist rate. At such sampling rates there is considerably more signal energy at frequencies near half the sampling frequency. The optimum value of the coefficient of the first order filter can be much less than unity or, even negative for certain high frequency fricative sounds, and the appropriate step size adaptation is thus quite different. Furthermore, a first-order predictor becomes inadquate for describing the spectrum shape and the argument in section IV.3 leading to the adoption of the prediction error power as the optimization criterion, does not hold well.

IV.5. A Discussion of the Relative Merits of the Direct and Lattice-Form of Encoders

A. Complexity:-

The encoder using the lattice-form of prediction filter is undoubtedly more complex than one using the direct-form of prediction filter. It not only uses twice the number of multipliers in the actual filter structure; but the adaptation algorithms for both the step size and predictor coefficients also involve more computations. However, as it will be shown later, the use of the lattice-form encoder, with its increased complexity, is not without its rewards.

B. Stability:-

Due to the non-linearity of the quantizer, and various approximations taken in developing the adaptation algorithms for the encoder paremeters; the adaptation algorithms, though seeking to minimize the prediction error variance, cannot guarantee the stability of the encoder. A few iterations in the wrong direction could be sufficient to trigger the whole encoder into unstability, especially when the reduction ratios had been adjusted to give the maximum convergence rate under more favourable conditions. With a lattice-form of prediction filter, at least the stability of the filter itself can be assured by keeping the coefficients within bounds of +1 and -1[65,67]. The checking of the stability of the direct-form filter is not such a simple matter. One way of checking the stability is to find the roots of the polynomial equation (Eq. (IV.4)) and keep them within the unit circle. A simplified method for doing this is given in Atal[63], but this method is still far too laborious for present purposes. In the simulations conducted, the encoder using the direct-form predictor was without a stability check.

C. Pole Sensitivity:-

It is generally acknowledged, in previous works[63] dealing with predictive coding of speech signals, where the predictor coefficients have to be transmitted, that the coefficients of the direct-form predictor have to be coded very accurately. A small deviation in these coefficients can cause a significant change in the pole positions of the received speech signal. The coefficients of the lattice-form predictor, on the other hand, can be coded more coarsely[67], for the same accuracy of pole positions. In the present application, though the predictor coefficient values do not have to be sent, steady-state or fluctuating errors from the optimum values can be resulted in the adapted coefficient values. Such errors can be caused by inaccuracies in the gradient estimation, non-linearity of the quantizer, interactions of concurrent optimizations or transmission errors. The resulting shifts in pole positions not only increase the noise power because of sub-optimal prediction, but also cause a perceptible deterioration of the quality of the regenerated speech. The lattice-form encoder, as it allows a wider tolerance in the coefficient values, has an advantage over the direct-form encoder in this respect.

D. Rate of Convergence:-

As demonatrated previously, each predictor coefficient, b_i, of the lattice-form encoder is adapted in one dimension. Assuming that the ensemble averages used to calculate the optimal gradients can be estimated exactly, theoretically the coefficient, b_i, can reach its optimum in one additional step after the coefficients of all its previous stages (b_1,----,b_{i-1}) had reached their optimum values. Or, the ith coefficient can reach its optimum in not more than i iterations. On the other hand, an infinite number of iterations would usually be required to reach the optimum for the multi-dimensional optimization of the predictor coefficients, a_i, ($i = 1,2,$----$,n$), of the direct-form encoder. The rate with which these coefficient values approach a certain proximity to the optimum point is dependent on the nature of the covariance matrix of the input signal. Other than in the unlikely case, where the matrix is orthogonal, the rate is much slower than ideally possible with the lattice-form predictor coefficients. Such is the idealized situation. In practice, the exact estimates of the ensemble averages cannot be obtained. The effect of the error in the estimation of the ensemble averages can be considered as noise added to the theoretical response, of the filter coefficients, to a change in the input speech spectrum. Such "noise" not only deviates the adaptation direction from that indicated by the optimal gradient, but also causes a fluctuation or "hunting" of the final coefficient values around the optimum. The magnitude of such fluctuatuon can be reduced by multiplying each iterative step by a reduction factor. This operation slows down the rate of convergence from that theoretically possible. If the same reduction factor is used for the two systems, their relative rates of convergence are still roughly in the same proportion as their "ideal" rates.If one system has to be slowed down appreciably more than the other for it to be useful, then the "usable" convergence rate is not only a function of the ideal rate, but is also a function of the reduction factor.

How much the convergence rate has to be slowed down depends on how much fluctuation can be tolerated, and on the magnitude of the fluctuation produced by the system. It has already been seen that the lattice-form predictor is less sensitive to the deviation of coefficient values from the optimum. However, the question remains as to whether its structure, and associated adaptation algorithms can give rise to larger fluctuations. Experimental evidences show that it does. To see whether this is compensated by the relative insensitivity of the system to such fluctuations, the following computer simulated experiment was performed, with, for convenience, the overall SNR being used as the performance measure. A white noise generator was used as a signal source, and used as input to an all-pole spectrum shaping filter of the lattice form (as Fig. IV.3.b) having fixed coefficient values of (b_1 = 0.7, $b_i = (-1)^{i-1} \cdot 0.1$, i = 2,3,----,n). The output of the filter was used as input to an encoder of the same order. The encoder could be of either the lattice- or direct-form. The initial values of the coefficients of the lattice-form encoder were set to the same values as the spectrum shaping filter. An equivalent set of coefficients a_i (i = 1,2,----,n) for the direct-form predictor were calculated using the recursive relation in Ref. (67) and these were used as the initial values for the direct-form encoder. The step size, σ, of the quantizer was set approximately to the final adapted value in an initial trial ran, and the total SNR figure was measured for a length of 512 sample, after allowing the parameters to settle down by running the set-up from initial condition for 1024 samples. It was found that, for the same reduction factor, the SNR is about the same irrespective of whether the lattice- or direct-form encoder is used. This implies that for the same resulting effect on the SNR caused by the coefficient fluctuations, the same reduction can be used; and, if the reduction factors used for the lattice- and direct-form encoders are the same, the convergence rates of the predictor coefficients of the two encoders should be in proportion to their ideal rates. This means that the coefficients of the lattice-form predictor converge much faster than that of the direct-form predictor. However, remembering the the actual reduction factor for the adaptation of each coefficient b_i of the lattice form is actually i times samller (i = 1,2,----,n), it can only safely be concluded that the coefficients of the first few stages of the lattice-form predictor actually converge faster than those of the direct-form predictor, while those of the last few stages may converge at an equal or a slower rate. As the greatest amount of redundancy in speech waves is extracted in the first few stages of a lattice-form predictor the convergence rates of the last few stages are of comparatively less importance to the overall performance of the encoder.

When using the direct-form encoder, one is faced with the dilema that if the order of the prediction filter is increased so as to model the input speech more accurately, the convergence rate becomes so slow that, when the statistics of the input speech change rapidly and last for only a short time (e.g. plosive sounds), the prediction breaks down completely; and this results in an extremely low SNR. Because of the rapid convergence of its first few coefficients, the lattice-form encoder copes with such situations much better and at least manages to provide a sub-optimal prediction.

IV.6. Some Practical Modifications

In the actual implementation of the encoders, there are a number of situations in which it is found that the optimization process fails. Such failures can be caused by transmission errors, instability, rapid transitions of the speech spectrum or sharp rises in the speech power, etc.. The consequences of a failure of the system to optimize are usually very severe, since incorrect prediction results and this leads to a disproportionately high increase in the noise power. In those cases in which the failures are rare and are of short duration, there is no serious deterioration of the overall SNR, but the perceptual effects resulting from the failures remain highly significant. In order to overcome these defects, a series of modifications have to be made to the encoders so that they can be used in practical situations. In carrying out the modification the emphasis is not so much on improving the SNR as on 'detuning' the optimization process so that a suboptimal set of parameters can be obtained under adverse conditions, even though this may result in slight deterioration of the SNR under more ideal conditions. The various causes of difficulty and methods for overcoming them are discussed below and a final modified set of parameter adaptation algorithms is given in Table IV.1.

A. Instability:-

The optimization algorithms, as they seek to reduce the power of the difference between the input and the predicted signal, tend basically towards a stable solution. However, instability can still result because the direction of the adaptation is not always correct on account of the errors in the gradient estimates. The greater the adaptation step, the larger is the effect of such an error. The adaptation step size could be reduced on a once-and-for-all basis by using a smaller reduction factor, or alternatively it could be reduced selectively in those cases where the step size is particularly large and therefore most likely to cause trouble. From the adaptation algorithms given in Eqs. (IV.12), (IV.13), (IV.14) and (IV.15), we see that the adapation step is large when the denominators (i.e. μ_a, μ_o, μ_c and μ_b respectively) approach zero, and that the step size can be reduced if a constant value is added to these denominators. In other words, the denominator μ in these equations is replaced by

$$\mu' = \mu + C \qquad \text{---- (IV.16)},$$

where C is an experimentally selected constant which may be different for each of the denominators.

It should be noted that instability is not a serious problem with an encoder using the lattice-form of predictor, since simply checking that the coefficient values lie within the bounds of +1 and -1 is sufficient to ensure stability of at least the predictor part of the encoder*. Although the lattice-form predictor is inherently stable, a constant C was still added, in the experimental work, to μ_b in order to avoid large numerical errors when μ_b becomes small.

B. Transmission errors:-

Errors in transmission can cause the encoder and decoder parameters to lose synchronism and the errors will propagate unless some means are provided to re-establish synchronism. We do this using the method suggested in Ref. (36). A clamping bias is applied to each predictor coefficient, that is, if the new coefficient value calculated by the adaptation algorithm is a it is replaced by

$$a' = a(1 - \delta) + K\delta ,$$

where K is the value of the clamping bias and δ is in the range of .005 to .01.

---

* See section IV.6.B.

The value a' is used as the 'old' coefficient value in the next iteration. During quiet passages in speech the denominators in the coefficient adaptation algorithms Eqs. (IV.12) and IV.15) become very small and the change in coefficient value due to this adaptation is very much reduced, because of the addition of the constant C to the denominator as in Eq. (IV.16). The clamping bias takes effect and gradually forces the coefficient value in both the encoder and decoder to the value of the clamping bias. The encoder and decoder are thus resynchronized after each quiet passage of sufficient duration. The quantizer step size generally does not have to be resynchronized, as a permanent change in this only alters the speech volume and does not affect its intelligibility or quality.

C. Rapid change of speech spectrum and speech power:-

In speech signals the situation is often encountered in which both the speech spectrum and speech power change simultaneously. As the quantizer step-size adaptation is highly dependent on the predictor coefficient values, incorrect predictor coefficient values can lead to incorrect quantizer step-size adjustment. Hence, the step-size adaptation cannot be done too quickly, if the actual step-size is to be prevented from becoming too small (or too large) before the predictor coefficients approach their optimal values. This limitation on the quantizer step-size adaptation rate, when coupled with the fact that during the adaptation the change in step-size is a fraction of the old step-size (see Eqs. (IV.13) and (IV.14)), severely limits the ability of the encoder and decoder to cope with a rapid rise in speech volume. This is not helped by the fact that a rise in volume is usually accompanied by a simultaneous change in spectrum shape, as occurs, for example, in the transition from a quiet passage to the beginning of a spoken passage. In this situation, the quantization step-size does not really start to increase until the predictor coefficients are close to their optimal values, and the situation is further aggravated by the fact that, under such circumstances, the predictor coefficients take longer than usual to reach their optimal values for two reasons: Firstly, the quantized prediction error $\hat{E}$ used as an approximation to the actual prediction error E in the coefficient adaptation algorithm is much smaller than E; and secondly, the constant added to the denominator also reduces the adaptation step when the signal level is low. In the case of a decline in speech volume such as occurs at the end of a spoken passage, it is easy to see that the encoder can cope much more effectively. The response of the quantizer step-size to changes in speech volume is thus seen to be contrary to good speech encoding practice which prefers a fast attack and slow decline.

It has been found possible to overcome the difficulty associated with the holding down of the quantizer step-size during upward changes in speech volume by applying to the step size some stimuli in the correct direction (i.e. upwards). The stimuli have to be applied automatically, and have therefore to be derived from information relating to the correct quantizer step-size. The relationship between the stimulus and information relating to the correct quantizer step-size does not have to be very exact. But it is imperative that the magnitude of the stimulus should not be limited by the quantizer step-size at the previous iteration. A quantity that appears to satisfy these specifications fairly well has been found and the method of deriving this quantity will now be considered.

Consider the circuit shown in Fig. IV.6, which is a simplified schematic drawing of a differential predictive encoder with a proportional gain λ added at the point between the feedback network and the comparator. The gradient of the prediction error variance $\langle E^2 \rangle$ with respect to this proportional gain λ can be found as follows:

$$E = S - \lambda P$$

thus,

$$\frac{\partial \langle E^2 \rangle}{\partial \lambda} = - \langle 2EP \rangle$$
$$\doteq - 2\hat{E}P$$

Assuming that λ is the only parameter to be optimized, the adaptative increment of λ to reduce the prediction error variance is found to be

$$\Delta\lambda = r_\lambda \frac{\hat{E}P}{\overline{P^2}} \quad ,$$

where r_λ is a reduction factor.

It is easy to see that λ is strongly related to the quantizer step-size. If λ needs to be increased, it means that $\hat{E}$ the quantizer output is consistently smaller than the prediction error E and vice versa. In the special case where the feedback network is time invariant, the proportional gain λ can be placed before the feedback network without it altering the circuit response and it has exactly the same effect as the quantizer step-size multiplier σ. The quantity $\Delta\lambda$ is not limited by the quantizer step-size, because if E is small, P will also be small thus the ratio EP to P^2 remains substantially unaffected.

In practice, a multiplier is, however, not actually used after a feedback network in the encoder structures, which means that λ is always equal to one. The quantity $\Delta\lambda$ is, however, added as a disturbance in the quantizer step-size adaptation algorithm (see Eqs. (IV.13) and (IV.14). For stability reasons, in the actual calculation of the value $\Delta\lambda$, a constant C_λ is added to the denominator. In order that this added constant shall not hold down the boost that $\Delta\lambda$ is supposed to give to the quantizer step-size during a rapid rise in signal level, two values C_λ^+ and C_λ^- are used selectively depending on whether the produce EP is positive or negative. The chosen relationship for C_λ^+ and C_λ^- was

$$C_\lambda^+ = C_\lambda^-/2$$

The effect of addition of $\Delta\lambda$ to the quantizer adaptation algorithm is demonstrated most vividly in Fig. IV.7. This figure, which was obtained when using real speech as input, shows the response of the quantizer step-size from a silent period to the beginning of an articulation. Fig. IV.7.a. is the response without the addition of $\Delta\lambda$. It can be seen that there is a delay before the quantizer step-size finally begins to increase. And, sometimes, a whole syllable is lost before this occurs. Fig. IV.7.b. shows the response with the addition of $\Delta\lambda$. Not only does the response rise more quickly, but it also does not suffer from a lengthy delay.

An additional advantage arising from the use of $\Delta\lambda$ is that when encoding voiced speech, the rise of quantizer step-size at the beginning of a pitch period is made much sharper. On account of this more realistic quantized prediction-error waveform, the SNR is actually improved with the addition of the disturbance $\Delta\lambda$. Under other situations, where the original quantizer adaptation algorithm would cope well, it was expected that some deterioration of performance would occur. However, it was found in the course of experiments that the only noticeable degradation was a slight increase in the idle-channel noise.

IV.7. Experimental Results

The SNR performance of the encoders described in this chapter was measured using real speech as input signal. The speech material consisted of five sentences taken from the Harvard Sentences, List No. 1, with each sentence spoken by a different speaker. The sentences used are listed in Table IV.2.

The encoders were simulated using a PDP 15 computer. The input speech was filtered using a 3,020 Hz charp-cut-off low-pass filter and then sampled and converted into digital signal using a 12-bit A/D convertor at various sampling rates selected by an external clock generator. These converted digital speech sentences were stored permanently on digital tapes. Any single sentence or sentences, at any desired sampling frequency, could thus be selected at will, from this pool of stored data and loaded into the computer disk store before processing*. During the processing, the speech signal was read from the disk and the processed signal was re-recorded and stored on another area on the disk. This meant that both the original signal, and the processed signal could be read from their respective disk areas and played back through the D/A convertor for subjective evaluation.

As a first step in assessment, the performance of the encoders were considered as a function of the number of predictor stages. The

---

* See also Chapter II.

sampling frequency was set arbitrarily at 8 KHz, and the clamping bias for the coefficient values were fixed at $K_1=0.7$ and $K_i=0$ $(i=2,3,----,n)$. The unfiltered SNR was measured for each test sentence. The overall average with respect to the five test sentences is shown in Fig. IV.8 as a function of the order n of the predictor, for both types of predictor structure. When n = 1 the two types of encoder are identical. However, with $n \geqslant 2$, the encoder with the lattice-form of predictor always has a superior performance. In connection with Fig. IV.8, a possibly significant observation is that the SNR of the lattice-type of encoder rises rapidly at first when the order n of the predictor is increased. After the inflexion point, at approximately n = 4, the rate of rise is very much reduced, but the increase in SNR is nevertheless monotonic with increased predictor order. On the other hand, the performance of the direct-type of encoder increases at a lower rate when the order n of the predictor is small and peaks when n is approximately equal to 7. It is believed that this peak is the trade-off point between the following two conflicting factors. A closer modelling of the speech signal is afforded by increasing the order of the predictor but the convergence rate of the multi-dimensional adaptation process of the coefficients of the direct-form predictor suffers as a consequence. In the case of the lattice-form of predictor, however, the predictor coefficients are adapted single dimensionally and the rate of convergence of each individual coefficient is not affected by the order of the filter. Hence, increasing the order of the predictor only serves to improve the SNR performance. However, inspection of Fig. IV.8 reveals that the SNR improvement resulting from increasing the order of the predictor beyond a certain point is not worthwhile.

In the evaluation of the encoders at various sampling rates, attention was focussed on low-bit-rate encoding in which less than 2 bits per Shannon sample were used. In the evaluation the clamping biases were set at the same values as before and the order of the predictors were chosen quite arbitrarily to be 4. The unfiltered average SNR's for the five test sentences are given in Table IV.3. In obtaining the values shown in Table IV.3, the following expression for SNR was used:

$$SNR = 10 \log \frac{\sum S_i^2}{\sum(\hat{S}_i - S_i)^2}$$

where S_i is the ith sample of the input signal and $\hat{S}_i$ is the ith sample of the unfiltered output signal.

Since, with encloders of the type being considered, the quantization error $(\hat{S}_i - S_i)$ is almost white, the use at the output of a sharp cut-off low-pass filter of bandwidth equal to the signal bandwidth should improve the SRN by a factor of approximately

$$\frac{\text{Sampling Rate}}{2 \times \text{Signal Bandwidth}}$$

These estimated SRN's are also shown in Table IV.3. In order to check the validity of the estimated SNR's, an attempt was made to measure the actual SNR at the filter output. This was done by storing $(\hat{S} - S)$ as processed signal, rather than S, playing this stored signal back through a D/A convertor and then filtering the output of the D/A convertor by a low-pass filter having a sharp cut-off at 3,020 Hz. The power of the signal at the filter output was measured using an RMS power meter of a long time constant.

The result was compared with the power of the similarly played back and filtered original digital speech. It was found that the SNR determined in this way was very close to the estimated SNR.

These SNR's are plotted in Fig. IV.9 as a function of the number of bits per Shannon sample. Also shown in the figure are the SNR's obtained by Gibson[34] for a multi-bit residual encoder and the SNR's obtained by Cohn[36] for a multi-bit residual encoder using additional compression. It should be noted that below approximately 1.5 bits per Nyquist sample, the SNR curves for the multi-bit encoders have been obtained by extrapolation. Also, it should be pointed out that many other factors such as the sample of speech material and the various time constants involved etc. could result in significant changes in the apparent SNR performance; and the multi-bit residual encoder curves should only be considered as indicative of the region of performance of the new encoders. The estimated SNR performances of the new encoders with the one-bit quantizer are seen to be between the two reference curves. The encoder using the direct-form of predictor appears, on average, to be approximately 0.7 dB inferior to the encoder using the lattice-form of predictor. The unfiltered SNR curves seem to be indicative of the way prediction is improved with over-sampling. They tend to flatten beyond the point of 2 bits per Shannon sample. It can be envisaged, thereafter, that the filtered SNR's of the one-bit residual encoders only improve linearly with the logarithm of the output bit rate, whilst those of the mulit-bit encoders continue to increase linearly with increasing output bit rate. This thus means that there is a point beyond which the performance of the one-bit encoders would become inferior even to that of the multi-bit encoder without further compression. Extrapolation from Fig. IV.9 indicates that this would not happen below 2.0 bits per Shannon sample.

The two performance curves for the new encoders tend to converge at higher sampling rates. This behaviour is hardly surprising because the use of higher sampling rates allows more time for the predictor coefficients to adapt to the optimal values. Therefore, the faster convergence rate of the coefficients of the lattice-form predictor makes little difference to its performance. Indeed, it was found in experiments that at sampling rates higher than 10 Ksamples/sec, it was difficult to distinguish the difference between the processed speech using these two types of predictor. However, the same could not be said of the subjective quality of the processed speech at bit rates below 8 Kbits/sec. At bit rates below 8 Kbits/sec, the perceptual difference appeared greater than that expected from the values of the SNR. The procssed speech using the direct-form predictor tended to be somewhat 'broken-up', with some short syllables disappearing completely. It was found that at low sampling rates, if, in an attempt to improve the adaptation rate of the encoder parameters, the reduction factors in the adaptation algorithms for the parameters were increased beyond the values that gave the lowest overall SNR, then the frequency with which syllables were lost was considerably reduced, but the encoder was then close to being transiently unstable and sometimes became so. Although the instability was often damped so that it was negligible after a short interval, the perceptual effect of this transient instability was a sudden loud burst of noise which was particularly irritating and was detrimental to the intelligibility of the reproduced speech.

The encoder using the lattice-form of predictor did not appear to suffer from these problems and it functioned well when operating with output bit rates as low as 7 Kbits/sec, and even lower.

Finally, it might be of interest to see, given an upper limit on the output bit ratio of, say, 7 Kbits/sec, and regardless of complexity and sparing little effort in searching for the optimal values of the various reduction factors and time constants, how good a performance could be obtained. Some tests were carried out and it was found that by using an 8th order lattice-form of predictor, an average unfiltered SNR of 9.38 dB could be obtained. This corresponds to an estimated filtered SNR of approximately 10 dB. A detailed breakdown of the SNR of each test sentence is given in Table IV.4 and an example of the original and reconstructed waveforms is shown in Fig. IV.10. The sound quality of the processed sentences could be described as smooth and clear.

IV.8. Discussions

It has been shown that it is possible to use a one-bit quantizer in a residual encoding scheme and achieve a reasonable quality of speech waveform transmission at bit rates very close to 1 bit per Shannon sample. The encoder with the lattice-form of predictor is especially suitable for such low-bit-rate applications. At 7 Kbits/sec, which is about 1.12 bits per Shannon sample, it achieves an average SNR of 9.22 dB with a 4th order predictor and about 10 db with an 8th order predictor. The adaptation algorithms may seem complicated at first sight. However, closer inspection reveals that many multiplication operations involve the multiplication between a variable and a constant value; and they can be replaced by much simpler shifting operations. It will also be noted that the algorithms are particularly suitable for hard-ware implementation. The parameter adaptation algorithms are independent of each other in the sense that the execution of one does not have to wait until the successful completion of another. Thus, the adaptation of the encoder parameters can be performed in parallel by using hardware circuits. This parallel operation very much simplifies the actual construction of the encoder as the circuits for predictor-coefficient adaptations will be the exact duplicates of one another. Furthermore, as each parameter adaptation has relatively few operations to perform, special high speed electronic devices will not be required for real-time realization, even with a predictor of very large order. Thus, on the whole, the one-bit residual encoder with the lattice type predictor appears to meet the requirements set forth in Section I.2. Admittedly, its complexity, although simple by the standards of adapative predictive encoders, is still somewhat forbidding when compared with even the most complex of fixed-predictor differential-encoding methods.

In comparing the lattice-form of predictor with the direct-form of predictor, it should be noted that, although the former is more complex than the latter, and more complex parameter adaptation algorithms have to be used, the convergence rate of its coefficients is faster than in the case of the direct-form predictor and stability can be maintained simply by keeping the coefficients within bounds of $\pm$ 1. Furhtermore, the fact that the positions of the poles of the processed signal do not appear to be as sensitive to small deviations of the coefficient values, also contributes towards the greater naturalness of speech reproduction. Below about a bit-rate of 1.4 bits per Shannon sample, the encoder with the direct-form of predictor has a tendency to miss some short speech syllables. Thus its use is not recommended at these bit-rates.

Between 1.4 and 1.8 bits per Shannon sample, the lattice-form of predictor still appears to have considerable advantages as compared with the direct-form of predictor. It can be seen from Fig. IV.8 that with values of n as low as 4, the performance of the lattice-form of encoder is already superior to that of the direct-form of any order. The problem of higher complexity is not therefore so strongly biased against the lattice-form of encoder and the lattice-form of encoder is still favoured on account of its better sound quality. At approximately 1.8 bits per Shannon sample, differences between the one-bit encoders are insignificant, both perceptually and in terms of SNR. This means that it is clearly advantageous to use the simpler direct-form predictor at such bit rates.

The comparison of the one-bit residual encoders with the multi-bit residual encoders is also of interest. It has been noted from Fig. IV.9. that the SNR performances of the two one-bit encoders are generally better than that of the multi-bit encoder without further compression on the output bit-stream and would be preferred to the latter not only because of the higher performance but also because of the simpler quantizer and the slightly easier adaptation procedure*. However, it should also be kept in mind that, above about 2.1 bits per Shannon sample, on account of the strong tapering of their SNR curves, the one-bit encoders would have lost their SNR advantage and would, in fact, be much inferior at even higher bit rates.

The multi-bit residual encoder with output bit-stream compression, on the other hand, is always superior to the one-bit encoders; but

---

* The one-bit quantizer output is only a polarity signal and a number of the multiplications can thus be replaced by a simple "straight-through or inversion" decision circuit.

it requires the additional expense of a large buffer store, its associated management circuitry and, of course, the code conversion circuit for the actual compression. This bit-stream-compression method has been successfully applied[36] for compressing the data rate of a 5-level quantizer output to a rate approximately equivalent to that of a 3-level quantizer output (i.e. from about 2.3 bits per Shannon sample to 1.5 bits per Shannon sample), with some sacrifice in performance. However, it is doubtful whether this method would be effective in achieving the same efficiency of compression when applied to the output of a 3-level quantizer, for the simple reason that the much reduced redundancies (as compared with the 5-level quantizer case) inherent in the output levels of the 3-level quantizer would be much more difficult to remove. Hence, it is felt that this method is again more suitable for higher bit-rates. Below 1.4 bits per Shannon sample, it appears that only the one-bit residual encoder with the lattice-form predictor could work satisfactorily.

Nontheless, the successful extraction of further redundancy from the output bit-stream of a residual encoder with a 5 (or more) - level quantizer indicates that the predictor network has not extracted all the redundancies in the input speech wave as it is meant to do, given a predictor of sufficient order. This thus suggests that there are some factors which prevent the predictor from performing at maximum efficiency. A close inspection of the detailed operation of a residual encoder has revealed one of the factors contributing to the reduced efficiency. This, and a method for overcoming the difficulty are discussed in the chapter which follows.

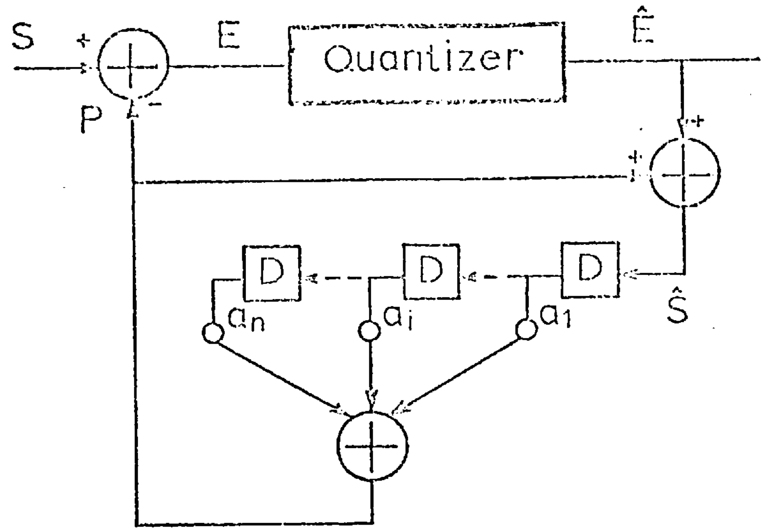
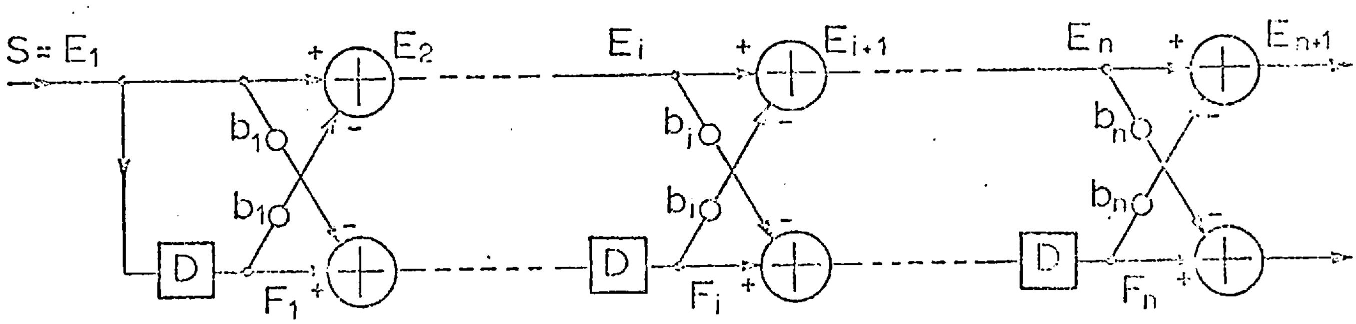
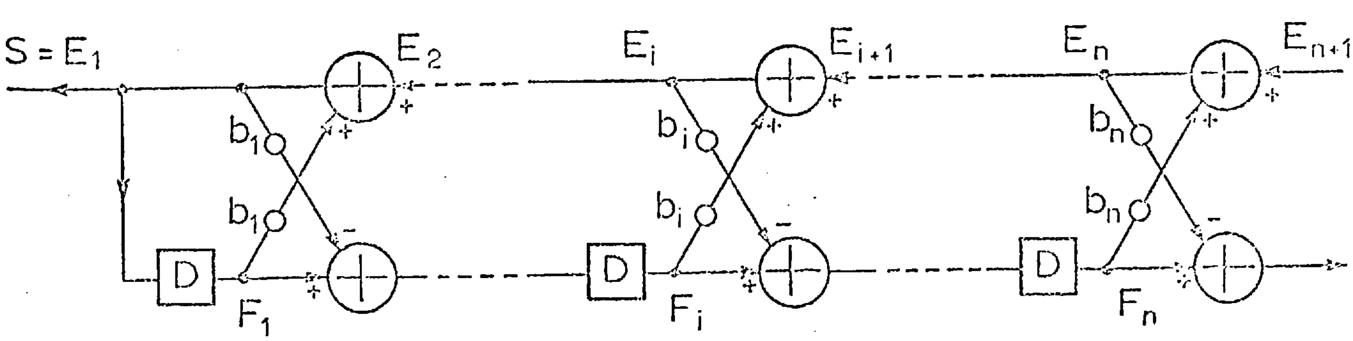


Fig. IV.1. Encoder using the direct-form predictor



a) The maximum-likelihood vocoder filter



b) The inverse filter of the above

Fig. IV.2.

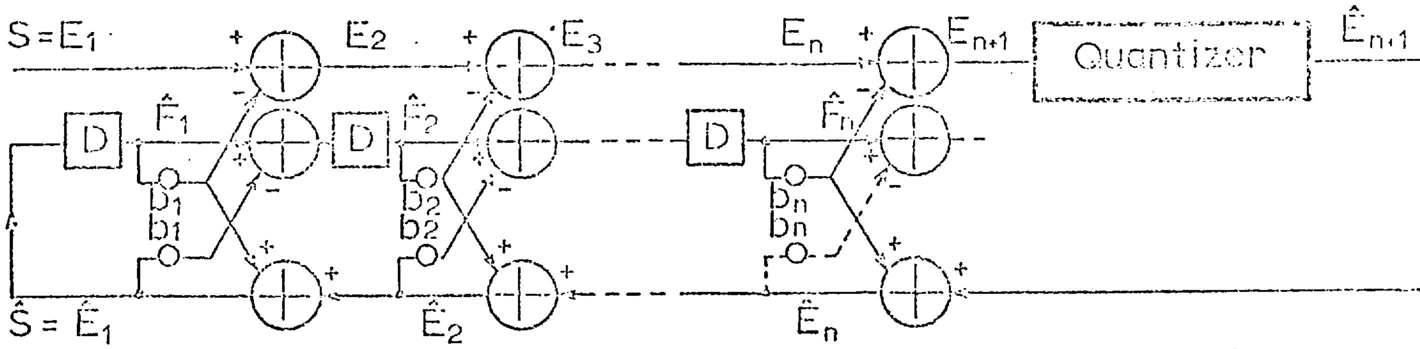


Fig. IV.3. Encoder using the lattice-form predictor

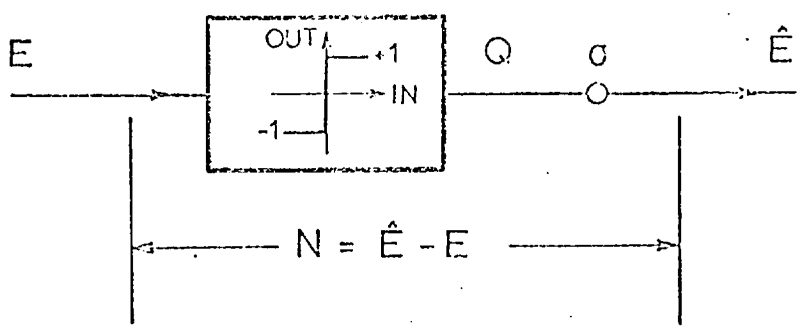


Fig. IV.4. A two-level quantizer with step size σ

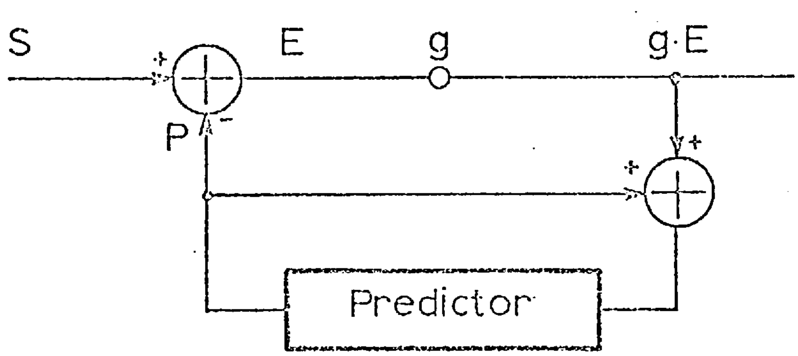


Fig. IV.5. Linearized model of the differential encoder

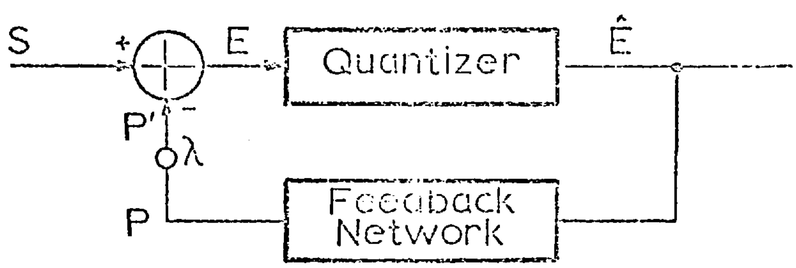
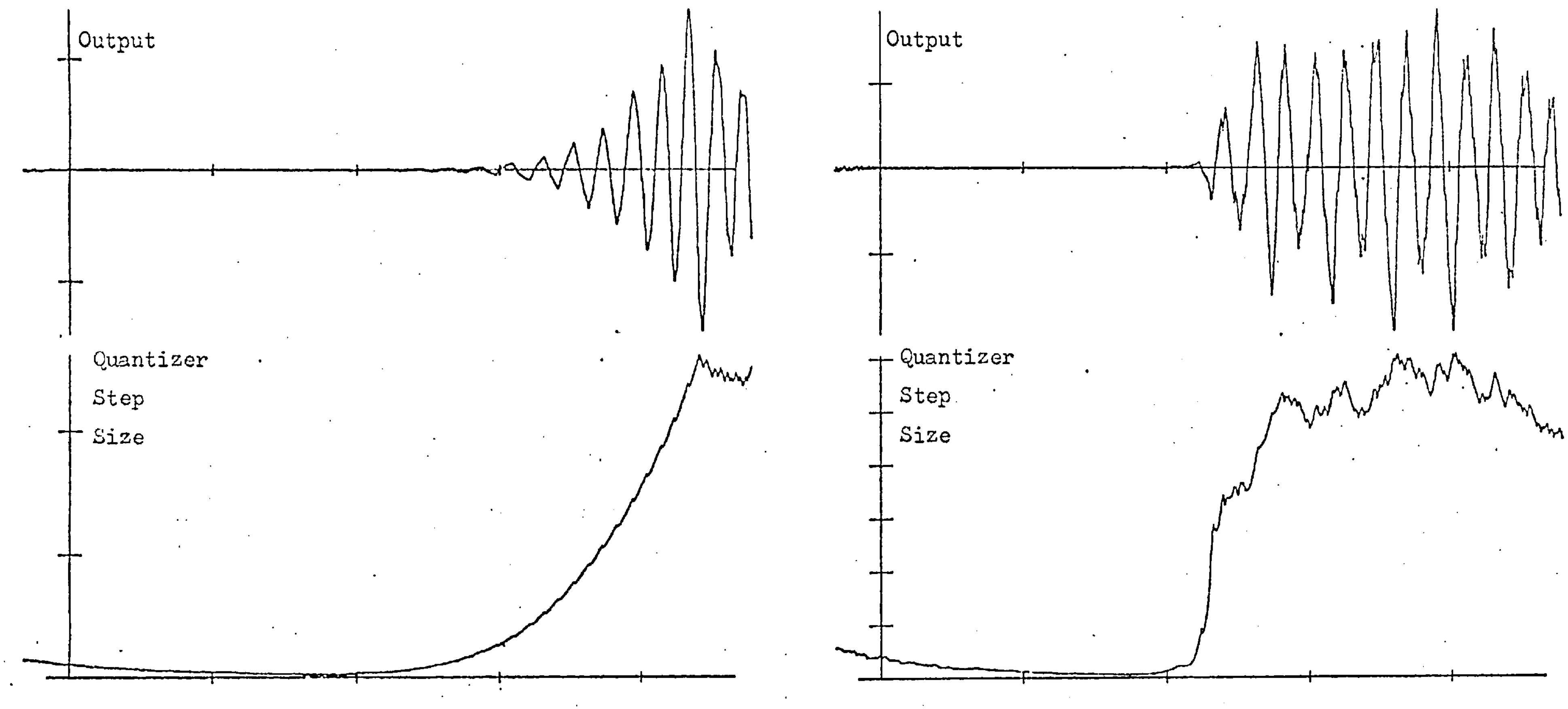


Fig. IV.6. A differential encoder incorperating a proportional gain λ after the feedback network



a) Without addition of Δλ

b) With addition of Δλ

<u>Fig. IV.7.</u> Effect of addition of Δλ to the quantizer step size

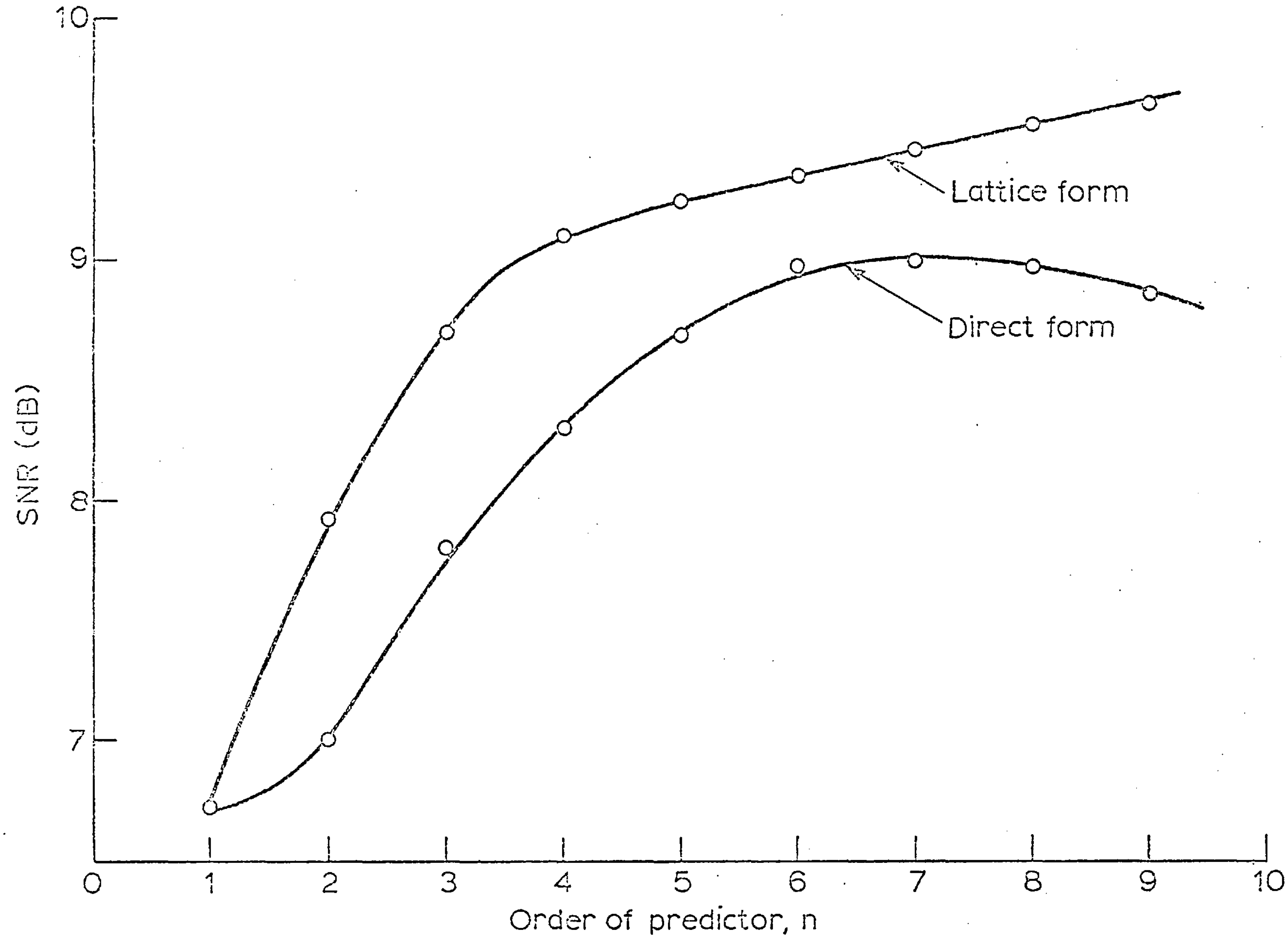


Fig. IV.8. Performance of the encoders versus predictor order
(Sampling frequency = 8 K samples/sec)

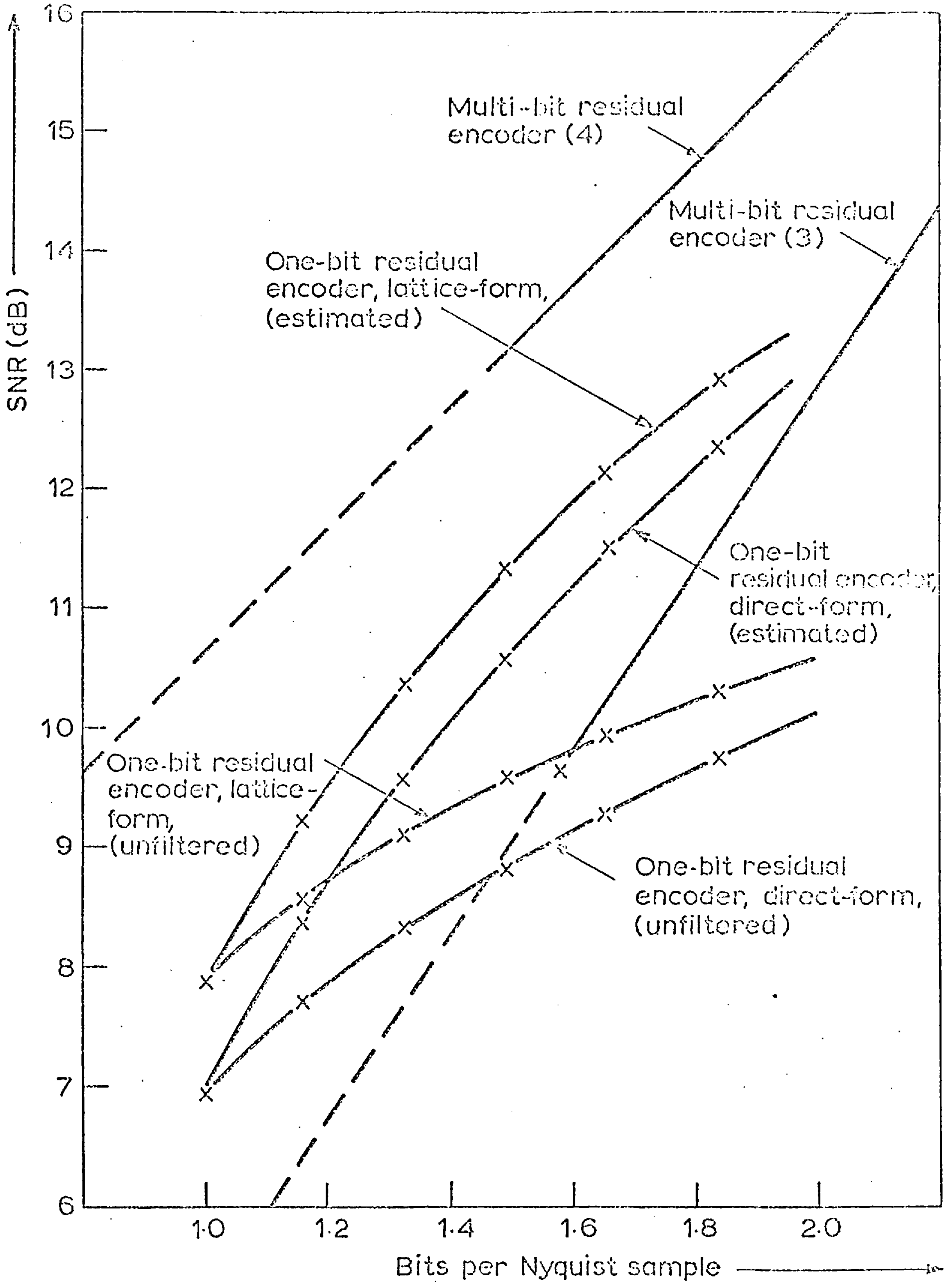
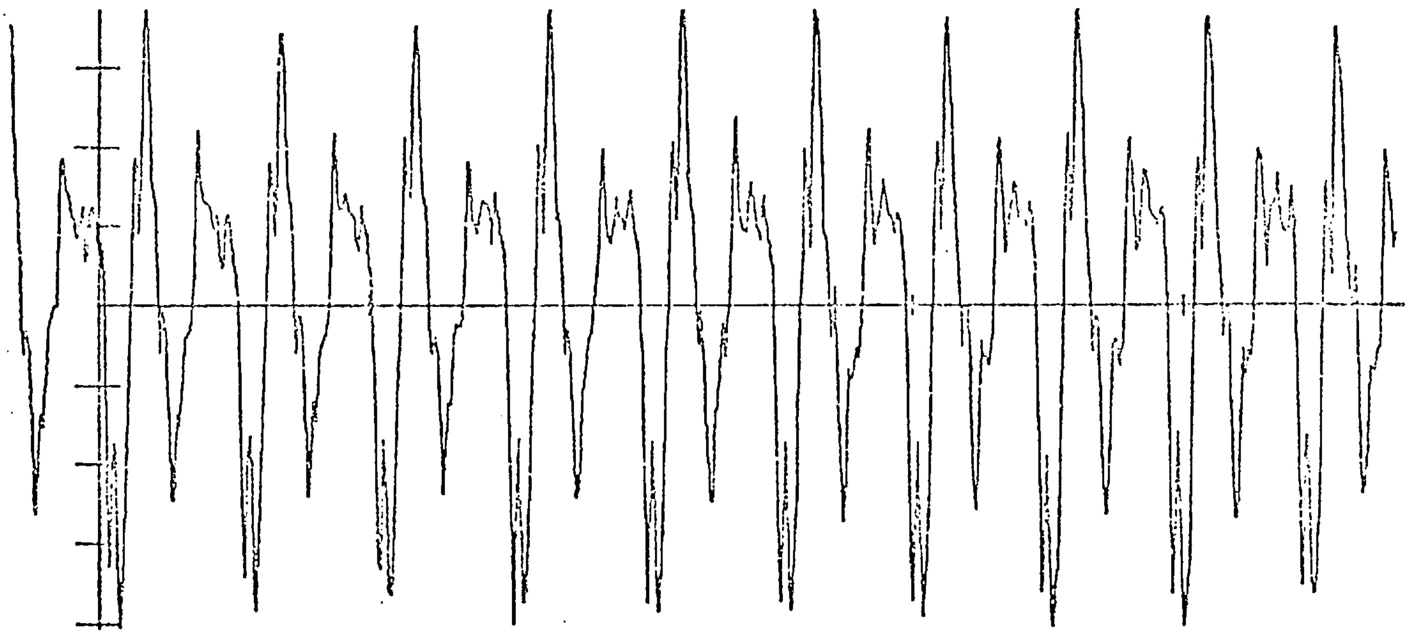
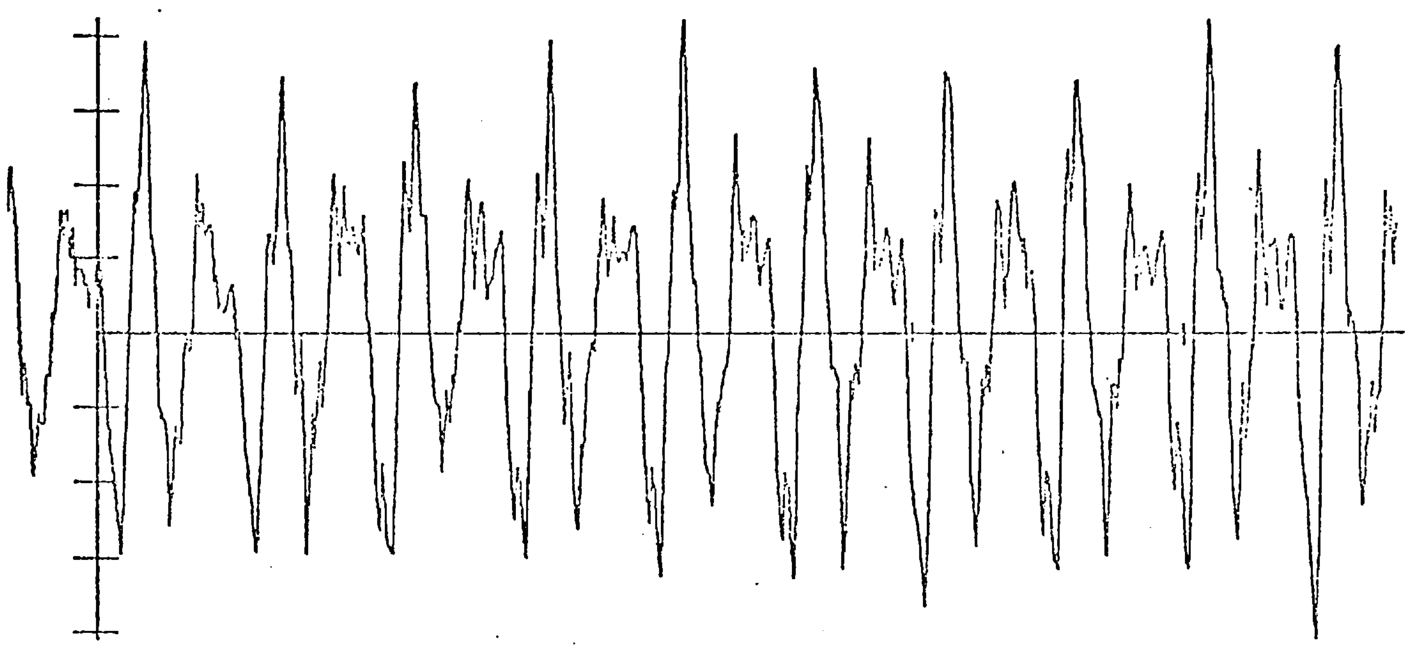


Fig. IV.9. Performance of the encoders versus sampling frequency (Predictor order = 4)



a) Input Waveform.



b) Output Waveform.

Fig. IV.10. An example of input and output waveforms at the data rate of 7 K bits/sec

Encoder with direct-form predictor	
$\sigma^{(k)} = \dfrac{\sigma^{(k-1)}}{1 - r_\sigma \dfrac{Q^{(k)} a_1^{(k)} Q^{(k-1)}}{\mu_\sigma^{(k)}}} + \Delta\lambda^{(k)}$	$\mu_\sigma^{(k)} = \overline{(a_1^{(k)} Q^{(k-1)})^2} + C_\sigma$
$\Delta\lambda^{(k+1)} = r_\lambda \dfrac{\sigma^{(k)} Q^{(k)} P^{(k)}}{\mu_\lambda^{(k)}}$	$\mu_\lambda^{(k)} = \overline{(P^{(k)})^2} + C_\lambda$ $C_\lambda = C_\lambda^+$ if $\sigma^{(k)} Q^{(k)} P^{(k)} > 0$ $C_\lambda = C_\lambda^- = 2C_\lambda^+$ if $\sigma^{(k)} Q^{(k)} P^{(k)} < 0$
$a_i^{(k+1)} = (a_i^{(k)} + r_a \dfrac{\sigma^{(k)} Q^{(k)} S_i^{(k)}}{u_a^{(k)}})(1 - \delta) + K_{a_i} \delta$, $i = 1,2,----,n.$	$u_a^{(k)} = \overline{S_i^2} + C_a$

Encoder with lattice-form predictor	
$\sigma^{(k)} = \dfrac{\sigma^{(k-1)}}{1 - r_\sigma \dfrac{\theta^{(k-1)} Q^{(k)}}{\mu_\sigma^{(k)}}} + \Delta\lambda^{(k)}$	$\theta^{(k)} = b_1^{(k+1)} Q^{(k)} - \sum_2^n b_i^{(k+1)} b_i^{(k)} Q^{(k)}$ $\mu_\sigma^{(k)} = \overline{(\theta^{(k-1)})^2} + C$
$\Delta\lambda^{(k+1)} = r_\lambda \dfrac{\sigma^{(k)} Q^{(k)} P^{(k)}}{\mu_\lambda^{(k)}}$	$\mu_\lambda^{(k)} = \overline{(P^{(k)})^2} + C_\lambda$ $C_\lambda = C_\lambda^+$ if $\sigma^{(k)} Q^{(k)} P^{(k)} > 0$ $C_\lambda = C_\lambda^- = 2C_\lambda^+$ if $\sigma^{(k)} Q^{(k)} P^{(k)} < 0$
$b_i^{(k+1)} = (b_i^{(k)} + r_{b_i} \dfrac{E_{i+1}^{(k)} F_i^{(k)}}{\mu_{b_i}^{(k)}})(1 - \delta) + K_{b_i} \delta$, $i = 1,2,----,n.$	$\mu_{b_i}^{(k)} = \overline{(F_i^{(k)})^2} + C_b$, $i = 1,2,----,n.$

Table IV.1. Parameter adaptation algorithms

Text	Speaker
1. It's easy to tell the depth of a well.	Male
2. These days a chicken leg is a rare dish.	Male
3. Glue the sheet to the dark blue background.	Male
4. The hog was fed chopped corn and garbage.	Female
5. Four hours of steady work faces us.	Female

Table IV.2. List of test sentences

Sampling Rate (Khz)	Bits per Nyquist Sample	SNR (dB)			
		Direct-Form Encoder		Lattice-Form Encoder	
		Unfiltered	Estimated	Unfiltered	Estimated
6	1.00	6.94	6.94	7.86	7.86
7	1.16	7.71	8.36	8.57	9.22
8	1.33	8.30	9.56	9.10	10.36
9	1.49	8.83	10.57	9.56	11.30
10	1.66	9.28	11.47	9.94	12.13
11	1.82	9.75	12.35	10.30	12.90

Table IV.3. SNR against sampling frequency

(Predictor order = 4)

Sentence No.	Unfiltered SNR (dB)
1	8.50
2	7.24
3	9.73
4	11.05
5	10.36
Average	9.38

Table IV.4. SNR of each test sentence with 7 KHz sampling rate (1.16 Bits/Nyquist sample) and N = 8

CHAPTER V

Phase Dithering Applied to Residual Encoding with a One-Bit Quantizer

The principle of adapative-predictive waveform-encoding, as developed by Atal and Schroeder[29], relies on the linear prediction of the incoming speech samples from the previous samples so that the power of the difference between the actual and the predicted signal, i.e. the error signal, is reduced. Theoretically, if the SNR of the quantizer (of B bits) remains constant, the improvement in SNR of the adaptive-predictive method over PCM (also of B bits) is equal to the ratio of the power of the input speech signal to the power of the prediction error*[29]. However, in the case of voiced input speech sounds, the greater the efficiency of the prediction process in reducing the power of the error signal, the more invalid is the assumption that the SNR of the quantizer will remain unchanged. Perfect (short-term) prediction of the voiced speech input would reduce the error signal to a periodic train of sharp pulses and these sharp pulses are very difficult to code accurately with ordinary amplitude quantizers. If the quantizer step size is made sufficiently large to handle these strong pitch pulses, then the overall quantization noise will be greatly increased. On the other hand, if these sharp pulses are allowed to be clipped by the quantizer, then severe non-linear distortion will result and this disrupts the linear prediction of the subsequent speech samples. In the scheme proposed by Atal and Schroeder, the problem associated with the pitch pulses is not encountered. In their scheme, periodic (long-term) redundancies are removed. This operation not only further reduces the error

---

* See also Chapter I, Section I.1, A.2.

signal power but also, to a large extent, eliminates the periodic sharp pulses in the error signal. However, the extraction of periodic redundancies requires a considerable amount of computation to determine the pitch period and the autocorrelation coefficients between input speech samples that are approximately a pitch period apart. Simplified adaptive predictive encoding schemes (e.g. that of Dunn(31)) often omit this periodic redundancy extraction. The ability of these simplified linear predictive methods, to attain the SRN gain that is theoretically possible with accurate prediction, is thus limited by the difficulty of coding the pitch pulses with an amplitude quantizer.

The disruption to the proper operation of a linear predictive encoder caused by the pitch pulses is worse for a special class of linear predictive encoders known as residual encoders*. In a residual encoder, information about the prediction residue is required for the adaptation of the predictor coefficients and the quantized prediction error is used as the estimate of the actual prediction residue. Correct adjustment of the predictor coefficients would require that the prediction residue (i.e. the residual fluctuations between pitch pulses) rather than the pitch pulse which is the source excitation, be coded accurately. For low-bit-rate applications the quantizer is necessarily coarse and if the step size of the coarse quantizer is adjusted to code the residue accurately, very severe clipping of the pitch pulses results. Hence, for residual encoders with coarse quantizers, usually very little can be gained by employing a large predictor order because even if the predictor is of sufficient order to fully model the input speech signal, perfect prediction will be hampered either by incorrect predictor coefficients

---

* See Chapter IV and Chapter I, Section I.1 A.3

or by a severely distorted excitation. In other words, the nature of the quantizer places a limit to the maximum achievable amplitude difference between the pitch pulse and the residual fluctations in the error signal. This means that the limit, to which the power of the error signal can be reduced, is set by the nature of the quantizer and not by the predictability of the input signal.

If the excitation source is noise, as is the case with unvoiced speech sounds, then the above problem is alleviated. Unlike the pitch pulse energy, the energy of a noise signal is spread over a long time interval, and it is thus much more suitable for amplitude quantization. In a residual encoder, although this spread out excitation introduces disturbances into the prediction residue, from which information for adjustment of the predictor coefficients has to be derived, the magnitude of the disturbances is not more than that which would be introduced by the quantization action. In other words, the amplitude of the prediction residue can at least be reduced to approximately the level of the noise excitation before this noise excitation will have any appreciable effect in disrupting the coefficient optimization process.

Thus, for a voiced speech input, if the periodic-pitch-pulse excitation can be converted into something resembling the noise excitation of unvoiced speech sounds (i.e. having the pulse energy spread out over a long time interval), then it can be digitized by a residual encoder with improved SNR. The network that performs this conversion should be a linear network since any non-linearity introduced might reduce the linear predictability of the resultant signal. The conversion required is one in which the flat spectrum pitch pulses are converted to flat spectrum noise-like fluctuations with the pitch-pulse energy being spread out. Thus it is clear that the conversion network should have a flat amplitude response and that the conversion operation can be obtained by a randomizing of the phase of the input signal and is thus referred to as "phase-dithering". A linear (time-invariant) phase-dithering network with a flat amplitude response has a number of valuable properties; among these are (i) the power spectrum of the input signal is preserved; (ii) a periodic input signal appears as periodic at the output; (iii) a non-periodic input signal appears at the output as a non-periodic signal. Hence, to the human perception system speech signals appear to sound the same, irrespective of whether they are processed by the phase dithering network or not. Thus, a restoration of the transmitted speech signal to its original phase is unnecessary.

This chapter is devoted to a description of an investigation into the application of the phase-dithering technique to one-bit residual encoding with a lattice-form of predictor structure of the type described in Chapter IV. It will be shown that a significant improvement in the SNR performance of the encoder can be obtained. Also, it will be shown that if some of the improvement of system performance is sacrificed, then it is possible to simplify considerably the parameter adaptation algorithms of the encoder. The resulting simplified encoder will be shown to require approximately 3n multiplications per iteration for a predictor of order n. This means that the system is only approximately an order of magnitude more complex than simple fixed-predictor differential encoders.

V.1. The Phase Dithering Network

A model for the production of voiced sounds is shown in part a. of Fig. V.I. The vocal tract is excited by a series of nearly periodic pulses produced by vibration of the vocal chords. The acoustical vibrations radiated from the mouth through the vocal tract form the final voiced speech wave. In general, in Fig. V.1, the term vocal tract transfer function is used loosely, as it represents the total effects of the spectrum shaping due to the vocal tract, lip radiation and the glottal pulse shape. As explained in the previous section, if the quasi-periodic pulse-like excitation of voiced speech is changed to acquire the properties of the noise-like excitation associated with unvoiced speech, the resultant waveform is rendered more suitable for residual encoding. In Fig. V.1.b, a network, X, is added between the excitation source and the vocal tract transfer function to perform the conversion. The network has to function so as to distribute the pulse energy more evenly over the time-axis and also to preserve the flat power spectral envelope of the excitation source. A network that approaches these requirements is a long non-recursive digital filter whose coefficients make up a cycle of a pseudorandom sequence. If an impulse is received as an input to the network then the output is exactly a cycle of this psuedo-random sequence. Fig. V.2 shows an example of a filter of this kind. The filter shown has, as coefficients, a maximal-length pseudo-random binary sequence (PRBS) of cycle length 7.

It is well known that the auto-correlation of a pseudo-random sequence is a sharp peak where the correlation delay is equal to an integer multiple of the cycle length and that it is zero, or nearly zero, elsewhere*. However, with one single cycle of the pseudo-random sequence, the auto-correlation function has a sharp peak at

---

* In the case of the maximal length PRBS, the magnitude of the auto-correlation function is equal to the cycle length at the sharp peaks and -1 elsewhere.

zero delay and some small fluctuations elsewhere. The magnitude of the fluctuations is dependent not only on the cycle length of the sequence but also on the starting point in the sequence from which the cycle is taken. To illustrate this, the auto-correlation functions for various single-cycles of a PRBS of cycle length 7 are shown in Table V.1, from which it can be seen that the sequence number 4 has the lowest values of non-zero-delay auto-correlations. In general, the longer the cycle length the higher is the magnitude ratio of the central peak to the highest non-zero-delay auto-correlation. The auto-correlation function of a single cycle of the PRBS is close to an impulse when the cycle length is large. This means that the impulse response of a non-recursive filter using this cycle of PRBS as coefficient values has an auto-correlation function that is close to being an impulse and, consequently, the impulse response has an almost white (flat) power spectrum. Thus the amplitude response of this filter is approximately flat.

With reference to Fig. V.1.b, it is clear that it is not possible in practice to place the phase dithering filter, X, between the excitation source and the vocal tract transfer function. However, as the phase dithering filter is a linear network, its position can be interchanged relative to that of the vocal tract transfer function without this having any effect on the final output. The result of this arrangement is shown in Fig. V.1.c. This is important, since it means that the desired alteration on the excitation can be achieved by passing the speech wave through the phase dithering filter.

It seems reasonable to assume that, to achieve an effective spreading out of the energy in the pitch pulses, the length of the one cycle of the PRBS should be at least a pitch period. Assuming.

a sampling frequence of 8 kHz and a minimum pitch frequency of 100 Hz, the longest pitch period is equal to 80 sampling-intervals. As the cycle length of a maximal-length PRBS is given by $(2^m - 1)$, where m is an integer, this means that m should not be less than 6 which gives a cycle length of 63. The cycles of PRBS with m = 6 and 7, that have the minimum "side lobe" fluctuation in their auto-correlation, are given in Table V.2.

The use of a digital filter whose number of stages is of the order of a hundred seems prohibitively costly at first sight. But, the use of PRBS as coefficients means that no multiplications are required, because the coefficients are either 1 or -1. Thus the additional expense in employing this phase dithering filter is only the cost of the delay elements.

V.2. One-Bit Residual Encoding with Phase-dithering

The block diagram of a phase-dithered one-bit residual encoding and decoding system is shown in Fig. V.3. Apart from the addition of the phase dithering filter placed before the encoder, the system is in fact a normal one-bit residual encoding and decoding scheme of the type described in Chapter IV. On the assumption that both the power spectrum and the voicing information of the input speech wave are preserved by the phase dithering filter, the signal leaving the phase dithering filter should sound the same as the original speech. Restoration of the transmitted speech to its original phase relationships can, therefore, be omitted. Also, since after the phase dithering the signal can be considered as equivalent to the original speech, the SNR performance of the encoder can be determined from the reconstructed speech, S_3, and the phase dithered speech, S_2, that is:

$$SNR = \frac{\overline{S_2^{\ 2}}}{\overline{(S_3 - S_2)^2}}$$

The system of residual encoding with phase dithering was simulated on a PDP 15 computer and the SNR of above was measured for the test sentence "It's easy to tell the depth of a well". The sentence, which is No. 1 in Table IV.2, was spoken by a male speaker. This particular sentence was used since experience gained in measuring the SNR performance of the non-dithered one-bit residual encoder had shown that although the sentence has an SNR which is slightly below the average SNR of the five test sentences of Table IV.2, the manner in which its SNR varies with the number of predictor stages is representative of the average behaviour with respect to this parameter. During the course of simulation two phase dithering filters were tested. PRBSs of cycle length 63 and 127 as shown in Table V.2 were used respectively as coefficients of the two filters. In the experiments, the test sentence was pre-filtered using a 3020 Hz sharp-cut-off low-pass filter and the output from the filter was sampled at 8 k samples/sec. The unfiltered SNRs were measured as a function of the order of the predictor. The results of the measurement are shown in Table V.3, where they are tabulated and compared with those of the original non-dithered one-bit residual encoder. The results are also shown graphically in Fig. 4. From Fig. 4, it can be seen that, irrespective of the predictor order the SNR of the residual encoder is significantly improved if the input signal is phase dithered. The SNR advantage with phase-dithering can be seen to increase with the order of the predictor and this thus appears to confirm the conjecture that the non-linear clipping of the pitch pulses in a non-dithered residual encoder is a factor which prevents the encoder from increasing the accuracy of its prediction as the order of the predictor is increased. It should be noted, however, that, even with phase dithering, the improvement of SNR with predictor order decreases for predictor of order greater than 6. This is probably a result of low-pass filtering of the original speech signal so that its highest frequency component is approximately 3 kHz. Usually, not more than 3 formants are present in speech that has been low-pass limited to 3 kHz and, therefore, a 6th order filter is sufficient to describe the 3 poles. The results of the test show that the advantage in SNR when using the phase dithering filter of 63 coefficients was less than that obtained when using the longer filter of 127 coefficients and its rate of increase with predictor order was slower. However, with predictor order greater than 6, both of these filters achieved an SNR improvement of about 1 dB.

The effect that phase dithering has on the waveform of the prediction error is clearly evident from Fig. V.5. Figs. V.5.a and V.5.b show respectively a section of a voiced speech sound before and after phase dithering. Fibs. V.5.c and V.5.d are the predictor error waveforms associated with the section of speech in Fig. V.5.a. Fig. V.5.c is the error waveform for an non-dithered one-bit residual encoder and Fig. V.5.d the error waveform for a dithered one-bit residual encoder. It should be noted that the prediction error of the undithered system has sharp peaks at regular intervals corresponding to the pitch pulses. The error waveform associated with the phase-dithered-system is relatively consistent in amplitude.

In order to verify that the SNR improvement was not something restricted to the particular sentence under test, the SNRs associated with the remaining sentences in Table IV.2 were measured. The tests were carried out using predictor of order 6 and the same prefiltering and sampling were used as when testing the first sentence. The previously used two lengths of phase-dithering filter were again employed. The results of the measurement are tabulated in Table V.4. On the average, the SNR improvement over non-phase-dithered one-bit residual encoder is 0.9 dB with the 63-coefficient filter and 1.51 dB with the 127-coefficient filter.

As predicted, the signal appearing at the output of the phase-dithered one-bit residual encoding system, does not sound different from the pre-phase-dithered speech, other than the small background noise introduced by the encoding process. The improvement in quality over non-phase-dithered but similarly residual-encoded speech is noticeable. However, with the dithering filter of 127 coefficients, the duration of some short speech syllables can be shorter than the energy spread caused by the dithering filter. Thus, such short syllables (e.g. the consonant "th" in "depth") are merged into their neighbouring sounds and appear slightly muffled. However, this is not as drastic a shortcoming of the phase dithering system as it may seem to be. It should be noted that such relatively short acoustic events are very difficult to be coded properly by the normal non-phase-dithered residual encoder in any case. For such a short duration sound, the predictor coefficients do not have sufficient time to adjust themselves to suit the statistics of the sound over the interval during which this sound lasts.

V.3. Phase Restoration

If a phase-dithering filter of 127 or more coefficients is to be used, then it appears likely that it may be advantageous to restore the phase of the phase-dithered signal (i.e. the accumulation of the dispersed energy back to one concentrated point). Although, in principle, this restoration could be performed by the exact inverse of the phase-dithering filter, this is not in fact possible because the inverse of the phase-dithering filter is unstable (except for the filter whose coefficients are the PRBS of cycle length 3, which has its roots exactly on the unit circle). However, an approximate reconstruction of the original phase can be obtained by using not the exact inverse filter but a linear non-recursive digital filter whose coefficients are the same cycle of PRBS as in the original filter, but in which the direction of signal flow is reversed. Alternatively, the direction of signal flow can be kept the same while the sequential order of the PRBS which forms the coefficient values is exactly reversed. An approximate restoration filter for the 7-coefficients phase-dithering filter of Fig. V.2 is shown in Fig. V.6. The impulse response of the cascade of the phase dithering filter and the approximate restoration filter is the auto-correlation function of the cycle of the PRBS. Thus, an impulse input produces a delayed impulse output plus some fluctuations around the peak. The ratio of the energy of the central spike in the impulse response to the sum total of the energies of the side-ways fluctuations is a measure of how well the dispersed energy is re-accumulated. For a random input, this is the average energy ratio between the part in a sample of the restored signal which has originally belonged to this sample and the part which is "spilled over" from other samples. Thus this ratio can be termed the "spill-over rejection ratio". For the cycles of PRBSs of lengths 63 and 127 shown in Table V.2, the "spill-over rejection ratios" are 6.67 dB and 6.25 dB respectively. With correlated inputs, because of the randomness of the values of the non-zero delay auto-correlations of the cycle of the PRBS, the spill-over energies tend to cancel and the ratio is thus a little higher. This degree of separation of the original energy to the spilled over energy is sufficient for the purpose of isolating the short acoustic events but it will significantly degrade the apparent SNR figure of the encoder, if the received and approximately phase-restored speech is compared with the original pre-phase-dithered speech input.

V.4. Phase-dithered one-bit Residual Encoding with Phase Restoration

In the testing of the phase-dithered one-bit residual encoder with phase-restoration , the effectiveness of the phase restoration network was evaluated first, with the cascade of the phase-dithering and the phase restoration networks shown in Fig. V.7 being simulated. It should be noted that if the length of the cycle of the PRBS used as coefficient values in the filters is L, then the output signal is delayed by L-1, and its amplitude is increased L times, relative to the input. This means that care should be exercised in any comparisons of the input and output to take account of these changes. It was found by listening that the output from the cascade of phase-dithering and phase restoring networks was virtually indistinguishable from the original. On the other hand, it was found that the apparent measured ratio of the power of the input signal to the power of the difference between the input and output was significant. With test sentence No. 1, for example, these "SNRS" were 7.09 dB and 6.69 dB respectively for the filters of 63 and 127 coefficients. It thus appears that extreme caution should be exercised in interpreting these SNR figures in terms of subjective quality.

As a next step, the effect of phase restoration on the phase-dithered one-bit residual encoded and decoded signal was examined. The system shown in block diagram in Fig. V.8 was simulated. Test sentence No. 1 was again used. The sampling frequency was fixed at 8 k Samples/sec and an 8th-order predictor was used. The results of a short series of listening tests indicated that the slight "muffling" of the very short consonant sound "th" in "depth" appeared to have been removed and the overall sentence sounded a good deal sharper, or more precise. However, though the restoration filter roughly reconcentrated the energy of the original speech that has been dispersed by the dithering filter, it, being another form of phase dithering filter, has the effect of spreading the noise introduced by the 1-bit residual encoding. Hence, the quantization noise produced by the coding of a speech syllable becomes, after phase restoration, out of step with that speech syllable. This effect seems to be more easily noticeable than the muffling of short acoustic events without phase restoration, but is so subtle that it would be necessary to use very elaborated subjective testing in order to determine whether it would have any effect on the intelligibility of the speech or whether it makes the received sound less pleasant to listen to.

The overall SNR of the complete system, including the phase-dithering and restoration filter, is given by (see Fig. V.8)

$$SNR = \frac{\overline{S_6^{\,2}}}{\overline{(S_5 - S_6)^2}}$$

Because of the limitation of the "spill-over" rejection ratio of the phase restoration network, this SNR appeared to be rather poor. For the test sentence No. 1, the SNR is 5.45 dB for filters of 63 coefficients and 5.30 dB for a filter of 127 coeffients.

As the signal, S4 (see Fig. V.7), after processing by the cascade of the phase dithering filter and restoration filter is indistinguishable from the original, S4 can be taken as the original signal and a more realistic SNR of the received speech is given by (see Fig. V.8)

$$SNR = \frac{\overline{S_4^{\,2}}}{\overline{(S_4' - S_4)^2}}$$

It was found that when using this expression, the measured SNR's were 9.44 dB and 9.67 dB respectively for filters of 63 and 127 coefficients and it will be noticed that these values are approximately the same as those measured in Section V.3.

V.5. Simplified Parameter Adaptation Algorithms

The improvement in SNR performance of the one-bit residual encoder for a phase-dithered input signal can be exploited so as to reduce the complexity of the encoder. The encoder is simplified by changing all parameter adjustments to fixed quantum-steps. This means that, at each iteration, each of the parameters (i.e. the quantizer step size and predictor coefficients) of the residual encoder is either increased or decreased by a fixed value (i.e. a quantum step). The best values for these quantum steps could be decided experimentally and fixed for all speech utterances. In this case, the only information needed for parameter adjustment is the information as to whether the particular parameter should be increased or decreased. As only polarity information is required, most of the expensive multiplying and dividing operations in the parameter adaptation algorithms can be replaced by "Exclusive-Or" type logic operations. In Table V.5, the simplified parameter adaptation algorithms for a one-bit residual encoder with the lattice-type predictor network are compared with the original algorithms developed in Chapter IV. The number of multiplications and divisions is seen to be reduced from $8n + 14$ to $2n - 1$, where n is the order of the predictor. If no count is taken of the multiplications that can be replaced by a shifting operation, the number of multiplications and divisions is seen to be reduced from $5n + 9$ to $n - 1$. As the actual predictor network uses another 2n multipliers, the total number of multiplications to be performed, per iteration of the simplified encoder, is approximately 3n.

The simplification results in a degradation in SNR of about 1 dB but, with such a simplification, the order of the predictor network no longer poses cost problems. The significance of this is that, to obtain a specific level of performance, it may be easier to use a simplified encoder together with a higher order predictor, than to use the unsimplified algorithms and a shorter predictor. In the absence of phase dithering, however, this idea does not hold good because the rate of increase of the SNR of the one-bit encoder is limited for predictor orders of 4 and above. With phase dithering, however, the limiting point is shifted to $n = 6$ and there is about 1 dB SNR advantage over the non-dithered version.

The ideas of immediately above were tested by simulation using the test sentence No. 1. Simplified phase dithered one-bit residual encoding without phase restoration, similar to the block diagram in Fig. V.3, was used. With an 8th order predictor and a 63-coefficient phase ditherer, the measured SNR was 8.79 dB. Comparing with the figures in Table V.3, it is seen to be better than the 8.62 dB of the unsimplified version of order 4 which in turn is better than the best that can be achieved without phase-dithering:- 8.56 dB with an 8th order predictor.

The viability of the simplification was further investigated by SNR measurements on the other test sentences in Table IV.2. The results, using the simplified dithered encoder of order 8 are listed in Table V.4 against those of the other methods employing an 8th order predictor. Not only is the SNR of the simplified phase-dithered version generally better than the non-dithered one-bit residual encoder, but the simplified version is likely to be cheaper to implement.

V.6. Discussions

Following the remarks made in the previous chapter, that the performance of residual encoders falls short of that that appears possible, the investigations reported on in this chapter were carried out. These investigations have confirmed that improvements are possible and have identified one important cause of impairment, namely, the inability of an amplitude quantizer to cope with the sharp pitch pulse in the error waveform. Because of this limitation, perfect prediction of a voiced speech waveform is never possible, even given a predictor of infinite order and given instantaneous convergence of the predictor coefficients.

The technique of phase dithering has been proposed as a means of overcoming this difficulty. The improvement in the SNR performance of the one-bit residual encoder that results from the use of the phase-dithering technique is very encouraging. The most significant achievement of the technique is, however, not so much in being able to further improve on a performance that is already satisfactory, but is in the simplification of the encoder (with accompanying loss of SNR) that can be afforded by making use of the performance "surplus". The simplified encoder only requires one additional multiplication per iteration per predictor stage for the adaptation of all the encoder parameters and this thus appears to be about the simplest form that a residual encoder with a lattice-form predictor can be reduced to. The phase dithering filter can be implemented with a single chip of a dynamic MOS shift register or, probably more simply, with a charge-coupled-device shift register of a relatively short length. Hence, the additional cost of the phase-dithering filter is relatively insignificant and a hardware impelementation of the complete encoder could be quite inexpensive and compact in size.

The method employed for phase dithering is satisfactory as far as preservation of the original speech power-spectrum and spreading out of the energy of the sharp pitch pulses are concerned. It is, however, far from perfect in respect of phase restoration. A search program, looking through all possible binary combinations given a fixed sequence-length, has been designed and, on searching through some short sequence lengths (up to 17) with this program, it has been found that some finite-length sequences of "spill-over-rejection-ratio" much higher than that of a cycle of a PRBS exist. If a sequence of high "spill-over-rejection-ratio" and long sequence length (say about 100) can also be found, it would be much more suitable to be used as the coefficients of the phase-dithering (and restoration) filter, than the PRBS currently employed.

On the other hand, the technique of phase dithering is not necessarily the only way of combatting the problem created by the sharp pitch pulses. A suitable run-length code, to take care of the infrequent situations where the magnitude of the error signal far exceeds the quantizer step size due to the occurrence of a pitch pulse, could be another solution. This question of how a pitch pulse could best be coded is open for very extensive investigations.

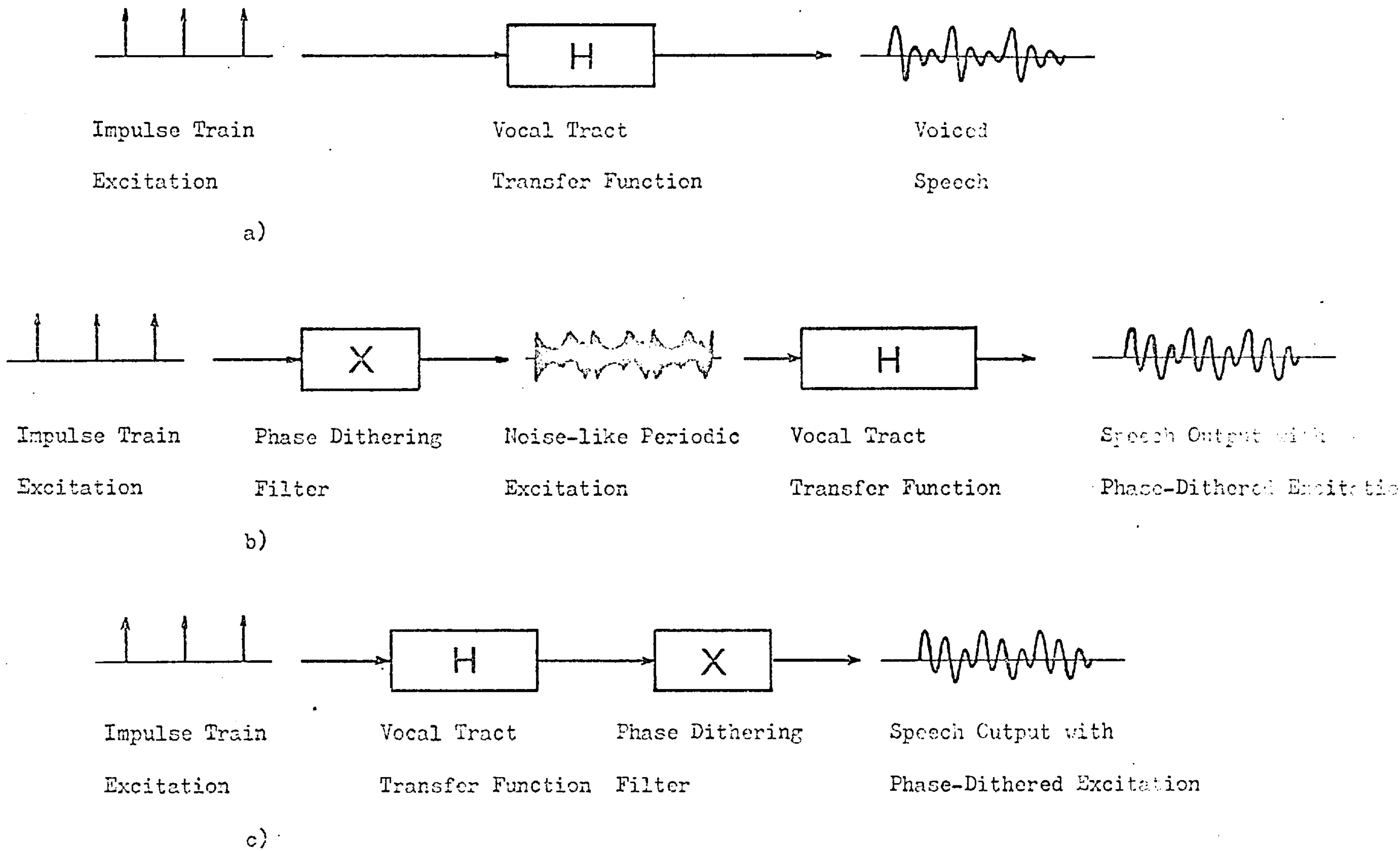

H
Impulse Train
Excitation
Vocal Tract
Transfer Function
Voiced
Speech
a)
X
H
Impulse Train
Excitation
Phase Dithering
Filter
Noise-like Periodic
Excitation
Vocal Tract
Transfer Function
Speech Output with
Phase-Dithered Excitation
b)
H
X
Impulse Train
Excitation
Vocal Tract
Transfer Function
Phase Dithering
Filter
Speech Output with
Phase-Dithered Excitation
c)

Fig. V.1.

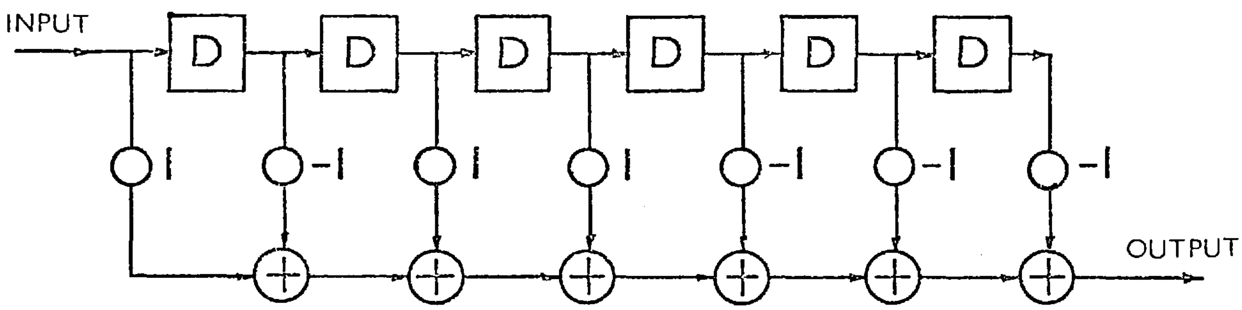


Fig. V.2. A non-recursive filter with a PRBS of cycle length 7 as coefficients

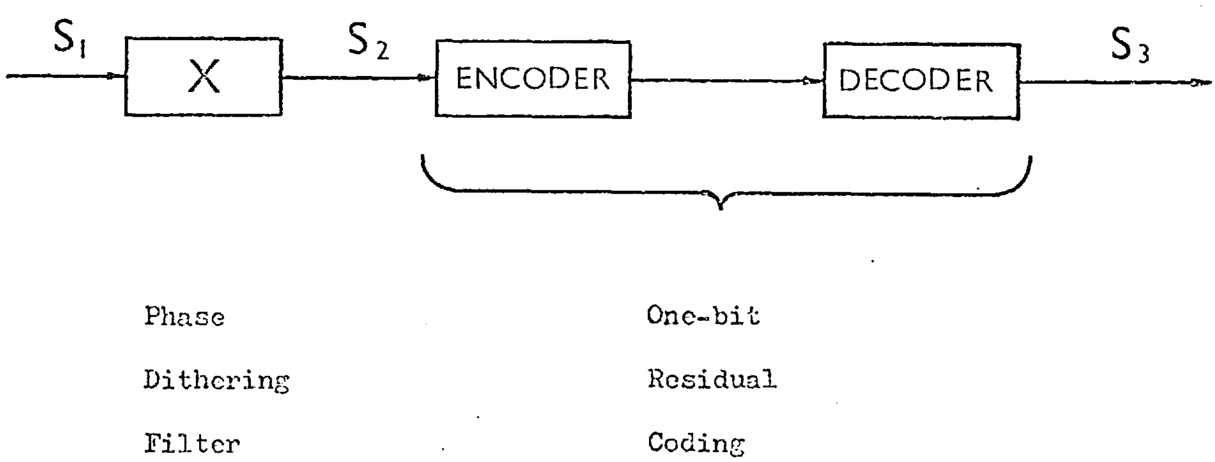


Fig. V.3. A phase-dithered one-bit residual coding scheme

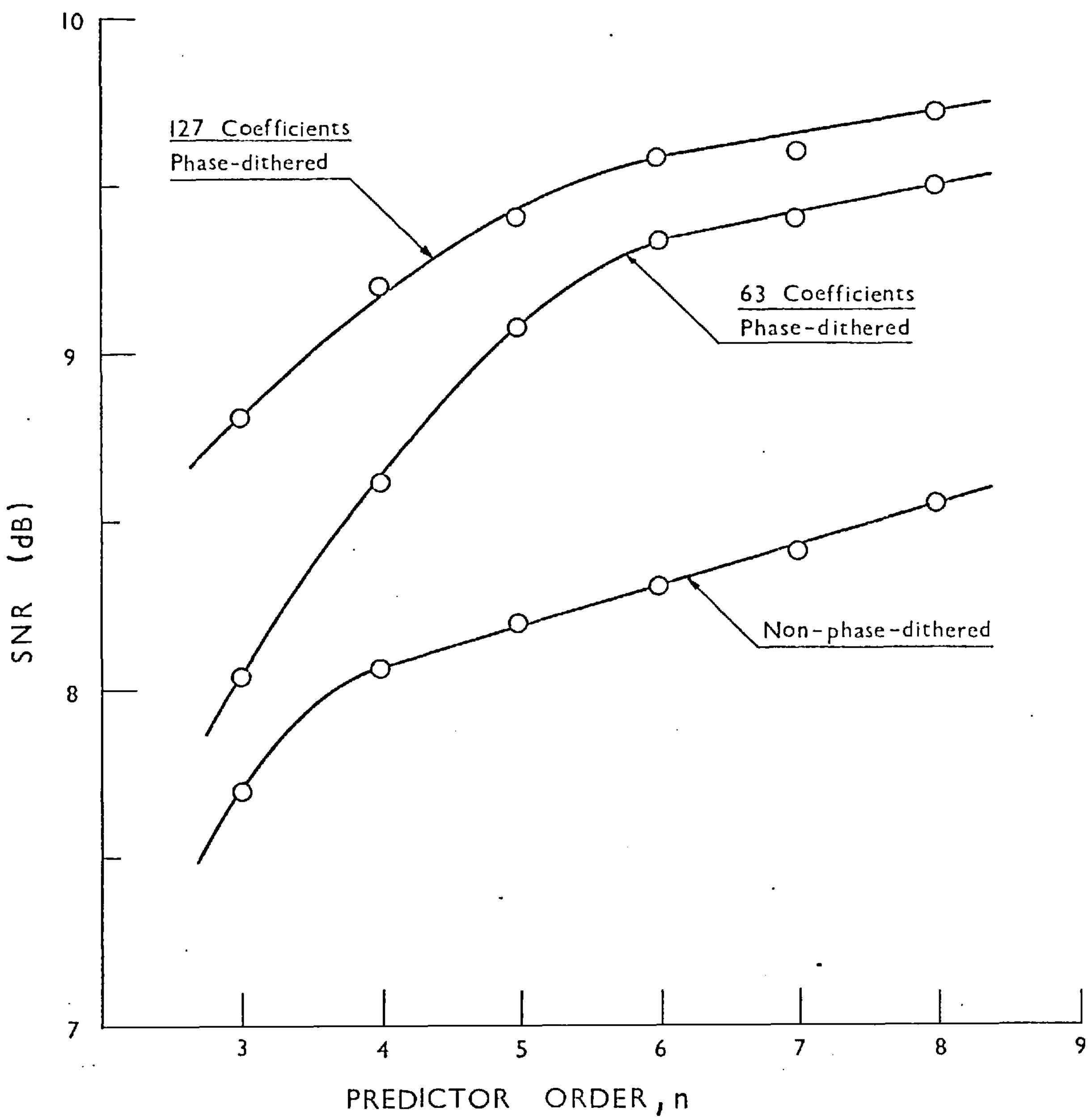
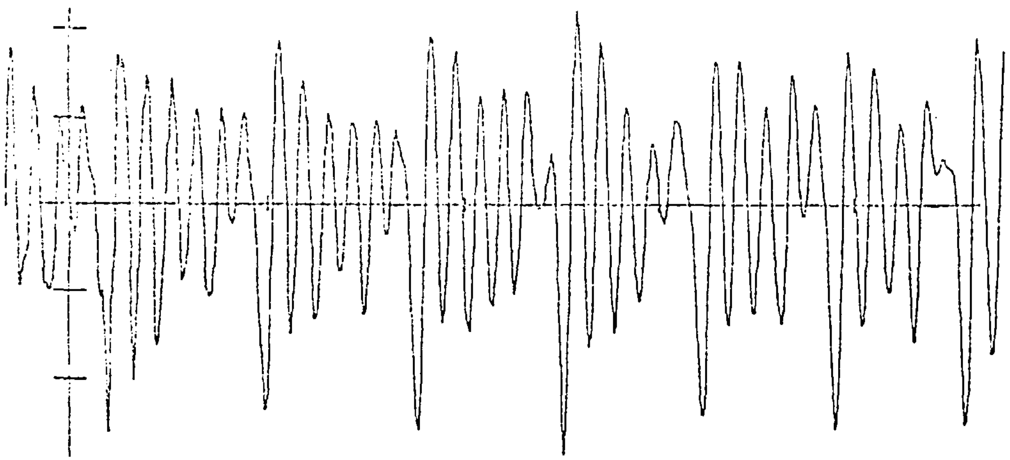
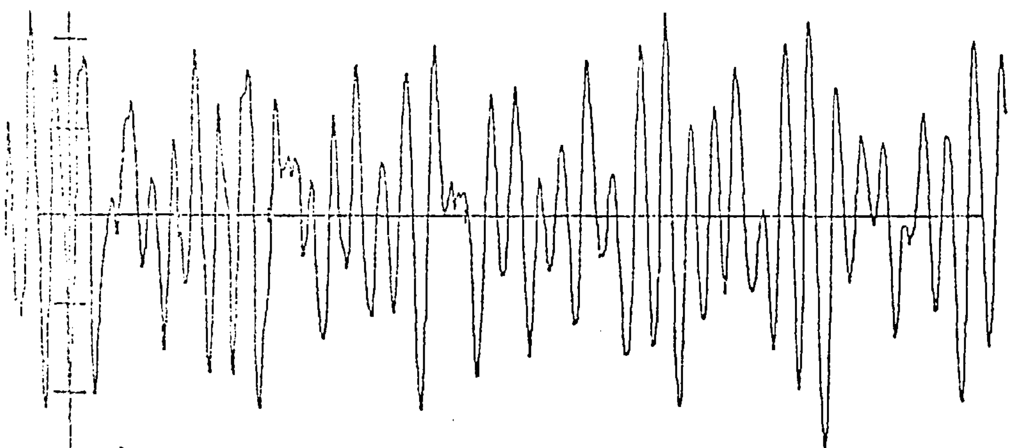


Fig. V.4. SNR of test sentence No. 1

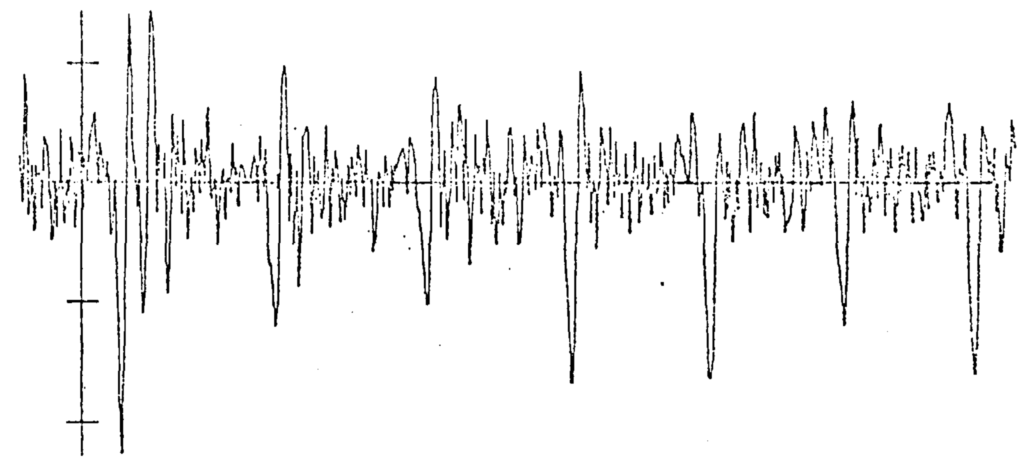
(Sampling Freq. = 8 K samples/sec)



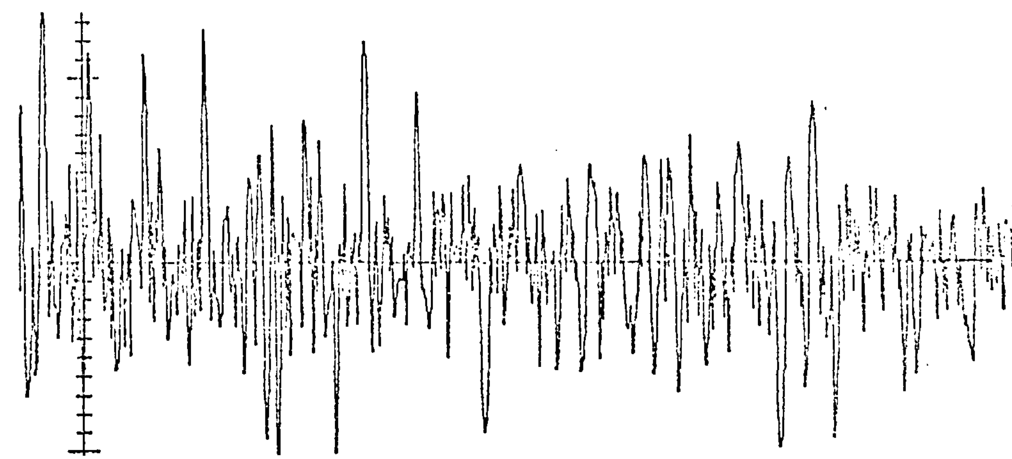
a) Original speech waveform



b) Phase dithered speech waveform



c) Error waveform without phase dithering



d) Error waveform with phase dithering

Fig. V.5. Illustrations of effects of phase dithering on the error waveform.

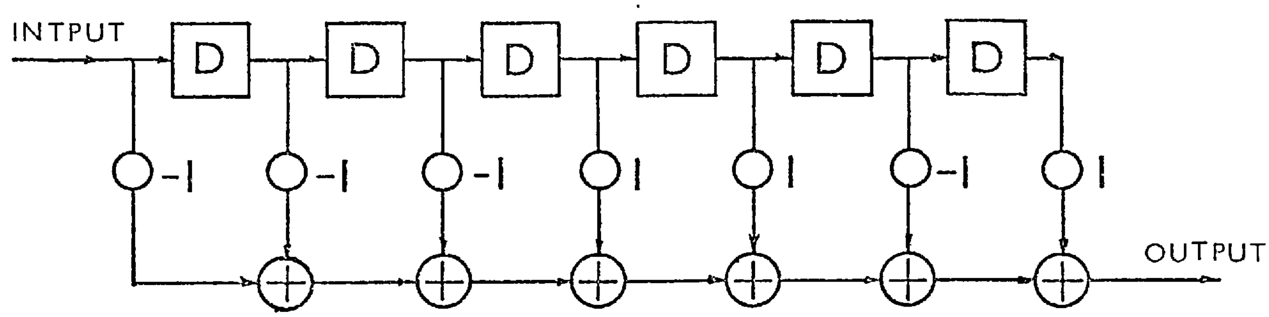


Fig. V.6. Approximate phase restoration network for the phase dithering filter of Fig. V.2

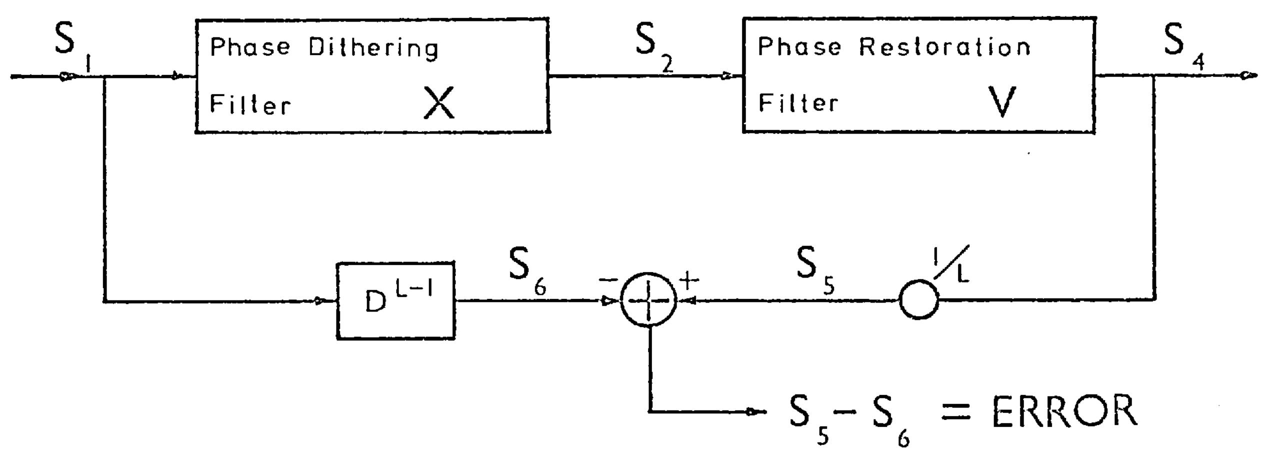


Fig. V.7. Comparison of the approx. phase restored waveform with the original

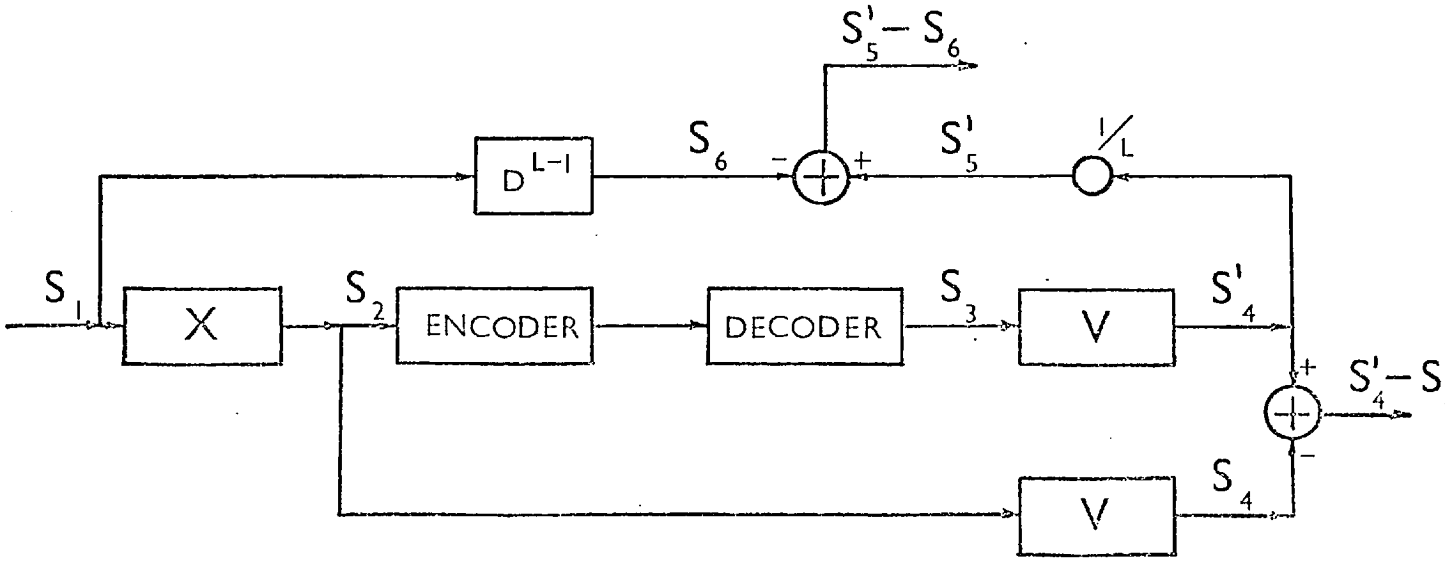


Fig. V.8. Phase dithered residual encoder with phase restoration

	Binary Sequences of Length 7	Auto-correlation (left hand side)	Max. Non-zero-delay Autocorrelation
1	-1, 1, 1, 1,-1, 1,-1	1,-2, 1,-2, 1,-2, 7	-2
2	-1,-1, 1, 1, 1,-1, 1	-1, 0, 1,-2,-1, 0, 7	-2
3	1,-1,-1, 1, 1, 1,-1	-1, 2, 1,-2,-3, 0, 7	-3
4	-1, 1,-1,-1, 1, 1, 1	-1, 0,-1, 0,-1, 0, 7	-1
5	1,-1, 1,-1,-1, 1, 1	1, 0,-1, 0,-1,-2, 7	-2
6	1, 1,-1, 1,-1,-1, 1	1, 0,-3, 2,-1,-2, 7	-3
7	1, 1, 1,-1, 1,-1,-1	-1,-2,-1, 0, 1, 0, 7	-2

Table V.1. The autocorrelation functions of various single cycles of a maximal length PRBS of cycle length 7

m = 6	Cycle Length = 63																			
Binary Sequence	-1	1	1	1	1	1	1	-1	1	-1	1	-1	1	1	-1	-1	1	1	-1	1
	1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	1	1	1	-1	-1	-1	1	-1
	1	1	1	1	-1	-1	1	-1	1	-1	-1	-1	1	1	-1	-1	-1	-1	1	-1
	-1	-1	-1																	
Auto-correlation (Left Hand Side)	1	0	-1	-2	-5	-2	-3	-2	-3	-4	-3	-2	1	-4	-1	0	-3	0	-3	-6
	-1	-4	3	-2	-1	0	3	0	-5	-4	3	-4	3	4	-1	-4	-1	0	1	-4
	3	0	5	2	-1	2	-1	0	3	-2	1	2	3	2	1	2	1	4	1	0
	-1	-2	63																	

Max. Non-zero-delay Auto-correlation = -6	Total Side-way Energy = 854	Energy of Central Peak = 3969

m = 7	Cycle Length = 127																			
Binary Sequence	-1	1	-1	-1	1	1	1	1	-1	1	1	1	-1	-1	-1	-1	1	1	1	1
	1	1	1	-1	-1	-1	1	1	1	-1	1	1	-1	-1	-1	1	-1	1	-1	-1
	1	-1	1	1	1	1	1	-1	1	-1	1	-1	1	-1	-1	-1	-1	1	-1	1
	1	-1	1	1	1	1	-1	-1	1	1	1	-1	-1	1	-1	1	-1	1	1	-1
	-1	1	1	-1	-1	-1	-1	-1	1	1	-1	1	1	-1	1	-1	1	1	1	-1
	1	-1	-1	-1	1	1	-1	-1	1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	1
	-1	-1	1	-1	-1	1	1													
Auto-correlation (Left Hand Side)	-1	0	1	-2	-1	6	1	-2	3	-2	1	4	-3	-2	-3	-2	-3	0	1	0
	-1	-4	-3	2	-5	0	-3	4	-3	-8	-7	0	-5	-4	-5	4	-3	-6	1	6
	-3	2	1	-4	9	-2	-3	-6	3	10	-3	-8	-1	-6	3	0	3	0	-1	-6
	-3	-8	-3	2	7	2	5	0	-1	-4	-1	-4	5	0	7	2	-11	-4	3	2
	1	-10	3	-2	-3	2	-7	-2	5	2	-5	4	3	4	-1	6	7	2	-5	2
	-1	4	-3	2	3	0	-1	-2	-1	2	1	2	1	2	-5	-2	1	-4	1	-2
	-7	0	1	-2	-1	0	127													

Max. Non-zero-delay Auto-correlation = -11	Total Side-way Energy = 3830	Energy of Central Peak = 16129

Table V.2. The maximal length PRBSs of cycle lengths 63 and 127. with lowest non-zero-delay auto-correlations

Predictor Order n	Unfiltered SNR (dB)		
	Non-Phase-Dithered	Phase-Dithered	
		63-Coef-Filter	127-Coef-Filter
3	7.70	8.04	8.81
4	8.07	8.12	9.20
5	8.21	9.08	9.41
6	8.31	9.34	9.58
7	8.41	9.40	9.60
8	8.56	9.50	9.72

Table V.3. SNR of test sentence No. 1

(Sampling freq. = 8 KHz)

Test Sentence No.	Unfiltered SNR (dB)			
	Non-Phase-Dithered	Phase-Dithered		
		63-Coefs	127-Coefs	Simplified (63-Coefs)
1	8.56	9.50	9.72	8.79
2	7.48	8.04	9.05	7.56
3	9.90	11.65	12.45	10.84
4	11.27	11.80	12.52	11.13
5	10.56	11.33	11.62	10.67
Average	9.56	10.46	11.07	9.80

Table V.4. SNR of test sentences

(Sampling freq. = 8 KHz,

Order of predictor = 8)

Encoder with lattice-form predictor	
$\sigma^{(k)} = \dfrac{\sigma^{(k-1)}}{1 - r_\sigma \dfrac{\theta^{(k-1)}Q^{(k)}}{\mu_\sigma^{(k)}}} + \Delta\lambda^{(k)}$	where $\theta^{(k)} = b_1^{(k+1)}Q^{(k)} - \sum_2^n b_i^{(k+1)} b_i^{(k)} Q^{(k)}$ $\mu_\sigma^{(k)} = \overline{(\theta^{(k-1)})^2} + C$
$\Delta\lambda^{(k+1)} = r_\lambda \dfrac{\sigma^{(k)}Q^{(k)}P^{(k)}}{\mu_\lambda^{(k)}}$	" $\mu_\lambda^{(k)} = \overline{(P^{(k)})^2} + C_\lambda$ $C_\lambda = C_\lambda^+$ if $\sigma^{(k)}Q^{(k)}P^{(k)} > 0$ $C_\lambda = C_\lambda^- = 2C_\lambda^+$ if $\sigma^{(k)}Q^{(k)}P^{(k)} < 0$
$b_i^{(k+1)} = \left(b_i^{(k)} + r_{b_i}\dfrac{\hat{E}_{i+1}^{(k)}\hat{F}_i^{(k)}}{\mu_{b_i}^{(k)}}\right)(1 - \delta) + K_{b_i}\delta$, $i = 1,2,----,n.$	" $\mu_{b_i}^{(k)} = \overline{(\hat{F}_i^{(k)})^2} + C_b$, $i = 1,2,----,n.$

Simplified algorithms for encoder with lattice-form predictor
$\sigma^{(k)} = \dfrac{\sigma^{(k-1)}}{1 - r_\sigma \mathrm{sign}(\Omega^{(k-1)})\mathrm{sign}(Q^{(k)})\mathrm{sign}(Q^{(k-1)})} + \Delta\lambda^{(k)}$ where $\Omega^{(k)} = b_1^{(k+1)} - \sum_2^n b_i^{(k+1)} b_i^{(k)}$
$\Delta\lambda^{(k+1)} = r_\lambda \mathrm{sign}(Q^{(k)})\mathrm{sign}(P^{(k)})$
$b_i^{(k+1)} = (b_i^{(k)} + r_{b_i}\mathrm{sign}(E_{i+1}^{(k)})\mathrm{sign}(F_i^{(k)}))(1 - \delta) + K_{b_i}\delta$, $i = 1,2,----,n.$

Table V.5. Comparison of the simplified parameter adaptation algorithms with the original

CHAPTER VI

Conclusions

In this thesis, a number of new techniques for the low-bit-rate coding of speech signals have been proposed and investigated. The motivation behind the development of the new techniques was discussed in detail in Chapter I. Very briefly, the motivation is the creation of a relatively simple speech coding system, that is capable of transmitting reasonably clear speech at a bit rate close to 1 bit per Shannon sample. Towards this end two alternative approaches were thought to be possible. The method of the first approach is to start from a complex vocoding method that is capable of transmitting good quality speech at extremely low bit-rates and look for methods that, by sacrificing some transmission bandwidth, could reduce the complexity of the coding system. The method of the second approach is to look for methods that, by introducing as little added complexity as possible, could either reduce the bit rate requirements of relatively simple waveform coders or improve their performance for a given bit rate. Basically, it is the second approach that is adopted in this research, although the principle of vocoding methods does have a strong influence in the inception of some of the new coding techniques. In this concluding chapter, the new techniques developed and the results of their application are summarized in Section 1. A critical assessment of the methods proposed in this thesis is presented in Section 2. The areas where the present treatment is inadequate or has been neglected, are pointed out. These are areas where further research is obviously necessary, but it should be borne in mind that some of the problems indicated may be relatively intractable and hence their solution may not be immediately possible. In Section 3, some more clearly developed proposals for future research are discussed.

II.1. Summary of Results :-

A. Amplitude Dithering (Chapter III)

Methods of amplitude dithering for speech quantization with zero crossing preservation were developed. Two of the methods were found to improve the intelligibility of the quantized speech signals. The improvement over conventional fixed-level PCM was found to be equivalent to an increase in bit rate of approximately 1 bit/sample.

B. Residual Encoding (Chapter IV)

1) A lattice form of predictor structure was developed. This lattice form of predictor was found to be superior to the conventional direct-form of predictor in the convergence characteristics of its coefficients, the ease with which its stability can be maintained and its relative insensitivity to small fluctuations in the coefficient values. When used with a one-bit quantizer, the residual encoder with the lattice-form of predictor exhibits improvement in SNR of approximately 1 dB over the same encoder with the direct-form of predictor.

2) A strategy for step-size adaptation from the output bit stream was developed for a one-bit quantizer in a residual encoding scheme. Algorithms for step-size adaptation were accordingly derived for both the residual encoding with the lattice-form of predictor and with the direct-form of predictor. The one-bit quantizer enables the lower limit of output bit rate of a residual encoding scheme

to be reduced to approximately 1 bit per Shannon sample. The lattice form of predictor was also found to be particularly suitable for use at such low bit rates. With the combination of the lattice form of predictor and the one-bit quantizer, it was found possible to transmit speech at a rate of 1.17 bits per Shannon sample with a resultant SNR of about 10 dB.

C. Phase Dithering (Chapter V)

1) The technique of phase dithering was introduced to further improve the performance of the one-bit residual encoder. By spreading out the concentrated energy of pitch pulses in voiced speech, the adverse effect of a coarse quantizer in clipping the pitch pulses is reduced. When applied to a one-bit residual encoder with the lattice-form predictor the improvement in the SNR of the encoder was found to be greater than 1 dB. Longer phase dithering filters that spread out the energy of the pitch pulses over a longer time interval were found to give a greater improvement in the SNR performance of the encoder.

2) Phase restoration was found to be unnecessary unless the length of the phase dithering filter is longer than the duration of some short speech syllables. In this case, the short syllable (or acoustic event) appears to be muffled by the energy spilled over from the neighbouring acoustic events. An approximate phase restoration filter which is the phase dithering filter in reverse is developed for the reconstruction of the muffled acoustic events.

3) A set of simplified parameter adaptation algorithms for the one-bit residual encoder with the lattice-form of predictor were developed. This set of simplified algorithms adapts the parameters

by fixed quantum steps. With this simplification, the complexity of the encoder is reduced so that it is only approximately an order of magnitude greater than common adaptive delta-modulation systems. Although this operation results in some degradation in the performance of the encoder, when used with the phase dithering, the SNR of the simplified residual encoder is still better than the unsimplified one without phase dithering.

VI.2. Critical Discussions :-

The research work described in this thesis is mainly concerned with the development of new ideas in order to create a coding system that meets the set of bit rate, complexity and quality specifications laid down at the beginning of the research. Hence effort was devoted to the search for new techniques that could:

1) reduce the bit rate requirement

2) improve the reproduced speech quality

or 3) reduce the coder complexity.

The research has been successful in so far as having shown that the proposed new techniques summarized in the previous section are workable and the initial specifications can be met with combinations of several of the new techniques. In particular, the coding system using the combination of phase dithering, one-bit residual encoding with the lattice-form of predictor and simplified parameter adaptation algorithms is of high potential, and likely to see practical application. There are, nonetheless, many questions that have arisen during the course of this research. Also, it is considered that many areas have either been treated insufficiently, or have been omitted in the study. These are described briefly overleaf.

A. Intelligibility Measurements

The investigation into the various methods of amplitude dithering have demonstrated clearly that the intelligibility of quantized speech at identical, or almost identical, SNRs can be very varied. Although the method of fully automatic computer administration developed for this investigation can reduce the chores of subjective intelligibility testing to a minimum, it is still far too time consuming to be applicable for any purpose other than the verification of the intelligibility of a more or less finalized speech coding system. It is difficult to envisage it being used during development of a coding system to establish whether a certain small modification could be beneficial. A question that thus arises is whether it is possible to develop a more elaborate form of articulation index whose value approximately corresponds to actual subjective intelligibility. This articulation index could be an empirical function of all known impairments to intelligibility, e.g. the noise power, the amount of formant shift and the change in pitch frequency etc. caused by the speech coding system on a standard set of input speech utterances. In order to develop the empirical function, it would seem that a considerable amount of subjective testing would have to be performed. On the other hand, however, if such an empirical function could be developed then the measurement of the articulation index should easily be within the capability of most computing facilities and the task of deriving a rough guide to the intelligibility of a speech coding system would be much simplified.

B. Step-Size Adaptation Strategy for the One-Bit Residual Encoder

In Chapter IV, Section 3, it was pointed out that to optimize the step size of a one-bit quantizer in a residual encoder, an effective step-size adaptation strategy is to adjust the step-size so as to minimize the prediction error. Also it was demonstrated that, given a differential encoding network with a predictor that compensates for the poles in the input signal, then if the input signal is truly all-pole, minimizing the prediction error is equivalent to minimizing the quantization error (or, more precisely, maximizing the SNR of the quantizer). If the input consists of both poles and zeros, then the situation is more complicated, but it can be argued that if the energy introduced by the zeros is small compared with that introduced by the poles, then the quantizer SNR will not be far from the maximum, if the prediction error is at a minimum. This adaptation strategy is substantiated by its successful implementation and (see section IV.5) by the analogy that can be drawn between the step-size adaptation algorithm resulting from this strategy and the step-size adaptation algorithms of well known adaptive delta modulators. However, what exactly happens when the input signal consists of poles and zeros is not well understood. For example, a question that remains to be answered is: "If zeros are present in the input signal, what is the exact value to which the step-size will be adjusted for minimum prediction-error power?", and another question is "At this step-size, how far from the maximum will the quantizer SNR be?". It is also worth considering how the overall SNR of the complete differential encoder will be affected by the reduction in the quantizer SNR. Eq. I.1. in Chapter I shows that the overall SNR is the product of two terms. The first term, $\overline{S^2}/\overline{E^2}$, is improved by the reduction of the prediction error power $\overline{E^2}$. The second term $\overline{E^2}/\overline{N^2}$, is the SNR of the quantizer. It will be interesting to see, under what condition, the improvement in the first term due to the minimization of $\overline{E^2}$ would be able to compensate for the reduction in the SNR of the quantizer.

C. Parameter Adaptation Algorithms for the One-Bit Residual Encoder

This is an area in which a number of approximations and assumptions have been used; and in which much intuitive reasoning and many empirical modifications have been made. An "in depth" study of the parameter adaptation process will almost certainly lead to greater physical insight and improvement in the parameter adaptation algorithms. The interaction between the quantizer step-size and the predictor coefficient adaptation processes deserves particular attention. With the present algorithms, the predictor-coefficient adaptation is most efficient when the quantizer step-size is optimal; but, if the predictor coefficients are not optimal in the first place, the quantizer step-size also has difficulties in rapidly reaching the optimal value. An intuitive modification to the step-size adaptation algorithm by applying some stimuli to the step-size was found to improve the capability of the encoder in handling rapid changes in amplitude of the input signal. It would be interesting to see, in a more rigorous mathematical analysis, how this modification affects the overall optimization process. Also, the present step-size adaptation algorithms adapt the step-size from information about the polarity of one previous quantizer output. An algorithm, if it could be derived, that makes step-size adaptation decisions from a memory of more than one previous quantizer output, could well improve further the response of the step-size to rapid signal amplitude changes. Furthermore, in the implementation of the residual encoder, various ensemble averages in the formulation are estimated by time averages of various time constants (including the time constant of zero). These time constants are arrived at by intuitive reasoning coupled with an experimental trial and error approach. There is no reason to believe that the best possible choices have been arrived at. Again, a better understanding of the adaptation process may lead to developments in this respect.

D. Effects of Transmission Errors

The effects of transmission errors on the performance of the encoder have been omitted in this study. It seems likely that the methods of dithered quantization would be no more susceptible to transmission errors then normal fixed-level PCM. On the other hand, transmission errors could have a serious effect on the performance of the residual encoding schemes. There are two reasons for this: Firstly, the output bit stream from the one-bit residual encoder is highly non-redundant and hence can tolerate fewer errors. Secondly, errors in the transmission can cause the parameter adaptation processes in the encoder and decoder to lose synchronism thereby resulting in the propagation of the errors. The first reason is perhaps so fundamental that it may be very difficult to improve any degradation in performance due to this cause. In the case of the second category of transmission-error impairment, some protection is, however, built into the present parameter adaptation algorithms. With this protection, the residual encoding system gradually recovers, after a finite length of time, from the damage caused by a transmission error. However, the protection does interfere with the optimization of the predictor coefficients; and the faster the recovery, the slower will be the rate of convergence. For a given channel noise condition, it is not clear what is the best compromise between deterioration due to transmission errors and deterioration due to slow convergence. For example there could be a certain transmission error rate above which the recovery rate has to be so fast that predictor coefficients just could not converge to the optimal values. At and above such error rates, the sophistication of residual encoding schemes would clearly not be worth contemplating. Furthermore, the various types of residual encoding schemes, with the one-bit or multi-bit quantizer, with the direct-form or lattice-form of predictor, with or without phase dithering and with or without phase restoration could behave rather differently through the same noisy channel. Conversely, transmission channel noise of different properties, like Gaussian noise, impulsive noise tec., can also have different effects on a type of residual encoding scheme. This is an area that would need to be investigated extensively if the residual encoding methods should gain wide acceptance.

VI.3. Suggestions for Further Research :-

In this section, some topics which are felt to be of more immediate research interest are discussed. Some suggestions are also presented as to how the problems should be approached.

A. The Coding of Waveforms with Sharp Impulsive Spikes

In Chapter V, the problems associated with the coding of the prediction-error waveform from a voiced input and the resulting limitation on the performance of a residual encoder were discussed in detail. As was explained, it is the difficulty of designing a coarse quantizer that is equally effective(in terms of the quantization error) on both the pitch pulses and the residual fluctuations. This difficulty can be avoided with phase dithering which transforms the sharp pitch pulses to a more uniform distribution along the time axis. The method of phase dithering that was adopted has an advantage when applied to speech signals since phase restoration is not necessary, provided the phase-dithering filter is of short length. However, a longer phase dithering filter results in greater performance improvements in the residual encoder. In order to exploit this advantage, phase restoration has to be employed. The present sequence of phase dithering filter coefficients is not perfect in the restoration aspect and other sequences have to be sought.

With a simple search program, some short binary sequences have been found whose correlation function is a sequence of alternate ones and zeros except for the central peak, with the zeros occurring at odd delays. Such sequences would be ideal for phase dithering and restoration. However, it is not certain if a long binary sequence having these properties can be found. As the number of searches to be made increases geometrically with the length of the sequence, the initial search program is so time consuming that it cannot be used to explore sequence lengths greater than about 20, it is likely that a more detailed knowledge of the properties of such sequences would allow a more sophisticated program to be developed. Possibly, such a program would drastically reduce the search time and a greater knowledge of the properties of such sequences may result in a method for their generation.

In principle, the phase dithering technique is very similar to the smearing-desmearing technique[94] used to combat impulsive channel noise in data transmission. It may be that a modification of these techniques can be adopted directly for phase-dithering, but there is some doubt about this. In data transmission, certain non-linearities such as alternate sampling are permissible, whereas the phase dithering technique, as used in Chapter V, requires explicitly that the phase-dithering filter be linear. Possibly, a modification of the phase-dithering technique so as to use a non-linear filter may be a solution. In any case, the binary self orthogonal sequences* commonly employed in smearing-desmearing techniques are themselves better phase dithering sequences than the maximal length sequences in Chapter V. The auto-correlation functions of the self orthogonal sequences are always zero at even delays while having non-zero values at odd delays.

Another approach is to abandon phase dithering altogther and devise a special code for the spiky prediction errors. It may be that the special code could be developed using the ideas of run-length codes while making use of the property that with a voiced speech input, the spikes in the prediction error waveform occur at regular intervals. A major difficulty may arise from the fact that the statistics of the error waveform that result from a voiced input are quite different from those resulting from an unvoiced input. Hence, it may be necessary to develop two sets of optimal codes, one for each situation, and to have some mechanism by which the correct code book can be selected at both the encoder and decoder.

B. Hardware Implementation of the Simplified One-Bit Residual Encoder with Phase Dithering

At this stage of the research, it is felt that it would be sensible to attempt to gain some information about the practical viability of the simplified one-bit residual encoder with phase dithering. It is felt that very little difficulty would arise in developing a hardware realization of the encoder, and it is considered that it would be worthwhile to attempt to see just how simply

---

* See Chapter IV of Ref. (94).

the technique can be realised. The phase dithering filter can be implemented easily as a binary transversal filter* using an MOS dynamic shift register preceded by a delta modulator operating at high sampling frequencies. It is a pity that the lattice form of predictor filter cannot be readily implemented in the form of a binary transversal filter. This makes the main body of the encoder relatively more complicated than the phase-dithering filter. Perhaps a study could be conducted to find a possible way of implementing the lattice filter structure by using binary storage elements and a delta modulator as the A/D convertor. On the other hand, the implementation of the direct-form of predictor filter in the form of a binary transversal filter is straight forward. Hence, alternatively, attempts can be made to find a simple method of maintaining the stability of the encoder using the direct-form of predictor filter, thus making possible the application of this predictor structure in the simplified residual encoder. The direct-form of predictor structure has another implementation advantage in that it saves approximately two multiplications per stage of the predictor.

With the hardware realization, a number of important questions regarding this residual encoder can be investigated. Its subjective intelligibility, resistance to transmission channel errors and ability to cope with various kinds of speech utterances are all important considerations. An empirical study can also be made on the best procedure to take in setting the rate of adaptation of the encoder parameters. A parallel computer simulation and analytical study can also be conducted, say, to find a way of adjusting the quantizer step-size from a memory of more than one previous quantizer output. If this method of step-size adjustment is found, it can be imagined that the incorporation of this in the hardware encoder would be a simple matter.

---

* See Ref. (25)

APPENDIX

Quantization-Error Variance of the Various Amplitude Dithering Schemes with Zero-crossing Preservation

In Chapter III, it was shown that for any given input signal X, with the normal method of dithered quantization, the dither-quantized signal $(X)_D$, can take any value between $(X-\Delta/2)$ and $(X+\Delta/2)$ and the PDF of $(X)_D$ is uniformly distributed over this range. In Fig. A.1, the PDF of a normal dither-quantized signal $(X)_D$ is shown in relation to the position of X. It has also been demonstrated that the variance of the quantization error can be calculated from the PDF and is given by

$$E^2 = \frac{\Delta^2}{12}$$

which is equal to the quantization error variance obtained with the method of fixed-level quantization. The error variances of the various dithered quantization schemes with preservation of zero crossings are analyzed below in a similar manner.

A.1. Quantization Scheme 2

With this scheme the quantization-error PDF is identical to that of a normal dithered quantization system except when the signal X lies in the regions 0 to $-\Delta/2$ and 0 to $+\Delta/2$. If X lies between 0 and $+\Delta/2$, it will always be quantized to $+\Delta/2$. Thus after subtraction of the decoder dithering noise, the resultant output $(X)_{D2}$ will have a PDF as shown in Fig. A.2. Therefore, the quantization error, given X, is

$$E\Big|_X = (X)_{D2} - X$$

and
$$\langle E^2{}_X \rangle = \int_{-X}^{\Delta - X} E^2 \cdot p(E)dE$$

$$= \frac{1}{\Delta} \cdot \left[\frac{E^3}{3} \right]_{-X}^{\Delta - X}$$

$$= \frac{1}{3\Delta} \left(\Delta^3 - 3\Delta^2 X + 3\Delta X^2 \right)$$

Assuming that the probability of X lying on any point between 0 and $\Delta/2$ is constant, then

$$\langle E^2 \big|_{0 < X < \Delta/2} \rangle = \int_0^{\Delta/2} \langle E^2 \big|_X \rangle \cdot p(X) \Big|_{0 < X < \Delta/2} dX$$

$$= \frac{1}{3\Delta} \int_0^{\Delta/2} (\Delta^3 - 3\Delta^2 X + 3\Delta X^2) \cdot \frac{2}{\Delta} dX$$

$$= \frac{1}{6} \Delta^2$$

Similarly, if X lies in the range 0 to $-\Delta/2$

$$\langle E^2 \big|_{-2/\Delta < X < 0} \rangle = \frac{1}{6} \Delta^2$$

For any other values of X outside the range, $-\Delta/2$ to $+\Delta/2$, the quantization-error variance is the same as the normal-dithered-quantization case. That is,

$$\left\langle E^2 \middle| X < -\Delta/2,\ X > \Delta/2 \right\rangle = \frac{1}{12}\,\Delta^2$$

Thus, on the average, the quantization error variance of this scheme is higher than that of a fixed-level quantizer.

A.2. Quantization Scheme 3

With this scheme, the PDF of the processed signal is exactly the same as the normal-dithered case, except when the magnitude of the input signal X is such that

(1) $0 < |X| < \Delta/2$,

or (2) $\Delta/2 < |X| < \Delta$,

or (3) $\Delta < |X| < 3\Delta/2$

As above, the behaviour of the quantization error when X is negative is the mirror image of that when X is positive. Thus, it suffices to consider only the case when X is positive. The PDFs of the processed signal with X in the positive parts of each of these three regions are as shown in Fig. A.3. These PDFs can easily be derived from a consideration of the quantization procedure as described in Chapter III. The variances of the quantization error with X falling in the positive parts of each of the regions are computed as follows:

$$\left\langle E^2 \middle| 0 < X < \Delta/2 \right\rangle = \int_0^{\Delta/2} \left\langle E^2\ X;\ 0 < X < \Delta/2 \right\rangle \cdot p(X)\Big|_{0 < X < \Delta/2}\, dX$$

$$= \int_0^{\Delta/2} \left\{ \frac{1}{2} \cdot (\Delta/2 - X)^2 + \int_{-X}^{\Delta/2 - X} E^2 \cdot \frac{1}{\Delta}\, dE \right\} \cdot \frac{2}{\Delta}\, dX$$

$$= \frac{1}{16}\,\Delta^2 < \frac{1}{12}\,\Delta^2$$

$$\langle E^2 |_{\Delta/2<X<\Delta} \rangle = \int_{\Delta/2}^{\Delta} \langle E^2 |_{X;\ \Delta/2<X<\Delta} \rangle \cdot p^{(X)} |_{\Delta/2<X<\Delta}\ dX$$

$$= \int_{\Delta/2}^{\Delta} \left\{ \frac{1}{2} \cdot (X - \Delta/2)^2 + \int_{-\Delta/2}^{\Delta/2-X} E^2 \cdot \frac{1}{\Delta}\, dE + \int_{\Delta-X}^{\Delta/2} E^2 \cdot \frac{1}{\Delta}\, dE \right\} \cdot \frac{2}{\Delta}\, dX$$

$$= \frac{5}{48} \Delta^2 > \frac{1}{12} \Delta^2$$

$$\langle E^2 |_{\Delta<X<3\Delta/2} \rangle = \int_{\Delta}^{3\Delta/2} \langle E^2 |_{X;\ \Delta<X<3\Delta/2} \rangle \cdot p^{(X)} |_{\Delta<X<3\Delta/2}\ dX$$

$$= \int_{\Delta}^{3\Delta/2} \left\{ \left(\frac{3}{2} - \frac{X}{\Delta}\right)\left(\frac{\Delta}{2} - X\right)^2 + \int_{\Delta-X}^{\Delta/2} E^2 \cdot \frac{1}{\Delta}\, dE \right\} \frac{2}{\Delta}\, dX$$

$$= \frac{1}{6} \Delta^2 > \frac{1}{12} \Delta^2$$

It can easily be seen that if X falls in the negative parts of any of the above specified 3 regions, the resulting error variance is the same as the corresponding situation in the positive parts of the regions. Thus, as compared with normal fixed-level quantization, the error variance is decreased in region (1) and increased in regions (2) and (3); and is the same when X falls anywhere else. If it is assumed that the probability of X falling in region (1) is approximately the same as that of it falling in region (2), then the increase and decrease cancel approximately and hence the average error variance for these two regions combined is close to

$$\frac{1}{2} \left(\frac{1}{16} + \frac{5}{48}\right)\Delta^2 = \frac{1}{12} \Delta ,$$

which is the same as the case of fixed-level quantization. Hence, the overall error variance of this scheme is slightly higher than that of normal fixed-level quantization due to the greater contributions from region (3).

A.3. Quantization Scheme 4

With this scheme, the PDF of the processed signal also differs from that of a normal dithered quantizer, in exactly the same regions as scheme 3. However, as demonstrated in Fig. A.4, the natures of the PDFs in the three regions are not identical to the above. The variances of the quantization error in the positive parts of each of the regions are computed as follows:

$$\left\langle E^2 \middle|_{0<X<\Delta/2} \right\rangle = \int_0^{\Delta/2} (\Delta/2 - X)^2 \, \frac{2}{\Delta} \, dX$$

$$= \frac{1}{12}\Delta^2$$

$$\left\langle E^2 \middle|_{\Delta/2<X<\Delta} \right\rangle = \int_{\Delta/2}^{\Delta} \left\{ \left(\frac{3}{2} - \frac{X}{\Delta}\right)(\Delta/2 - X)^2 + \int_{\Delta-X}^{\Delta/2} E^2 \cdot \frac{1}{\Delta}\, dE \right\} \frac{2}{\Delta}\, dX$$

$$= \frac{1}{12}\Delta^2$$

$$\left\langle E^2 \middle|_{\Delta<X<3\Delta/2} \right\rangle = \int_{\Delta}^{3\Delta/2} \left\{ \left(\frac{3}{2} - \frac{X}{\Delta}\right)\left(\frac{\Delta}{2} - X\right)^2 + \int_{\Delta-X}^{\Delta/2} E^2 \, \frac{1}{\Delta}\, dE \right\} \frac{2}{\Delta}\, dX$$

$$= \frac{1}{6}\Delta^2 > \frac{1}{12}\Delta^2$$

With this quantization scheme, the quantization-error variance is identical to that of normal fixed-level quantization, if X should fall in regions (1) and (2), but is higher if X is in region (3). Thus the overall error variance is slightly increased.

A.4. Quantization Scheme 5

With this scheme, the PDF of the processed signal differs from that of normal dithered quantization only when X falls into the region of $-\Delta/2$ to $+\Delta/2$. The quantization-error variance for X in this region can be calculated by considering the positive part of the region only. Fig. A.5 shows the PDF of the processed signal when X is in the region 0 to $+\Delta/2$. The error variance is computed as follows:

$$\langle E^2 \big|_{0<X<\Delta/2} \rangle = \int_{\Delta}^{\Delta/2} \left\{ (\tfrac{1}{2} - \tfrac{X}{\Delta})(\Delta/2 - X)^2 + \int_{-X}^{\Delta/2} E^2 \cdot \tfrac{1}{\Delta} \, dE \right\} 2/\Delta \quad dX$$

$$= \frac{1}{12} \Delta^2$$

Thus the quantization error variance is the same as that of normal fixed-level quantization if X should fall in the range of 0 to $+\Delta/2$. The same is true if X is within the range of $-\Delta/2$ to 0. As the PDF of the processed signal is the same as that with normal dithering for all other values of X, the error variance is identical to that of normal dithered quantization. Thus the overall error variance of the scheme is the same as that of normal dithered or non-dithered quantization.

A.5. Quantization Scheme 6

With this scheme, the PDF of the processed signal is again only different from that of normal dithered quantization, when X is in the range of $-\Delta/2$ and $+\Delta/2$. The PDF, when X is in the positive half of this range is shown in Fig. A.6. The error variance is computed as follows:

$$\left\langle E^2 \middle| 0 < X < \Delta/2 \right\rangle = \int_{\Delta}^{\Delta/2} \left\{ \int_{-X}^{\Delta/2-2X} E^2 \cdot \frac{2}{\Delta}\, dE + \int_{\Delta/2-2X}^{\Delta/2} E^2 \cdot \frac{1}{\Delta}\, dE \right\} \frac{2}{\Delta}\, dX$$

$$= \frac{1}{16}\Delta^2 < \frac{1}{12}\Delta^2 .$$

The same is true if X falls in the range of $-\Delta/2$ to 0. For X anywhere else, the error variance is identical to that of normal dithered quantization. Thus the overall error variance is slightly less than that of normal dithered, or fixed-level quantization.

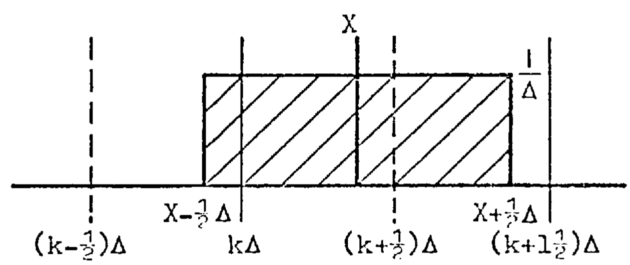


Fig. A.1. PDF of $(X)_D$ given X,

for a normal dithered-quantization-system.

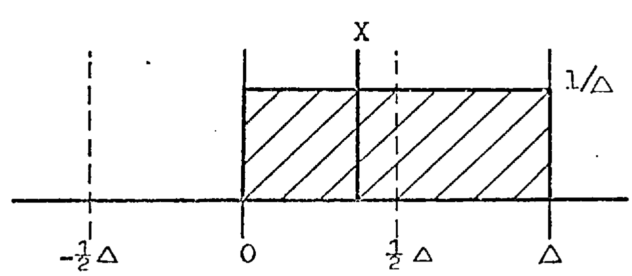
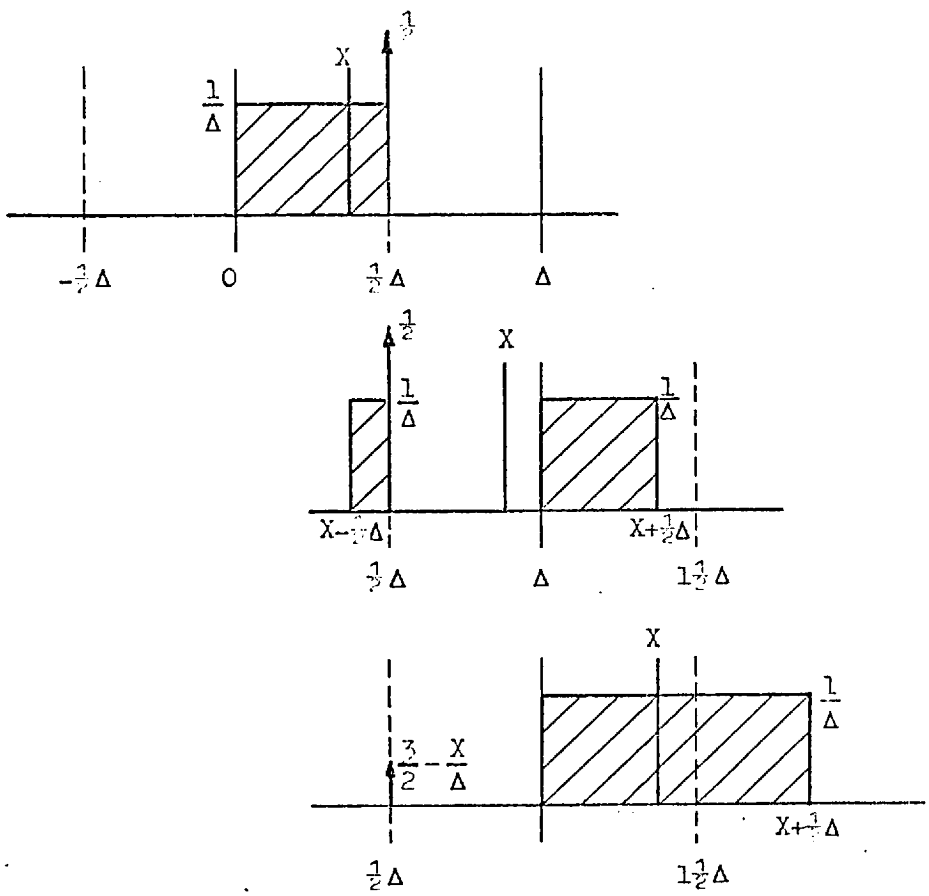


Fig. A.2. PDF of $(X)_{D2}$ given X,

for quantization scheme 2.



<u>Fig. A.3.</u> PDFs of $(X)_{D3}$ given X,

for quantization scheme 3.

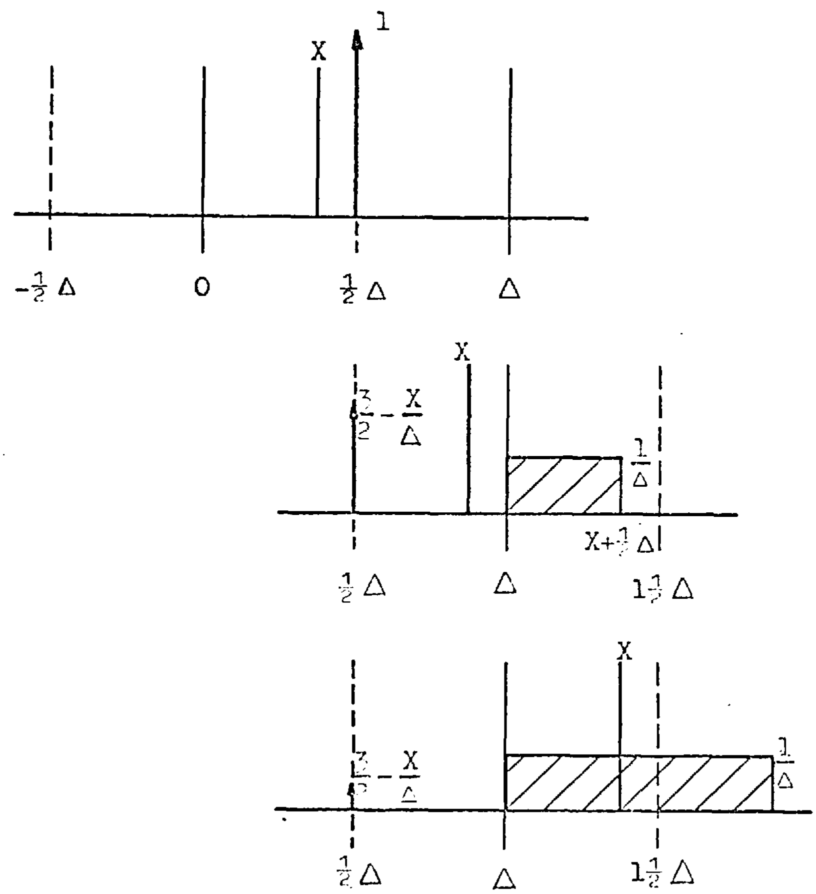


Fig. A.4. PDFs of $(X)_{D4}$ given X, for quantization scheme 4.

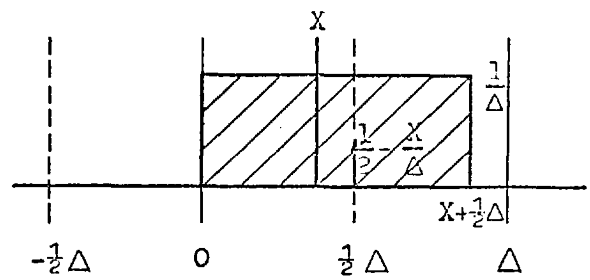


Fig. A.5. PDF of $(X)_{D5}$ given X, for quantization scheme 5.

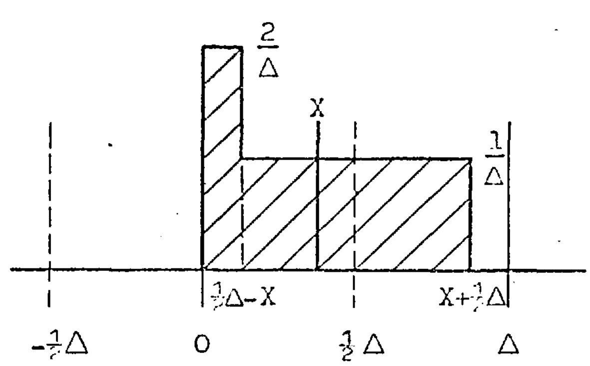


Fig. A.6. PDF of $(X)_{D6}$ given X, for quantization scheme 6.

REFERENCES

(1) JAYANT, N.S., "Digital Coding of Speech Waveforms: PCM, DPCM and DM Quantizers", Proc. IEEE, vol. 62, may 1974, pp. 611-632.

(2) RICHARDS, D.L., "Transmission Performance of Telephone Networks Containing PCM Links", Proc. IEE, vol. 115, Sept. 1968, pp. 1245-1258.

(3) CATTERMOLE, K.W., Principles of Pulse Code Modulation, Iliffe Books, London, 1969.

(4) PURTON, R.F., "A Survey of Telephone Speech-Signal Statistics and Their Significance in the Choice of a PCM Companding Law", Proc. IEE, vol. 109B, Jan. 1962, pp. 60-66.

(5) SMITH, B., " Instantaneous Companding of Quantized Signals", Bell Syst. Tech. J., 1957, pp. 653-709.

(6) MAX, J., " Quantizing for Minimum Distortion", IRE Trans. Inform. Theory, vol. IT-6, Mar. 1960, pp.7-12.

(7) JAYANT, N.S., " Adaptive Quantization with a One-Word Memory", Bell Syst, Tech. J., Sept. 1973, pp. 1119-1144.

(8) SCHLINK, W., " A Redundancy Reducing PCM System for Speech Signals", in Proc. Int. Surich Seminar Integrated Systems for Speech, Video and Data communications, Mar. 1972, pp. F41.1-F41.4.

(9) GOODMAN, D.J. and GERSHO, A., "Theory of an Adaptive Quantizer" IEEE Trans. Comm., vol. COM-22, Aug. 1974, pp. 1037-1045.

(10) WILKINSON, R.M., " An Adaptive Pulse Code Modulator for Speech", in Proc. IEEE Int. Conf. Communications (Montreal, Canada), June 1971, pp 1.11-1.15.

(11) JAYANT, N.S. and RABINER, L.R., "The Application of Dither to the Quantization of Speech Signals", Bell Syst. Tech. J., July-Aug. 1972, pp. 1293-1304.

(12) RABINER, L.R. and JOHNSON, J.A., "Perceptual Evaluation of the Effects of Dither on Low Bit Rate PCM Systems", Bell Syst. Tech. J., Sept. 1972, pp. 1487-1494.

(13) WOOD, R.G. and TURNER, L.F., "Pseudorandomly Dithered Quantization of Speech Samples inPCM Transmission Systems", Proc. IEE, vol. 119, May 1972, pp. 512-518.

(14) CHEN, M. and TURNER, L.F., "Zero-Crossing Preservation in The Low-Bit-Rate Pseudorandomly Dithered Quantization of Speech Signals", Proc. IEE, vol 122, Feb. 1975, pp. 119-123.

(15) DE JAGER, F., "Delta Modulation, a Method of PCM Transmission Using a 1-Unit Code", Philips Res. Pep., Dec. 1952, pp. 442-466.

(16) NIELSEN, P.T., "On the Stability of a Double Integration Delta Modulator", IEEE Trans. Comm. Tech., vol. COM-19, June 1971, pp. 364-366.

(17) O'NEAL, J.B.JR. and STROH, R.W., "Differential PCM for Speech and Data Signals", IEEE Trans. Comm., vol. COM-20, Oct. 1972, pp. 900-912.

(18) JAYANT, N.S., "Adaptive Delta Mudulation with a One-Bit Memory", Bell Syst. Tech. J., Mar. 1970, pp. 321-342.

(19) GREEFKES, J.A. and RIEMENS, K., "Code Modulation with Digitally Controlled Companding for Speech Transmission", Philips Tech. Rev., vol. 31, No. 11/12, 1970, pp. 335-353.

(20) WINKLER, M.R., "High Information Delta Modulation", in IEEE Int. Conv. Rec., pt. 8, Mar. 1970, pp. 321-342.

(21) GREEFKES, J.A. and DE JAGER, F. "Continuous Delta Modulation", Philips Res. Rep., vol. 23/2, 1968, pp. 233-246.

(22) CUMMISKEY, P., JAYANT, N.S., and FLANAGAN, J.L., "Adaptive Quantization in Differential PCM Coding of Speech", Bell Syst. Tech. J., Sept. 1973, pp. 1105-1118.

(23) JAYANT, N.S., "Adaptive Quantization with a One-Word Memory", Bell Syst. Tech. J., vol. 52, Sept. 1973, pp. 1119-1144.

(24) COHN, D.L. and MELSA, J.L., "The Relationship between an Adaptive Quantizer and a Variance Estimator", IEEE Trans. Inform. Theory, vol. IT-21, Nov. 1975, 669-671.

(25) STEELE, R., Delta Modulation Systems, Pentech Press, 1975.

(26) GOODMAN, D.J. and GREENSTEIN, L.J., "Quantization Noise of DM/PCM Encoders", Bell Syst. Tech. J., Feb. 1973, pp. 183-204.

(27) NOLL, P.W., "Nonadaptive and Adaptive Differential Pulse Code Modulation of Speech Signals", Overdruk Int Polytech. Tijdschr, Ed. Elektrotechnick/Electronica (The Netherlands) Nr. 19, 1972.

(28) DONALDSON, R.W., DOUVILLEI, R.J., "Analysis, Subjective Evaluation, Optimization and Comparison of the Performance Capabilities of PCM, DPCM, DM and PM Voice Communication Systems", IEEE Trans. on Commun. Tech. , vol. COM-17, 1969, pp. 421-431.

(29) ATAL, B.S. and SCHROEDER, M.R., "Adaptive Predictive Coding of Speech Signals", Bell Syst. Tech. J., vol. 49, Oct. 1970, pp. 1973-1986.

(30) GOLDBERG, A.J. and SHAFFER, H., "A Real-Time Adaptive Predictive Coder using Small Computers", Private Communication, also to be published in Trans. IEEE on Commun.

(31) DUNN, J.G., "An Experimental 9600 bits/s Voice Digitizer Employing Adaptive Prediction", IEEE Trans. on Commun. Tech., vol. COM-19, No. [illegible], Dec. 1971, pp. 1021-1032.

(32) SCAGLIOLA, C., "Automatic Vocal Tract Parameter Estimation by an Iterative Algorithm", in Proc. 8th International Congress on Acoustics, London, July 1974, vol. 1, p. 224.

(33) ASTROM, K.J. and EYKHOFF, P., "System Identification - a Survey", Automatica, vol. 7, 1971, pp. 123-162.

(34) GIBSON, J.D., JONES, S.K. and MELSA, J.L., "Sequentially Adaptive Prediction and Coding of Speech Signals", IEEE Trans. on Commun., vol. COM-22, Nov. 1974, pp. 1789-1797.

(35) STROH, R.W., "Optimum and Adaptive Differential PCM", Ph.D. dissertation, Polytech. Inst. Brooklyn, Farmingdale, N.Y., 1970.

(36) COHN, D.L. and MELSA, J.L., "The Residual Encoder - An Improved ADPCM System for Speech Digitization", IEEE Trans. on Comm., vol. COM-23, No. 9, Sept. 1975, pp. 935-941.

(37) FLANAGAN, J.L., Speech Analysis Synthesis and Perception, Springer-Verlag, 1965.

(38) HOLMES, J.N., Speech Synthesis, Mills and Boon, London, 1972.

(39) FLANAGAN, J.L. and RABINER, L.R. (Ed.), Speech Synthesis, Dowden, Hutchinson and Ross, Inc. 1973.

(40) FANT, G., Acoustic Theory of Speech Production, Mouton, 1970.

(41) VON BEKESY, G. Experiments in Hearing, McGraw Hill, 1960.

(42) UNWIN, P., "A Correlational Theory of Hearing", Hi-Fi News and Rec. Rev. Annual, 1974, pp. 42-49.

(43) STENNING, L.C. "Slow Fluid Waves in the Cochlea: Hypothesis on their Action in Pitch, Feedback and Lateralisation", Acustica, vol. 26, Heft 4, 1972, pp. 200-212.

(44) HOLMES, J.N., "The Influence of Glottal Waveform on the Naturalness of Speech from a Parallel Formant Synthesizer", IEEE Trans. on Audio and Electroacoust. vol. Au-21, No. 3, pp. 298-305.

(45) DUDLEY, H., "The Vocoder", Bell Lab. Record 17, 1939, pp. 122-126.

(46) VILKIG, F. and HAASE, K., "Some Systems for Speech-Band Compression", J. Acoust. Soc. Am., vol. 28, 1956, pp. 573-577.

(47) BAYLESS, J.W., CAMPANELLA S.J. and GOLDBERG A.J., "Voice Signals: Bit-by-Bit", IEEE Spectrum, Oct. 1973, pp. 28-34.

(48) SCHAFER, R.W., RABINER, L.R. and HERRMAN, O., "Uniform and Non-uniform Digital Filter Banks for Speech Analysis", in Proc. 8th Internation Congress on Acoustics, London, July, 1974, vol. 1, p. 242.

(49) DAVID, E.E., SCHROEDER, M.R., LOGAN, B.F. and PRESTIGIACOMO, A.J., "New Applications of Voice-Excitation to Vocoders", in Proc. Stockholm Speech Comm. Seminar, R.I.T., Stockholm, Sept. 1962.

(50) FLANAGAN, J.L., "Resonance-vocoder and Baseband Complement", IRE. Trans. on Audio, AU-8, 1960, pp. 95-102.

(51) BOGERT, B., HEALY, M. and TUKEY, J., in Time Series Analysis, Wiley, 1965, ch. 15.

(52) NOLL, A.M., "Short-time Spectrum and 'Cepstrum' Techniques for Vocal Pitch Detection", J. Acoust. Soc. Am., vol. 36, Feb. 1964, pp. 296-302.

(53) OPPENHEIM, A.V., SCHAFER, R.W. and STOCKHAM, T.G., Jr., "The Non-linear Filtering of Multiplied and Convolved Signals", Proc. IEEE, vol. 56, Aug. 1968, pp. 1264-1291.

(54) OPPENHEIM, A.V., and SCHAFER, R.W., "Homomorphic Analysis of Speech", IEEE Trans. Audio Electroacoust, vol. AU-16, June 19 8, pp. 221-226.

(55) OPPENHEIM, A.V., "A Speech Analysis-Synthesis System Based on Homomorphic Filtering", J. Acoust. Soc. Am., vol. 45, No. 2, 1969, pp. 458-465.

(56) FAKHRO, Y., "A Method for the Finite Discrete Spectral Analysis of Finite Real Power Systems Waveforms", Internal Research Rep., Imperial College, London, Dec. 1975.

(57) FANT, G., "On the Predictability of Formant Levels and Spectrum Envelopes from Formant Frequencies", in For Roman Jackobson, 'S-Gravenhage, 1956, pp. 109-120.

(58) FLANAGAN, J.L., "Automatic Extraction of Formant Frequencies from Continuous Speech", J. Acoust. Soc. Am., vol. 28, 1956, pp. 110-118.

(59) FLANAGAN, J.L. and HOUSE, A.S., "Development and Testing of a Formant-Coding Speech Compression System", J. Acoust. Soc. Am., vol. 28, 1956, pp. 1099-1106.

(60) COKER, C.H., "Computer-simulated Analyzer for a Formant Vocoder", J. Acoust, Soc. Am., vol. 35, Oct. 1963, p. 1911.

(61) JUDD, M.W. A Seminar given in Imperial College, Nov. 1974.

(62) MARKEL, J.D., "Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation", IEEE, Trans. Audio Electroacoust., vol. AU-20, 1972, pp. 129-137.

(63) ATAL, B.S. and HANAUER, S.L., "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", J. Acoust. Soc. Am., vol. 50, No. 2 (part 2), Aug. 1971, pp. 637-655.

(64) CHANDRA, S. and LIN, W.C., "Experimental Comparison between Stationary and Nonstationary Formulations of Linear Prediction Applied to Voiced Speech Analysis", IEEE Trans. on Acoust. Speech and Signal Processing, vol. ASSP-22, No. 6, Dec. 1974, pp. 403-415.

(65) ITAKURA, F. and SAITO, S., "On the Optimum Quantization of Feature Parameters in the PARCOR Speech Synthesizer", in IEEE Proc. Conf. Speech Commun. Process, 1972, pp. 434-437.

(66) WAKITA , H., "Estimation of the Vocal Tract Shape by Optimal Inverse Filtering and Acoustic Articulatory Conversion Methods", Ph.D. dissertation, Univ. California, Santa Barbara, June 1972; also see Speech Commun. Res. Lab., Santa Barbara, Calif. SCRL Monograph 9, July 1972.

(67) MARKEL, J.D. and GREY, A.H., Jr., "On Auto-correlation Equations as Applied to Speech Analysis", IEEE Trans. Audio and Electroacoust., vol. AU-21, No. 2, April, 1973, pp. 68-79.

(68) GIBSON, J.D., MELSA, J.L. and JONES, S.K., "Digital Speech Analysis using Sequential Estimation Techniques", IEEE Trans. Acoust., Speech and Signal Process., vol. ASSP-23, No. 4, Aug. 1975, pp. 362-369.

(69) MAKHOUL, J., "Spectral Analysis of Speech by Linear Prediction", IEEE Trans. Audioelectroacoust., vol. AU-21, June 1973, pp. 140-148.

(70) MAKHOUL, J., "Spectral Linear Prediction : Properties and Applications", IEEE Trans. Acoust., Speech and Signal Process., vol. ASSP-23, No. 3, June 1975, pp. 283-296.

(71) KNUDSEN, M.J., "Real-time Linear-Predictive Coding of Speech on the SPS-41 Triple-Microprocessor Machine", IEEE Trans. on Acoust., Speech and Signal Process, vol. ASSP-23, Feb. 1975, pp. 140-145.

(72) MORRIS, L.R. and GOUGH, H.S., "An Economical Hardware Realization of a Digital Linear Predictive Speech Synthesizer" in Proc. IEEE Int. Conf. Communications (Seattle, Wash. June 1973), pp. 46-23 - 46-28.

(73) PARPOULIS, A., Probability, Random Variables and Stochastic Processes, McGraw-Hill, 1965, p. 483.

(74) WILKINSON, R.M., A talk given in the Colloquium on Differential Quantizing systems, Loughborough University, Oct. 1975. Refer also to - "An Adaptive Pulse Code Modulator for Speech Signals" SRDE Report 72001, Jan. 1972.

(75) TURNER, L.F., A Talk given in the Colloquium on Differential Quantizing Systems, Loughborough University. Refer also to - "Application of Data Compression to an Experimental 9.6 Kbits/s Adaptive PCM Digital Speech System", Proc. IEE., vol. 123, No. 2, Feb., 1976, pp. 109-113.

(76) FREI, A.H., SCHINDLER, H.R., VETTIGER, P., and VON FELTEN, E., "Adaptive Predictive Speech Coding Based on Pitch-Controlled Interruption/ Reiteration Techniques", in Proc. IEEE Int. Conf. Communications (Seattle, Wash. June 1973), pp. 46.12 - 46.16.

(77) AGRAWAL, A. and Lin, W.C., "An On-Line Speech Intelligibility Measurement System", IEEE Trans. on Acoust., Speech and Signal Process., vol. ASSP-22, June 1974, pp. 203-206.

(78) COOLEY, J.W. and TUKEY, J.W., "An Algorithm for the Machine Computation of Complex Fourier Series", Math. Comput., vol. 19, April 1965, pp. 297-301.

(79) GOLD, B. and RADAR, C.M., Digital Processing of Signals, McGraw-Hill, 1969.

(80) SILVERMAN, H.F. and DIXON, N.R., "A Parametrically Controlled Spectral Analysis System for Speech", IEEE Transaction on Acoustics, Speech and Signal Processing, vol. ASSP-22, Oct. 1974, pp. 362-381.

(81) GOODMAN, D.J., GOODMAN, J.S. and CHEN, M., "Relationship of Intelligibility and Subjective Quality of Digitally Coded Speech", Paper presented at the 90th Meeting of the Acoust. Soc. of Am., San. Franc., Nov. 1975.

(82) LOPRESTI, P.V. and SURI, H.L., "A Fast Algorithm for the Estimation of Autocorrelation Functions", IEEE Trans. on Acoust., Speech and Signal Process., vol. ASSP-22, Dec. 1974, pp. 449-453.

(83) ROBERTS, L.G. "Picture Coding Using Pseudo-Random Noise", IRE Trans. on Inform. Theory, vol. IT - 8, Feb. 1962, pp. 145-154.

(84) LICKLIDER, J.C.R., BINDRA, D., and POLLACK, I., "The Intelligibility of Rectangular Speech Waves", Am. J. Psych., vol. 51, 1948, pp. 1-20.

(85) LICKLIDER, J.C.R., "The Intelligibility of Amplitude Dichotomized, Time Quantized Speech Waves", J. Acoust. Soc. Am., vol. 22, 1950, pp. 820-823.

(86) MORRIS, L.R., "The Role of Zero Crossings in Speech Recognition and Processing", Ph.D. thesis, Imperial College, London, 1970.

(87) HAMMING, R.W. Numerical Methods for Scientists and Engineers, McGraw-Hill, 1962, pp. 384-388.

(88) HOUSE, A.S., WILLIAMS, C.E., HECKER, M.H.L. and KRYTER, K.D., "Articulation Testing Methods : Consonant Differentiation with a Closed-Response Set", J. Acoust. Soc. Am., vol. 37, Jan. 1964, pp. 158-166.

(89) GUENTHER, W.C., Analysis of Variance, Prentice Hall, 1968.

(90) EDWARDS, A.L., Experimental Design in Psychological Research, Holt International, 1970, pp. 131-135.

(91) ITAKURA, F., SAITO, S., KOIKE, T., SAWABE, H., and NISHIKAWA, M., "An Audio Response Unit Based on Partial Autocorrelation", IEEE Trans. Comm., vol. COM-20, Aug. 1972, pp. 792-797.

(92) AOKI, M., Introduction to Otpimization Techniques, The Macmillan Company, 1971, p. 147.

(93) TOMOJAWA, A., and KANEKO, H., "Companded Delta Modulation for Telephone Transmission", IEEE Trans. Commun. Tech., vol. COM-16, Feb. 1968, pp. 149-157.

(94) ESTEVES, N.L., "Data Transmission through Channels Perturbed by Impulsive Noise", Ph.D. Thesis, Imperial College, London, 1975.