LOW BIT RATE DIGITAL SPEECH SIGNAL PROCESSING SYSTEMS

by

SAEED AHMADI, B.SC. (HON), M.Sc., DIC.

September 1980

A Thesis submitted for the
Degree of Doctor of Philosophy
in the Faculty of Engineering,
University of London.

DEPARTMENT OF ELECTRIC , ENGINEERING.
IMPERIAL COLLEGE OF SCIEN : AND TECHNOLOGY,
UNIVERSITY OF LONDON.

ABSTRACT

The purpose of the research is to develop relatively simple to implement and efficient systems for low bit rate digital speech transmission. The approach taken for the work contained in this thesis is to sample speech signals at a regular rate and examine successive groups of samples for the extraction of information bearing attributes that are characteristic of each group. The groups of samples are all taken to be of a fixed length. The information bearing attributes are then extracted from a transformed version of the time domain segments of the speech signal where there are essentially either conventional formants as defined in the frequency domain or similar "formant" behaviour as exhibited in other orthogonal domains for the voiced sounds. Amongst the variety of different orthogonal transforms that can be employed to this end, the Walsh-Hadamard is one of the most desirable transforms particularly when the final systems are to be realized in a digital form. The following three alternative systems arise from the study in an evolutionary way.

(i) From the formant approach, a simple and efficient system is proposed which produces intelligible digital speech transmission with rates of the order of 4 K bits/sec. The principle of the system is based on the fact that the perceptually important parameters of the voiced speech segments in time are formants which are obtained from the Walsh-Hadamard domain in this system where at the receiver end the voiced speech is recovered by replacing the

formants at their appropriate positions and inverse transformation of each block.

The unvoiced sound segments are coded according to their statistical behaviour.

(ii) Further developments are achieved on the above system by combining the well-known linear prediction methods with the formant approach in the transform domain and from such considerations a system is proposed and examined. The principal idea in this system is the fact that formants show high degree of correlation among themselves and hence the linear prediction method is used to exploit this correlation and predict higher formants from a knowledge of the first one, in the transform domain. The unvoiced sounds are processed in accordance with their statistical behaviour. Intelligible digital speech transmission is achieved with bit rates of around 2 Kbits/sec.

(iii) The above two systems show the importance of formants in the perception of the voiced sounds. The formants are also of great importance in the recognition of vowels, The major part of the voiced sounds, and because the complete set of voiced sounds includes not only vowels but voiced consonant-like utterances, a technique is developed for the classification of all the voiced sounds based on their formant positions. A system is designed and simulated using this technique leading to successful digital speech transmission rates of approximately 1 K bits/sec.

ACKNOWLEDGEMENTS

The author wishes to express his gratitude to his supervisor, Dr. A.G. Constantinides of the Electrical Engineering Department, Imperial College, for his encouragement, help and advice during the course of the work.

Thanks are due to the staff and colleagues of the Communication Section who have been helpful in many respects, in particular, Dr. M.R. Ashouri and Mr. L.F. Mujica for their useful comments on the subject, and Mr. R. Sulley for his assistance with the computer.

Particular thanks are due to the late Professor Colin Cherry, the founder of the Communication Section, and the late Dr. W. Saraga, the recent deaths of whom were a great loss to the section. Also thanks are due to Mrs. Jo Vogt for her skillful typing of this thesis.

Finally, the author's deepest gratitude is due to his parents for their moral and financial support during his studies.

C O N T E N T S

PAGE

TITLE PAGE 1

ABSTRACT 2

ACKNOWLEDGEMENTS 4

CONTENTS 5

LIST OF FIGURES 10

LIST OF TABLES 13

Chapter One: INTRODUCTION

1.1. Speech production..... 14

1.2. Engineering Approach and definition of the problem. 16

1.3. Aim of the Project. 17

1.4. Thesis structure..... 20

PAGE

Chapter Two: Review of the work in speech processing systems

2.1. History of speech processing..... 24
2.2. Different speech processing systems..... 28
2.3. Time Domain Systems..... 31
2.3.1. Pulse Code Modulation (PCM)..... 31
2.3.2. Differential Pulse Code Modulation (DPCM)..... 33
2.3.3. Delta-Modulation (DM)..... 34
2.3.4. Adaptive Predictive Coding (APC)..... 36
2.4. Transform Domain Systems..... 37
2.4.1. Channel Vocoder..... 37
2.4.2. Formant Vocoder..... 39
2.4.3. Sub-band Coder..... 39
2.4.4. Adaptive Transform Coding 41
2.5. Linear Predictive Coding 42
2.6. Comments on different systems 46

Chapter Three: Signal Processing in the Transform Domain.

3.1. Transform Domain Analysis of the Signals..... 58
3.2. Orthogonal Transform Domains..... 60

PAGE

3.2.1. Karhunen - Loève transform (KLT)..... 60

3.2.2. Discrete Cosine Transform (DCT)..... 61

3.2.3. Discrete Fourier Transform (DFT)..... 63

3.2.4. Walsh-Hadamard Transform (WHT)..... 65

3.2.5. Haar Transform (HT)..... 70

3.2.6. Slant Transform (ST)..... 72

3.2.7 Number Theoretic Transform (NTT)..... 74

3.3. Comparison of Transform Domains..... 75

3.4. Speech Signals in the Transform Domain..... 75

3.5. Special Features of the Voiced Sounds..... 76

Chapter Four: Digital Speech Processing based on Formants.

4.1. Formants in the transform domain..... 87

4.2. Different formant detection methods in the transform domain..... 87

4.2.1. Peak Picking Approach..... 89

4.2.2. Analysis by synthesis approach..... 91

4.2.3.. The Spectral Method..... 91

4.2.4. Moments Approach..... 92

4.2.5. The Linear Predictive Method..... 93

4.3. Speech Processing Systems based on Formants..... 94

4.4. A new Formant based system for speech processing..... 96

4.4.1. The choice of Transform Domain..... 96

4.4.2. Similarity of Walsh and Fourier Transforms. 98

4.4.3. Block Size Decision..... 99

		PAGE
4.4.4	Voiced-Unvoiced Decision.....	100
4.4.5.	Formant Detection Technique.....	103
4.4.6.	Unvoiced Sound Processing Method.....	104
4.4.7.	Coefficient Quantization.....	105
4.4.8.	Overall System Description.....	107
4.4.9.	Computer Simulation Results.....	109

Chapter Five: Linear Prediction of Formants for Speech Processing.

5.1.	Correlation between Speech Samples in the Transform Domain.....	120
5.2.	Formant Correlation in the Transform Domain.....	122
5.3.	Linear Prediction Techniques in the Transform Domain.....	124
5.4.	Linear Prediction of Formants.....	128
5.5.	Computer Simulation Results of the Formant Prediction System.....	130

Chapter Six: Clustering Approach for Speech Processing Systems.

6.1.	Introduction.....	145
6.2.	Vowels and Consonants.....	146
6.3.	Formant Transitions in Consonant Sounds.	147
6.4.	Voiced Sounds Short time Spectrum.....	149
6.4.1.	Formation of the Library for Spectrum Shapes.....	151
6.5.	Comparison of Spectrum Shapes, "The best choice".....	152
6.5.1.	Least Mean Square Error Approach.....	153

PAGE

6.5.2. Moment Approach..... 154

6.6. Speech processing System based on Clustering Approach,..... 156

6.6.1. Walsh-Hadamard domain System Approach..... 157

6.6.2. Frequency domain System Approach..... 164

Chapter Seven: Conclusions and Suggestions for Further Research.

7.1. Conclusions..... 181

7.2. Suggestions for further Research..... 185

REFERENCES 187

LIST OF FIGURES: PAGE

Fig. 1.1. Anterior - Posterior Section of the Larynx..... 22

Fig. 1.2. The vocal organs..... 23

Fig. 2.1. A PCM communication system..... 48

Fig. 2.2. Block diagram of a DPCM System..... 49

Fig. 2.3. A delta-Modulation system of Communication..... 50

Fig. 2.4. Adaptive predictive Coding System..... 51

Fig. 2.5. Block diagram of a typical channel vocoder..... 52

Fig. 2.6. Block diagram of a Formant Synthesizer..... 53

Fig. 2.7. Block diagram of a Variable-band Coder..... 54

Fig. 2.8. Structure of an ATC system..... 55

Fig. 2.9. Linear Prediction Vocoder System..... 56

Fig. 3.1. Block diagram of transform coder..... 78

Fig. 3.2. Walsh functions of order 8..... 79

Fig. 3.3. Signal flow graph of fast Walsh Computation..... 80

Fig. 3.4. Slant-transform basis waveforms..... 80

Fig. 3.5. Transform domain behaviour for different speech sounds..... 81

Fig. 3.6. Typical Spectrum shapes for Voiced Sounds in Frequency and Walsh domains..... 82

Fig. 3.7. Typical spectrum shapes for unvoiced sounds in frequency

Page

and Walsh domains..... 83

Fig. 4.1. F_1-F_2 Space for different vowels..... 114

Fig. 4.2. Block diagram of a LPC dependent method for determining Formant Frequencies and Amplitudes..... 115

Fig. 4.3. Energy distribution of Walsh coefficients for different threshold levels..... 116

Fig. 4.4. Block diagram of the Formant Sequency System..... 117

Fig. 4.5. Time domain representation of part of the original and the reconstructed speech signal.using the formant sequency approach..... 118

Fig. 4.6. The Walsh domain representation of part of the original and the reconstructed speech signal using the formant sequency approach..... 119

Fig. 5.1. Representation of part of the Average Covariance Matrix obtained from the Walsh-Hadamard domain..... 137

Fig. 5.2. Probability distribution graph of the first, second and third formant locations..... 138

Fig. 5.3. Conditional probability distribution graph of the second and third formant locations for given first formant locations..... 139

Fig. 5.4. Block diagram of the "Linear Prediction of formants" System..... 140

Page

Fig. 5.5. Time domain representation of part of the original and reconstructed speech signal using"the linear prediction of of formants" approach..... 141

Fig. 5.6. The Walsh domain representation of part of the original and reconstructed speech signal using the "Linear Prediction of formants" approach..... 142

Fig. 6.1. $F_1 - F_2$ graph for vowel identification... 170

Fig. 6.2. Typical formant transitions in the case of consonant voiced sounds,..... 171

Fig. 6.3. Formant Transitions with different slopes and their staircase approximations for the consonant sounds,..... 172

Fig. 6.4. A tree representation of the library spectrum shapes,..... 173

Fig. 6.5. Block diagram of the clustering system based on the Walsh domain..... 174

Fig. 6.6. Time domain representation of part of the original and reconstructed speech signal using the clustering approach..... 175

Fig. 6.7. The Walsh domain representation of part of the original and the reconstructed speech signal using the clustering approach,..... 176

PAGE

Fig. 6.8. The illustration of the symmetrical real input method for FFT usage..... 177

Fig. 6.9. Block diagram of the clustering system based on the frequency domain..... 178

Fig. 6.10. Time domain representation of part of the original and reconstructed speech signal using the clustering approach (Frequency domain case)..... 179

Fig. 6.11 The frequency domain representation of part of the original and reconstructed speech signal using the clustering approach..... 180

LIST OF TABLES:

Table 2.1. Merits of different speech coding systems..... 57

Table 3.1. Comparison of Transform domains..... 84

Table 3.2. List of Vowels and Consonants..... 86

Table 5.1. Weighting coefficients for the 64 point Walsh-Hadamard domain..... 143

Table 5.2. Prediction table of speech in the Walsh-Hadamard domain..... 144

CHAPTER ONE

I N T R O D U C T I O N

1.1 SPEECH PRODUCTION

Speech is the most natural means of communication between humans. Only man has developed the vocal means for coding and conveying information beyond an elementary stage. Although speech is a familiar means of communication, it is nevertheless a complex process with many aspects to be studied. It can be considered as a process containing various aspects of communication such as the intentions of the speaker, the movements of the vocal organs, transmission of the speech sounds, detection of the sounds by the hearing organs of the listener, and the listener's interpretation, and many others.

Acoustically speech may be considered as consisting of two types of sounds, namely voiced and unvoiced sounds.

The class of voiced sounds is produced by passing air from the speaker's lungs through the larynx, a passage that contains the vocal cords. A cross-section of the larynx is shown in Figure 1.1. As air is passed through the vocal cords, they vibrate causing bursts of air to escape into the aural cavity which consists of the throat, nasal cavity and mouth as shown in Figure 1.2.

By a combination of muscular tension in the cords and the lowering pressure in the glottis, the vocal cords are drawn back to their starting position and the air flow ceases. The sub-glottal pressure then forces the cords

apart again and the whole cycle is repeated. It can be seen therefore that the vocal cords act as an intermittent barrier to the flow of air from the lungs, and in fact chop the air stream so that a discrete set of bursts is produced.

The vocal cord vibration period is a function of vocal cord mass, tension and sub-glottal pressure. For normal male talkers these bursts of air are produced with a frequency referred to as fundamental frequency typically in the range of 50-250 Hz and extend to about 500 Hz and higher for women and children.

Harmonics of this fundamental frequency are expected to be present in the spectrum of voiced sounds. Due to the irregular shape of the aural cavity, peaks at distinct points of the spectrum can be observed.

The other class is the unvoiced class of sounds during which the vocal cords are held wide apart and the air stream from the lungs is forced through a constriction between the tongue and the teeth. The unvoiced sounds do not exhibit the harmonic structure as the voiced sounds, The unvoiced spectrum has in general noise-like energy concentration.

Voiced sounds include all the vowels and some of the consonants, whereas the unvoiced sounds include only some consonants. (A list of vowels and consonants are given in Section 3.5). The unvoiced sounds are important to the intelligibility of speech and in general it can be concluded that speech is the continuous production of voiced and unvoiced sounds with appropriate pauses to add clarity and distinctness.

1.2. ENGINEERING APPROACH AND DEFINITION OF THE PROBLEM

Speech can be considered from different points of view. Linguists, physiologists, phoneticians and engineers would consider the process of speech in different ways. From the point of view of a communication engineer speech is an analogue signal having bandwidth that extends from the low frequency audio range up to 8 to 10 KHz; channel bandwidth of the same order is then required for the direct transmission of a speech waveform. However, lower bandwidths can be used for transmission of fairly good quality speech if the redundancies inherently present in the signal are taken into account.

The digital transmission of the speech waveform directly, using the Pulse Code Modulation method for example, requires rates of 20 K bits/sec to 100 K bits/sec. depending on the quality of speech required.

Reduction of the transmission bandwidth, however can be achieved by considering the fact that the speech waveform is produced by relatively slow movements of the articulatory tract, and these movements can be described, with sufficient accuracy, by signals of much lower information rate than the expected rate of transmitting the original acoustic signal. Then in order to achieve low bandwidth transmission, the extraction of information-bearing parameters from the speech waveform is necessary and can be obtained by studying the production of speech and its behaviour in different domains. These parameters can then be transmitted with a much lower bandwidth, or equivalently with a much lower bit rate in the case of digital systems.

At the receiver speech can be reconstructed using these parameters usually with some degradation in quality but without any serious loss in intelligibility.

1.3 AIM OF THE PROJECT

Digital Speech Processing systems available can be divided into two major catagories: those with simple system structure but high transmission-rate requirements, such as PCM, DPCM, DM with bit rates higher than 16 K bits/sec and those that require lower transmission-rates but much more complex structures, such as vocoders, with bit-rates around 1·2 K bits/sec.

The aim of the work contained in this thesis is to investigate alternative possibilities for systems with simpler structures but transmission rates similar to the second type of speech processing system mentioned above.

To achieve these aims, the signal processing systems will be considered to be in digital form for the obvious advantages over their analogue counterparts, such as less sensitivity to noise, cross-talk, distortion and more efficiency in switching, multiplexing, security and other advantages.

The parameters considered to be vital for the perception of the voiced sounds are the resonances of the vocal tract which are referred to as Formants. To obtain these formants for a specific voiced segment of speech a transform domain analysis of the speech signal is necessary rather than a time domain analysis which in addition to

highlighting the resonances, offers a non-uniform energy distribution for the speech signal and easy detection of the desired parameters. Different orthogonal transforms can be used in this respect. Amongst them the Walsh domain is more attractive due to the simplicity of implementation for digital systems. Frequency domain techniques are also used in the design of such systems.

In any speech processing system the perceptually important parameters of speech sounds are chosen and used to achieve an efficient processing method for low bit rate intelligible speech transmission. Such parameters have proved to be the formants and regions around them, that is to say the distinct peaks obtainable from the transform domain spectrum of speech voiced sounds. The aim of the investigation presented in this thesis is to develop simple and efficient systems for speech processing leading to low bandwidth (or equivalent low bit-rate requirement) processing. This could be best achieved by making use of the speech characteristics to their fullest extent for removing redundancies. Based on the above principles a system is proposed in Chapter 4 operating in the Walsh-Hadamard domain for simplicity and efficiency and achieving intelligible low bit rate digital speech transmission. Further developments can be made by using another well-known method for removing the redundancies of the speech signals, i.e. that of linear prediction (53,54).

The linear prediction methods have been used in this context essentially in the time domain. The possibilities of using this method in transform domain in combination with formants is studied in the investigation contained in Chapter 5. This combination appears to be useful for the removal of speech redundancies in a more efficient way. A system, then, based on the above principles of linear formant prediction is proposed and assessed, where intelligible low bit rate speech transmission is achieved. In essence the above methods are based on the fact that formants play a crucial role in the perception of voiced sound. However, they are also of paramount importance in the recognition of vowels, which in fact form the major part of the voiced sounds in speech.(75).

The complete set of voiced sounds, however includes not only vowels but voiced consonant like utterances hitherto unexamined. Such an examination and study is carried out in Chapter 6 and techniques are developed for the classification of these sounds based on the formant positions.

This pattern recognition approach is shown to be possible provided that short segments of voiced sounds are considered for such classifications.

A system based on this approach is then designed and studied leading to successful very low transmission rate requirements with intelligible speech recovery.

1.4 THESIS STRUCTURE

The aim of the work presented here is to consider the design of digital speech processing systems having simple structures and low information rate requirements.

To achieve this, investigations have been carried out, the results of which are arranged in this thesis as follows:

In Chapter 2, a brief review of the existing speech processing systems is given in which time domain approaches such as Pulse Code Modulation (PCM), Differential Pulse Code Modulation (DPCM), Delta Modulation (DM) and Adaptive Predictive Coding (APC) are reviewed, and transform domain based systems such as Channel Vocoders, Formant Vocoders, Sub-band Coder, and Adaptive Transform Coding (ATC) as well as linear Predictive Coding (LPC) are briefly discussed.

In Chapter 3, signal processing in the transform domain is considered and its advantages are discussed. Different transform domains such as Karhunen-Lòeve (KL), Discrete Cosine (DC), Fourier, Walsh-Hadamard (WH), Haar (H), Slant (S) and Number Theoretic (NT) transforms are reviewed. The presence in such transformed domains of different types of sounds (voiced/unvoiced) is discussed together with the existence of formants in these domains.

Chapter 4 is devoted to the behaviour of formants in transform domains. Different methods of detecting formants is presented and discussed in this chapter. Then a simple system based on the Walsh domain is proposed and simulated results for this approach are discussed.

In Chapter 5 another approach is presented based on the well-known method of Linear Prediction Coding in the transform domain in combination with formants.

A system based on this approach is designed and simulated in achieving a better digital speech processing. The description and performance of such a system is given in this chapter.

Chapter 6 describes a novel approach based on short time spectrum classification of speech signal segments from the formant locations point of view. A digital low bit-rate speech processing system using this approach is designed and simulated. This is done both in the frequency and the Walsh domains.

Finally in Chapter 7 the results stemming from the studies in Chapters 4, 5 and 6 are assessed. Further research suggestions on similar lines are given in this chapter.

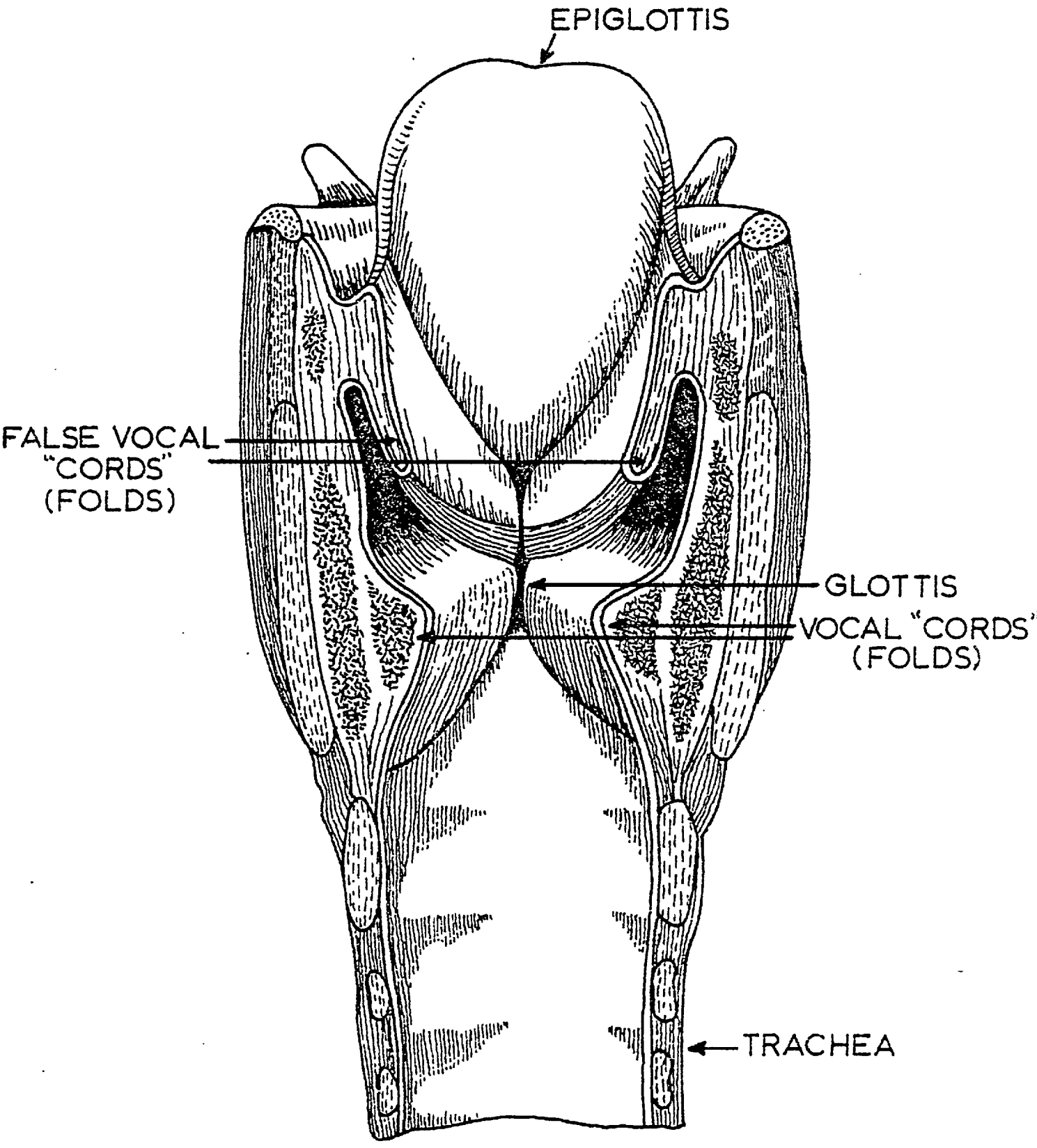


Fig.1.1: Anterior-Posterior section of the Larynx

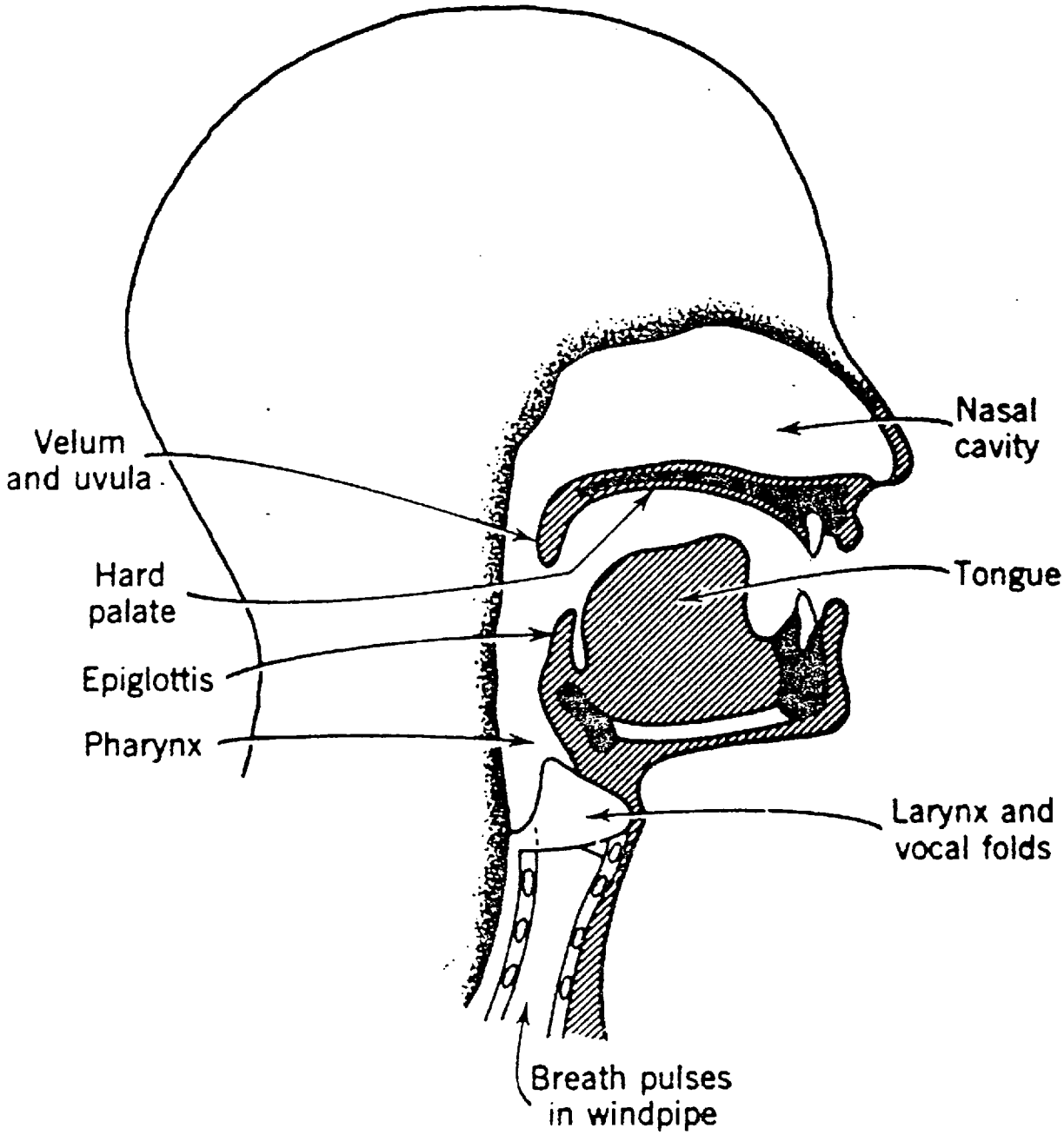


Fig. 1.2: The Vocal Organs

CHAPTER TWO

REVIEW OF SPEECH PROCESSING SYSTEMS

2.1. HISTORY OF SPEECH PROCESSING

Speech has played a dominant role in the development of human civilisations. For many years man was able to communicate verbally through short distances only, due to rapid diminuation of acoustic energy in the atmosphere.

It was only recently (about 100 years ago) that by converting the acoustical energy into electrical form that man could extend the limitations which were imposed on him by nature.

A brief survey of such electrical speech processing systems is presented here.

Work on speech processing systems started in the eighteenth century with Wolfgang von Kempelen's design of the early vocal tract model which was able to produce speech sounds by mechanical movements (1).

The theoretical work on vowel production was started by Willis, Helmoltz (1829) and Wheatstone (1837) (2), and followed by the work of Alexander Graham Bell in the late nineteenth century which led to the invention of the telephone, the first long distance speech transmission system in the history of mankind. The design of phonautograph, speech in visible form, by Bell (1873) was also of great use to the speech researchers at the time (3). Sir Richard Paget's work (4) early in the century (1930) on human speech gave more insight into the field. In 1939 Dudley announced

the completion of the channel vocoder which was the first system of its kind (5).

Pulse Code Modulation (PCM) system was proposed by A.H. Reeves (7) in 1939.

The first spectrogram machine was reported by Potter et al in 1945 (5).

The modeling of the vocal tract was the main research area at that time. An electrical vocal tract model was built and tested by Dunn (8) in 1950. Shannon's major contribution to the field of information theory was announced around 1950 (9) and a great deal of investigation on dynamic models of the vocal tract was carried out at the Massachusetts Institute of Technology (10). Researchers at the Haskins Laboratories were concentrating on systems of speech processing based on formants and their significance in speech perception and recognition. Flanagan in a similar area of research (1956) studied formant extraction and its use in speech compression systems (11). Another famous researcher in this field, Fant, investigated Formants, their predictability and their use in speech coding (12). In 1957, Smith (13) reported the use of P.C.M. in the Speech Coding systems and in 1960 Flanagan introduced a resonance vocoder (14). The first computer simulation of a speech processing system was reported in an article by Golden in 1963 (15). The Cepstrum, a method of analyzing periodic signals, was introduced into the speech processing field by Noll (16) and Kulya's method of

spectrum compression of speech signals was published in 1964 (17).

The theoretical and practical aspects of the time domain based systems such as Delta Modulation were studied by Mcdonald (18), O'Neal (34) and others in 1966.

Linear transformation, a transform domain approach, of the speech signals was introduced in 1966 by Crowther and Rader (19).

Homomorphic analysis of speech, a method for deconvolution of speech signals was first reported by Oppenheim and Schafer (20) in 1968.

Then the introduction of signal prediction techniques gave a new insight and impetus into the speech processing field.

Atal (21) studied the adaptive prediction coding of speech in the late sixties.

In the time domain based systems of speech processing the work of Cattermole (22) on Pulse Code Modulation Systems is impressive.

The studies of Shanks, Bowyer (23) and Klein[24] are known in domains other than time for speech processing. Frequency domain analysis has been used extensively in the systems of speech processing. In the late sixties the trend was towards the digitization of speech processing systems.

Time and frequency domain analysis techniques were modified to suit the new digital approach. With the introduction of digital systems and computers the need for an alternative and perhaps a more efficient analyzing domain was felt. In this respect the Walsh domain offered itself as one of the possibilities.

Many researchers started investigations into different orthogonal domains, of whom the work of Harmuth (25) in 1972, Ahmed and Rao (26) are important and typical. The use of the Walsh domain in the area of speech processing has been studied by researchers such as Shum and Elliott (27).

The possible use of the Discrete Cosine transform was reported by Ahmed et al (28) in 1974 whilst the use of the Slant transform was considered for image processing by Pratt et al (29). Recently the number theoretic transform has come under investigation as a new transform technique for digital signal processing (30). A good review of the frequency domain based on speech processing systems has been carried out very recently by Tribolet et al (31). Linear Prediction techniques for speech processing have recently attracted considerable attention for which typical references are Markel and Grey (32), Makhoul (33).

Adaptive schemes have been considered very efficient in the field of signal processing in general and speech processing in particular. Thus Adaptive Pulse Code Modulation (APCM) (43), Adaptive Delta Modulation (ADM) (38), Adaptive Delta Pulse Code Modulation (ADPCM) (38), Adaptive Predictive Coding (38), Adaptive Transform Coding (40), Adaptive Bit Allocation (41), Adaptive Sub-band Coding (42) and other similar techniques are examples of such systems developed recently.

Time domain based systems of speech processing have also been studied recently. Delta Pulse Code Modulation

systems have been investigated and improved by O'Neal and Stroh in 1972 (34). Cummiskey, Jayant and Flanagan have studied Adaptive Delta Pulse Code Modulation (ADPCM) schemes (35). In 1973 Jayant introduced an interesting quantization technique suited to digital speech processing (36). A good comparison between PCM, DPCM and DM have been presented by Jayant in 1974 (37). Noll also has compared some of the time domain quantization techniques extensively (38). Recently a time domain based speech processing system has been proposed by King and Go sling (39) the so-called time-encoded speech system, which is currently examined by many people before being put into practical use.

2.2 DIFFERENT SPEECH PROCESSING SYSTEMS

Digital speech coding systems have been developed drastically over the last twenty years. Researchers have tried to design systems with the highest possible efficiency for specific applications. Where complexity is a problem, simple systems have been designed but with high bandwidth requirements, the converse being low bandwidth systems with relatively complex structures. Speech researchers discovered, at a later development stage, that speech contained a high degree of redundancy. For the best results, the speech characteristics must be used efficiently. It was at this stage that circuit designers realised the perceptual properties of speech and the performance of ear and brain in this respect. Relatively little is known about this area of research in the sense that many facts about the

perception of speech by the human brain are still unclear. Further research in this field will give more insight into the problem of speech coding system design which will enable the development of more efficient systems.

The existing digital speech coders can be classified in different ways. Of the more useful classifications, the following can be mentioned:

(1) One way is to classify the coders into two groups: waveform coders and source coders (vocoders). The waveform coders tend to follow the signal waveform and can therefore be used in a variety of signals. Although they are signal independent, some signal statistics can be used for improving design. Normally they have simple circuits compared to the source coders but require more bandwidth in their implementation. These are usually time domain based systems such as PCM, DPCM DM, ADM, APC and similar ones.

The source coders, on the other hand, are speech dependent systems which make use of a priori knowledge of the mechanism for speech generation and also take into account the specific characteristics of the speech signals such as perceptual clues and important spectrum regions. This type of system is usually based on some sort of speech production modeling. Different vocal tract models have been suggested and used such as: (a) Parametric Articulatory Model, (b). Transmission line model, (c) Acoustic-oscillator Model (non-linear) (d) Pole-Zero Model, (e) All-pole Model, (f) Self-

oscillating Model, (g) Two-mass Model. More discussion of these models are given in references (14), (67), (88). The most practically successful model so far has been the All-pole Filter Model. The source coders are usually more complex than the waveform coders but their signals occupy less transmission bandwidth. They are normally based on a transform domain analysis, similar to the channel vocoder, Formant vocoder and others. It is also possible to combine the waveform and source coders to obtain a more efficient system to take advantage of the two mentioned types. In this case the coder tends to preserve the short-time amplitude spectrum of the speech signals.

(2) Another method of classification of digital speech coders is based on the merits of their transmission rate, complexity and quality of reconstructed speech. This,in fact, is useful in order to find a suitable system for a specific application. An example of such a classification is given in table 2.1.

(3) Digital speech coders can also be classified simply by their mode of operation either in time or transform domain. This is a more straightforward classification within which the merits of each system can be discussed. In the work presented here, the digital speech coders are classified into time and transform domains. In the following sections, the most interesting systems from each group are discussed.

2.3. TIME DOMAIN SYSTEMS

This group of the digital speech coders can be regarded as relatively simple in complexity, but with high transmission rate requirement. When the bandwidth is not the major problem but complexity is, a time domain based system is used, but when complexity can be tolerated to achieve bandwidth reduction, this type is no longer suitable. In the following review, digital time domain based systems such as PCM, DPCM, DM and APC are discussed briefly, each for its merits in this category.

2.3.1. PULSE CODE MODULATION (P C M)

The Pulse Code Modulation System was the first fully digitized waveform coder. It represents the signal in digital form both in time and amplitude. It was first proposed by A.H. Reeves of the ITT Corporation in a French patent specification in 1939 (7) and in an American specification in 1942 (7).

Sampling, quantizing, companding, pulse shaping and coding are the essential elements involved in the PCM processes.

One of the major problems of the practical PCM is the wide range of power levels to be handled. Companding is used in a somewhat complementary manner to lift the level on line of the low level signals while maintaining the peaks of the high level signals within a predetermined power limit. This improves the low level signal-to-noise ratio at the expense of changing

a constant level noise into one related to signal level in accordance with the companding law. In order to secure a linear relationship between quantization noise and signal level, a non-linear (logarithmic)relationship between the number of quantisation steps and the signal level must be used.

A truly logarithmic curve would not pass through the origin and there have been many curves examined which offer a reasonable compromise. Two have been the subject of debate by the C C I T T. These may be expressed as:

A-Law

$$\begin{cases} F(x) = \dfrac{1+ \log Ax}{1+ \log A} & \dfrac{1}{A} \leqslant x \leqslant 1 \\ F(x) = \dfrac{1+ Ax}{1+ \log A} & 0 \leqslant x \leqslant \dfrac{1}{A} \end{cases}$$

μ-Law :

$$F(x) = \frac{\log (1 + \mu X)}{\log (1+ \mu)}$$

The A-law is in effect truly logarithmic for higher levels with a linear bottom section comprising of the tangent through the origin.

The μ-law has a comparable overall form but is nowhere truly logarithmic and nowhere truly linear, though it approximates to these characteristics at the extremes. The use of the A-law with a value "A" of approximately 100 enables these requirements to be met within a specified error assuming an eight bit word digits (i.e. $\pm$ 128 levels).

The process of companding may be executed by a compression operation on the speech waveform followed by linear quantisation, the inverse operation being performed at the receiver. PCM representation of the speech waveform low pass filtered to 4KHz and sampled at 8 KHz, requires 56 K bits/second if a logarithmic quantization of 7 bits/ sample is used.

Another way for achieving better results and a reduction of the quantization noise for signals of large dynamic range such as speech, is to use a truly optimal quantization strategy that uses a time-varying or adaptive quantization level. Adaptive (APCM) quantization utilizes a quantizer characteristic (uniform or non-uniform) that shrinks or expands in time. Generally APCM codes good quality speech at bit rates greater than 20 K bits/sec. The APCM implementation, therefore, is more complex than the normal PCM system, but requires less bandwidth., An efficient adaptive quantization method was proposed by Jayant in 1973. Details on PCM coders can be found in Cattermole's book (22) and on APCM systems in Jayant and Wilkinson's papers (37, 43). Block diagram of a typical PCM is shown in Figure 2.1.

2.3.2. DIFFERENTIAL PULSE CODE MODULATION (DPCM)

The main drawback of the PCM systems is their high bandwidth requirements. For achieving the bandwidth reduction, the methods of non-uniform quantization and adaptive quantization were mentioned before. A further

reduction of the bandwidth which is required can be achieved if the difference between the input signal and a predicted version of this signal is quantized and transmitted based on the correlation between successive input samples. Because adjacent amplitudes in speech waveforms are highly correlated, the variance of the difference between successive speech samples is much smaller than the variance of the signal itself. This suggests that the DPCM method could be efficiently used for coding of speech signals. This method uses feedback to reduce signal redundancy as shown in Figure 2.2.

The predicted signal is taken as the previous signal sample and is used for comparison with the input sample. In a similar manner to the PCM case, adaptive as well as fixed quantization techniques can be used. The adaptive strategy exploits the short time autocorrelation function statistics of the signal, that is, the predicted samples are adapted to the changing spectral properties of the speech signals. Although this adaptation enables a gain of up to 16 dB over log-PCM, it results in a more complex system. The lowest bit-rate achievements in differential encoding of speech have depended on predictions that are both spectrum adaptive and pitch adaptive. (35,38, 44).

2.3.3. DELTA MODULATION (DM):

Following the development of the PCM coders, researchers looked for systems with less complex y. This resulted in the introduction of Delta Modulation systems at I T T French laboratories in 1946 (46). Delta Modulation is a form of

differential PCM with a simpler structure; as it quantizes the difference signal by one bit (2-level) only, as shown in Figure 2.3. Despite its basic similarity to differential PCM, there are many detailed differences. Of these, the sampling rate is a major one as the DPCM normally works at the Nyquist sampling rate whereas Delta modulation employs much higher sampling rates. The step size adaptations are also different for DM and DPCM. As in the previous cases, adaptive as well as fixed quantization can be employed, due to the fact that a fixed step size introduces problems such as slope overload distortion and granular noise. Good quality speech can be achieved by DM with a fixed step-size but with a relatively high transmission rate (of the order of 32 K bits/sec). Using the adaptive DM, one can reduce the bit rate, whilst keeping the quality, by reducing the sampling rate and adapting the step size to the requirements of the input signal.

The quantization noise in DM coding has components outside the transmission bandwidth due to input oversampling, hence by employing a low pass filter at the output stage the noise components can be eliminated. There are many versions of Delta Modulation discussed in the literature, such as Delta-Sigma Modulation, Exponential Delta Modulation, Exponential Delta-Sigma Modulation, Continuous Delta Modulation. Those systems which use a companding method to improve the performance of the Delta Modulator are: Externally Companded DM, Digitally Companded DM, Discrete DM, High Information DM, and other

similar ones. A good survey of these systems can be found in Steele's book (45) and in references (36) (37).

2.3.4. ADAPTIVE PREDICTIVE CODING (APC)

The previous section highlights the need for adaptive systems in the low-bit rate transmission of speech signals. The adaptive predictive coding technique has been developed for low bit rate systems and a brief explanation is given in this section.

The Adaptive Predictive Coding approach for speech signals can be designed using two stages to exploit the correlations between successive speech samples and also to exploit the quasi-periodic nature of the voiced speech. In this coding method, both the transmitter and the receiver estimate the current value of the signals by linear prediction on the previously transmitted signal. The difference between this estimate and the true value of the signal is quantized, coded and transmitted to the receiver. At the receiver end, the decoded difference signal is added to the predicted signal to reproduce the input speech signal. This approach gives an improvement over the previously mentioned systems PCM, DPCM, DM. A block diagram of an APC coder is shown in figure 2.4.

The predictor of the coder can be similar to the DPCM encoder form. One of the main difficulties in this system is 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 too complex. The degree of prediction depends on the correlation between the adjacent pitch periods.

Such correlations vary considerably across speech sounds and speakers. Good work on APC coders has been done by Atal and Schroeder (21) and Noll (38).

2.4 TRANSFORM DOMAIN SYSTEMS

A further reduction in bandwidth of coding systems is possible if the coding is done in domains other than time. In this sense the information bearing parameters can be coded with less bandwidth than the time domain case. This reduction of bandwidth is compromised by an increase in the system complexity. Among the transform domains, the frequency is the most well known one, which is used by most transform domain based systems.

In the following section the Channel Vocoder, the Formant Vocoder, the Sub-band Coder and Adaptive Transform Coding systems are discussed.

2.4.1 CHANNEL VOCODER

Homer Dudley announced the invention of the channel vocoder, the first of its kind, in 1939.(5) It is based on an analysis-synthesis approach to speech signals. The frequency domain is used as the analysis domain. At the transmitting end, some information regarding the spectrum shape of the speech is extracted from the speech signal (analysis) and at the receiver end this information is used to reconstruct the speech (synthesis). The aim in the channel vocoder system is to preserve the power spectrum and the pitch period of the speech signals. The power spectrum is sampled at discrete frequency points and

information about these points is coded and transmitted at a much slower rate than would be required for the time domain waveform. In this case a great bandwidth saving can be achieved but at the price of a more complex system compared to a time domain based system. To extract suitable information from the power spectrum, early channel vocoders used a bank of band-pass filters (normally 10 to 20 filters). An introduction of digital techniques has led to digital filters being used for this job. The Fast Fourier Transform (FFT) can also be used instead of direct filtering. The low-pass filtered signals obtained at the output of the analyser are connected to the transmission path, and at the receiver end are used to control dynamically the gain of each synthesis filter. The synthesis filters are supplied with a wide-band (pulse or random noise) excitation signal controlled from the analyser and each synthesis filter selects a band of excitation about its centre frequency. Summation of the output signals from the synthesis filters and further equalization yields the synthetic speech. A block diagram of a channel vocoder is given in Figure 2.5.

As this vocoder transmits almost all of the power spectrum it cannot be regarded as the most efficient system due to the known fact that certain regions in the frequency domain are more important perceptually than others. These regions can be found by psycho-acoustic experimentation. (47). The bandwidth required, is about one tenth of the original speech bandwidth requirement..

2.4.2 FORMANT VOCODER

A more efficient system in the frequency domain for achieving low transmission-bandwidth is the Formant Vocoder. In this class of vocoders, the most important regions perceptually of the spectrum are transmitted. These regions have been proved to be the spectral peaks (from extensive psycho-acoustic studies) of the speech signals (47). These peaks are called the formants and correspond to the poles in the transfer function of the vocal tract. In the formant vocoders, the position of the formants, normally the first three, together with their intensities are transmitted. At the receiver end these parameters are used to control a formant synthesizer for the reconstruction of the speech waveform. A block diagram of a typical formant vocoder is given in Figure 2.6. Different methods can be used for extraction of the formants such as the Zero-Crossing method, peak picking approach, spectral matching, spectral envelope method, moments approach, LPC. These methods are discussed in more detail, in section 4.2. Digital systems, as well as analogue ones, can be designed for speech processing based on Formants (48). The bandwidth required for speech processing is around 1kHz in the analogue case and similarly in digital systems based on Formants. A comparison of this system with others can be seen from Table 1.1 (Chapter 1).

2.4.3. SUB-BAND CODER

In the wake of the formant vocoder, a more adaptive

coding system is the Sub-band Coder. In this approach regions around the formant positions are transmitted.by adjusting the frequency of the filters in accordance with the change in formant locations. The sub-band coder, in digital form was first proposed by Crochiere, Webber, and Flanagan in 1976 (42) (49).

As the quantization distortion is not equally detectable at all frequencies a non-uniform bit allocation strategy is adopted based upon perceptual criteria to reduce the audible noise level. The spectrum bands are partitioned such that each sub-band contributes equally to the so-called Articulation Index, AI. (47). The AI concept is based upon a non-uniform division of the frequency scale for the speech spectrum. Then the introduction of some gaps into the spectrum would be allowed provided that their contribution to the AI is negligible. So to design an adaptive and efficient system, the centre frequency of the upper bands can be allowed to vary in accordance with the dynamic movements of the first, second and third formants.

This would result in a reduction of the information-rate compared to the fixed band alternative, although a more complex system is employed. The complexity can be reduced by using a simple zero-crossing measurement technique for formant detection (42, 49).

This would allow the reduction of the bit rate to around 4.8 k bits/sec. The block diagram of a variable sub-band coder is given in Figure 2.7.

Sub-band coding has shown to be an efficient method of exploiting the short-time correlations due to the formant structure in speech. By allowing the independent variations of the quantizer step sizes in each band, the equivalent of a short time prediction can be achieved. Although this prediction is obtained in a coarse, piece-wise manner as a function of frequency, it can match the performance of adaptive time domain coding methods with fully adaptive short-time predictors.

2.4.4. ADAPTIVE TRANSFORM CODING

The systems mentioned above, were either time-domain based or frequency domain based systems. However, alternative transform domains can also be employed in signal analysis. Thus the speech waveform can be transformed into domains other than time, which include frequency, Discrete Cosine, Walsh, Haar, Slant or others and then adaptive strateges can be employed to achieve an efficient speech processing system. Among the orthogonal transforms, the Cosine Transform leads to a nearly optimum performance for almost all speech sounds (40). (More details on orthogonal transforms are given in chapter 3). The Discrete Fourier transform is used mainly for its familiarity and the availability of fast algorithm, the Walsh Hadamard transform is attractive for its simplicity in digital applications and other orthogonal transforms are used in specific applications for which they are suitable. Generally, the speech characteristics are taken into account in the desired transform domain and optimum methods are designed to suit that particular domain. In the Adaptive

Transform Coding method (40, 41), due to the non-uniform energy distribution of the speech waveform in the transform domain, adaptive bit assignment and adaptive quantization techniques are employed in order to exploit the signal energy distribution changes which result in a more effective coding system. The adaptation is controlled by a short-term basis spectrum which is derived from the transform coefficients prior to coding and transmission and is sent as side information to the receiver. The approach is a non-pitch tracking coding method where bit-rates of 12 Kbit/sec. can be achieved. At the receiver end, the quantized coefficients are replaced at their appropriate positions and an inverse transformation is performed to produce a replica of the original input segment. Successful segments when joined together represent the input speech signal. An analysis window is used to control block boundary effects. A block diagram of the adaptive transform coding system is shown in Figure 2.8.

2.5 LINEAR PREDICTIVE CODING (LPC)

The speech processing systems discussed so far reconstruct the speech either by coding the speech waveform in the time domain directly or by exploiting some of the speech spectrum redundancies in the transform domain. In the case of time domain coding of speech, high bandwidth is needed as the redundancies are not fully exploited. In the spectral analysis case, although it is a well known technique for signal analysis, limitations arise in the case of speech as it is a non-stationary and quasi-periodic signal.

To avoid the above mentioned problems, modeling of the speech waveform has been suggested as a possibilitty, (50), (51), (52), (53),(54).

The time domain based suggested method can be considered as an Adaptive Predictive Coding (APC) system in which the prediction residual is replaced by pulse and noise sources. In this method the speech spectrum is approximated by a number of prediction coefficients. The method of preserving the peaks of the speech spectrum using a direct method of peak picking has been applied in the formant vocoders, whereas in the LPC method the peaks are preserved by an indirect method of all-pole modeling. So in this approach the speech signals are represented in terms of time varying parameters related to the transfer function of the vocal tract and the characteristics of the source function (excitation). The speech waveform is realized by predicting the present sample as a linear combination of the n (normally 12) previous samples. The predictor coefficients are determined by minimizing the mean-squared error between the actual and the predicted values of the speech samples. Different methods have been suggested and used for obtaining the predictor coefficients which are based on the all-pole - modeling of the speech waveforms. The various methods of linear prediction have in common the assumption as already mentioned that a speech sample $S(n\tau)$, can be predicted approximately from a linearly weighted summation of a number of immediately preceeding samples, which can be formulated

as:*

$$\hat{S}(n) = - \sum_{K=1}^{P} A_K \, S(n-K)$$

where $\hat{S}(n)$ is the value of the signal predicted from the S(n-K) previous samples for different K values. The A_K coefficients are the prediction coefficients (weighting factors). They are obtained in such a way that the error between the actual (original) and predicted value of the sample is minimum.

Atal and Hanauer (53) used a knowledge of the auto-covariances of the input speech signal (referred to as non-stationary formulation) to obtain the prediction coefficients. In the covariance method, the error is minimized over a finite interval, i.e. over zero to N-1 samples. Markel (52) suggested the auto-correlation method to determine the prediction coefficient values (referred to as stationary formulations). In the auto-correlation method, the error is minimized over the infinite duration, from $-\infty$ to $+\infty$. therefore the speech signal must be windowed (59) before any analysis.

The main difference between the covariance and auto-correlation methods is that in the auto-correlation matrix the elements along the lines parallel to the diagonal are equal while in the covariance matrix this is not normally the case. Less storage and computation time is needed for the auto-correlation method compared to the covariance approach. The covariance method is preferred when the analysis interval is relatively short, since it leads to

* This is an all-pole model.

a smaller mean squared error compared to the auto-correlation method. The covariance method tends to the auto-correlation method as the length of the analysis window tends to infinity.

Itakara and Saito (55) introduced a method which uses an all-zero inverse filtering which differ from the above two methods in that it employs a special lattice filter structure in which the redundancies of the speech waveform are removed stage by stage. The coefficients in this method are called Partial Correlation Coefficients (PARCOR). In the covariance and auto correlation method the correlation is removed simultaneously by solving a system of simultaneous equations for prediction coefficients whereas in the PARCOR approach the correlation is removed stage by stage.

Atal and Hanaver (50) have 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, using the fact that a zero can be approximated by a large number of poles. An all-pole function approximation of the spectrum can be used to match the original spectrum and the reconstructed one.

An all pole function is of the form,

$$S(Z) = \frac{1}{1 + \sum_{i=1}^{n} A_i Z^{-i}}$$

Where n is the order of the function and the A's are the fil ter coefficients.

Given an inverse filter of the same order, a one to one relationship exists between the PARCOR coefficients and the direct form coefficients of the inverse filter method mentioned above. It has been shown 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 (52) (55).

The advantages of the Linear Prediction Coding system are improved computation speed and accuracy, considerable saving over direct sampling, and low bit-rate achievements. The disadvantages are given as sensitivity to noise, errors and it's complexity. The method of LPC can also be used for estimating the spectral envelope, formant analysis, pitch detection and others (52).

The prediction coefficients, as well as the so called partial correlations, pseudo-area coefficients (52), log area coefficients (52), in fact, show a high degree of correlation among themselves that can be exploited by transformations or similar methods.

Block diagram of a typical LPC system is shown in Figure 2.9.

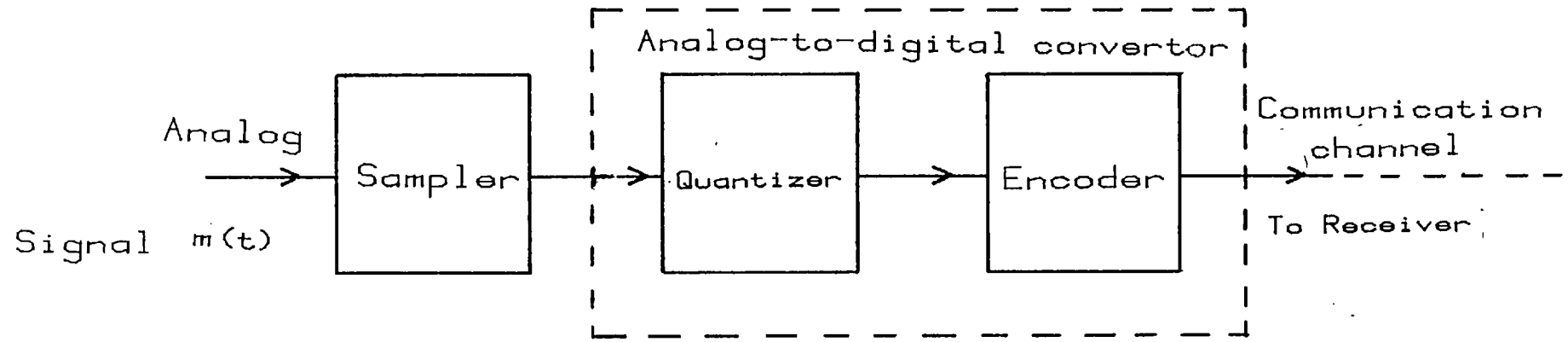
2.6. COMMENTS ON DIFFERENT SYSTEMS

The speech processing systems discussed so far have different merits which can be exploited for specific applications. Simple structured time-domain systems can be used where the complexity is a problem and more complex systems can be used where the bandwidth is restricted.

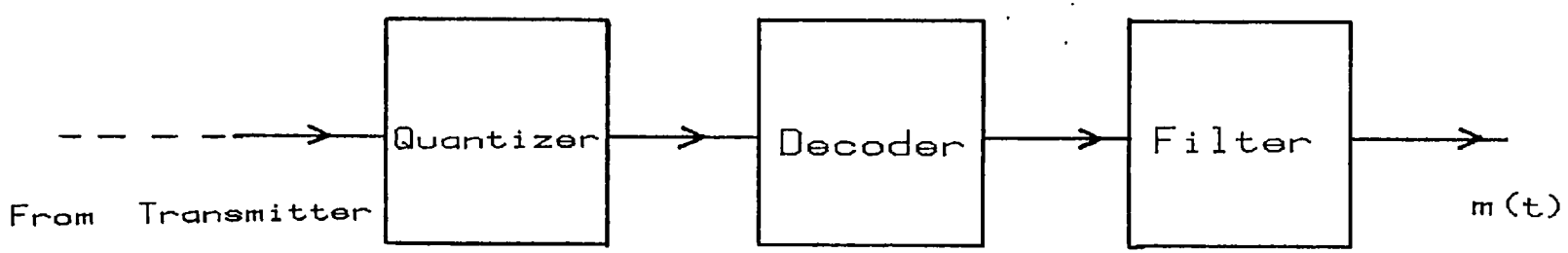
The time domain systems have relatively simple structure but occupy a relatively high transmission bandwidth. Amongst them, Delta Modulators are very efficient and adaptive type systems are nearer to the optimum case. For digital transmission of speech signals, at bit rates of around 32 Kbits/sec or less Differential PCM and DM provide SNR as well as perceptual advantages over PCM. At 16 Kbit/sec, the performance of DPCM and DM are not significantly different from one another. At bit rates below 16 Kbits/sec, the differential coder performance cannot be adequately described by a simple SNR value, but depends instead on a variety of parameters that characterize DPCM or DM noise structure, as in the case of low bit-rate PCM. A more detailed comparison between the time domain systems can be found in reference (37) (38).

The transform domain systems of speech processing have more complex structures than their time domain counterparts but occupy less transmission bandwidth. In this case more speech characteristics are exploited and therefore the systems are more efficient than time domain ones. The adaptive systems are more attractive as they lead to an optimum performance. A detailed study of the important frequency domain systems is given in reference (31).

A general review of the speech coding systems is given in reference (66).



(a) TRANSMITTER



(b) RECEIVER

Fig. 2.1: A PCM Communication System

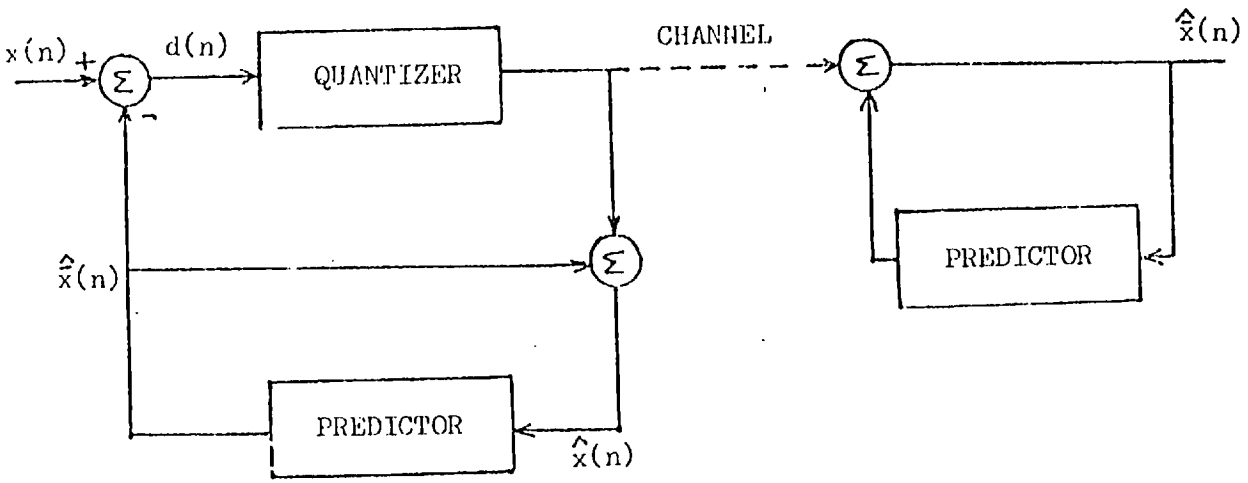


Fig. 2.2: Block diagram of a DPCM system

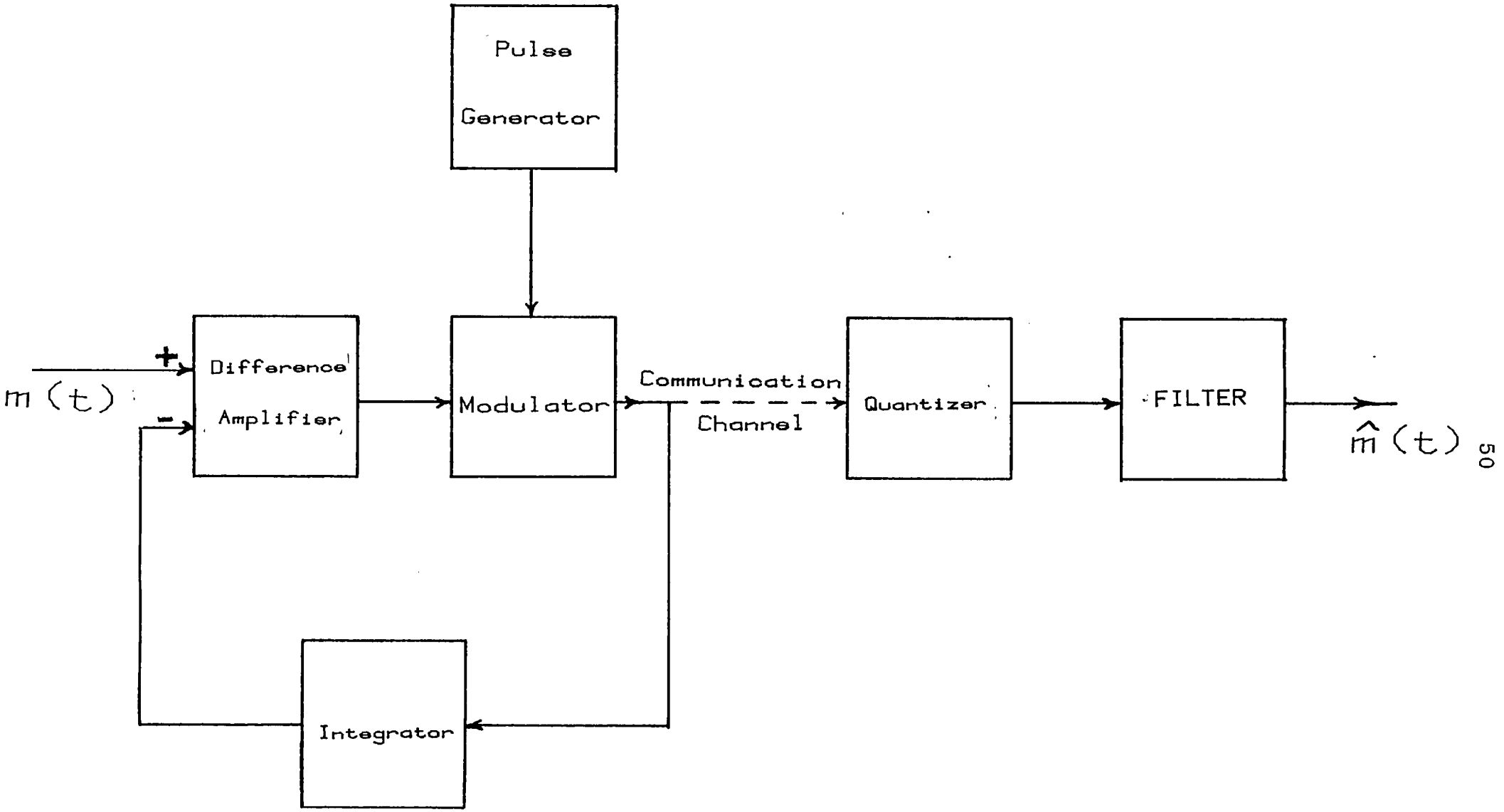


Fig. 2.3: A Delta-Modulation system of communication

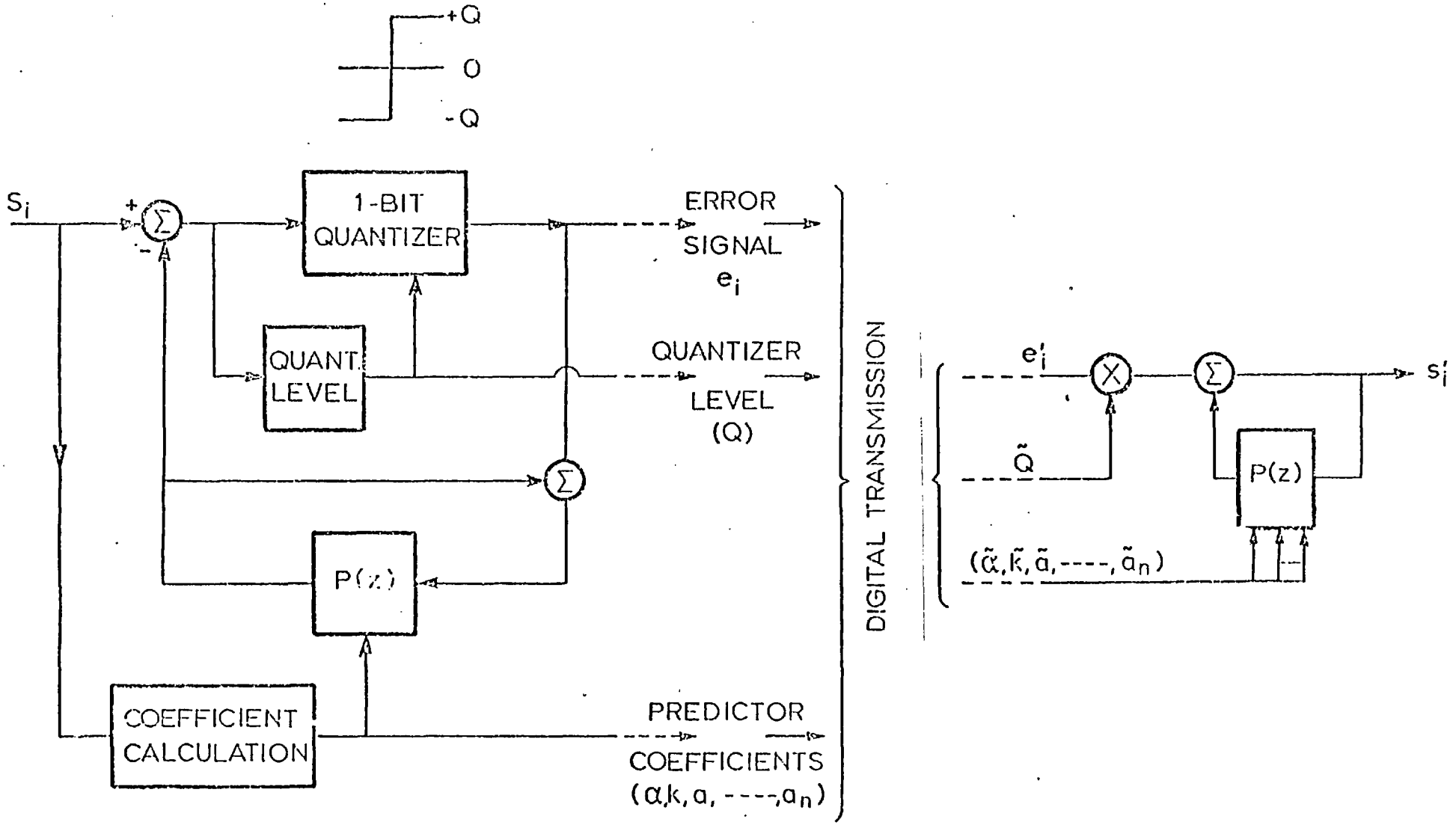


Fig. 2.4: Adaptive Predictive Coding System

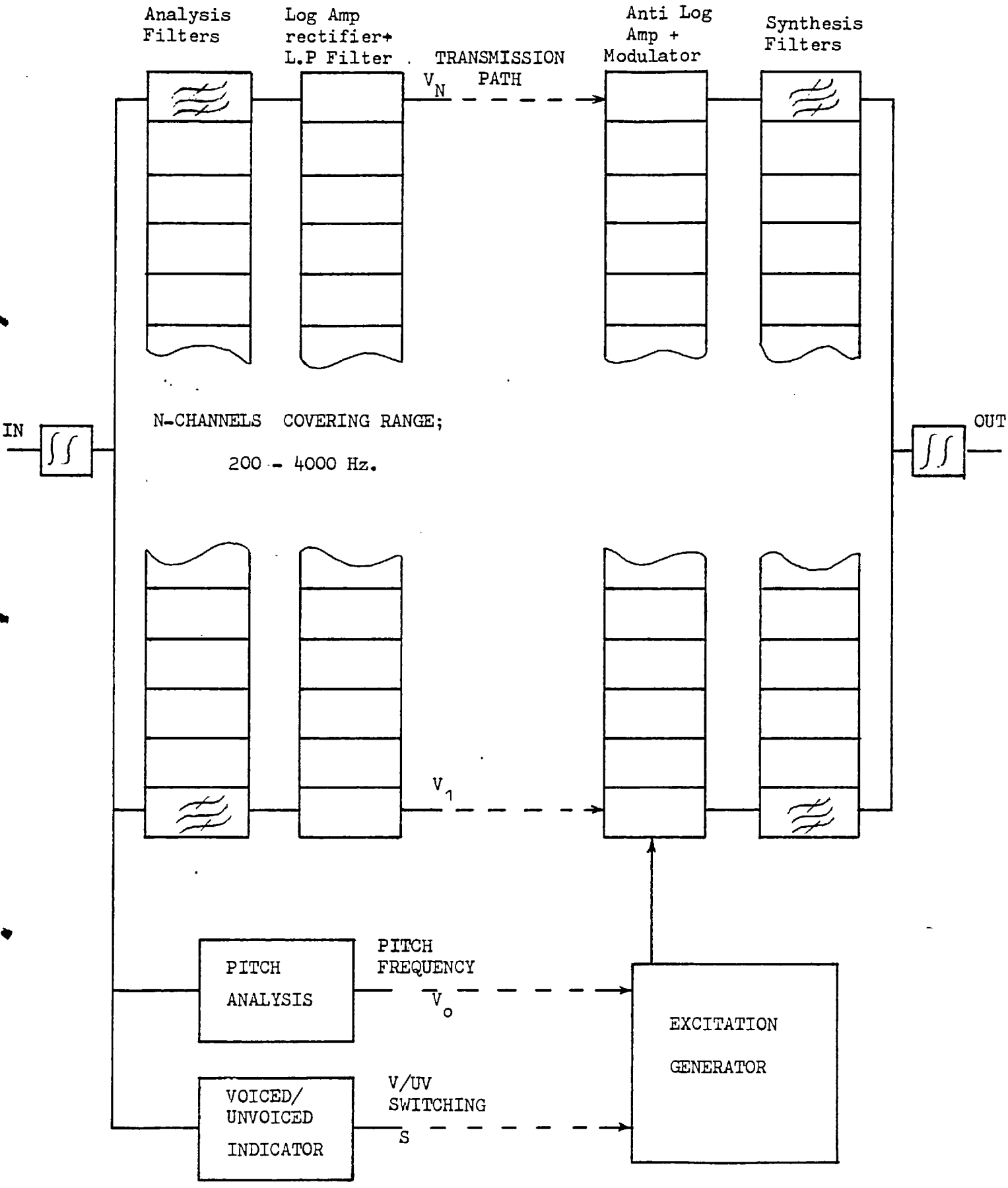


Fig. 2.5: Block diagram of a typical

Channel Vocoder

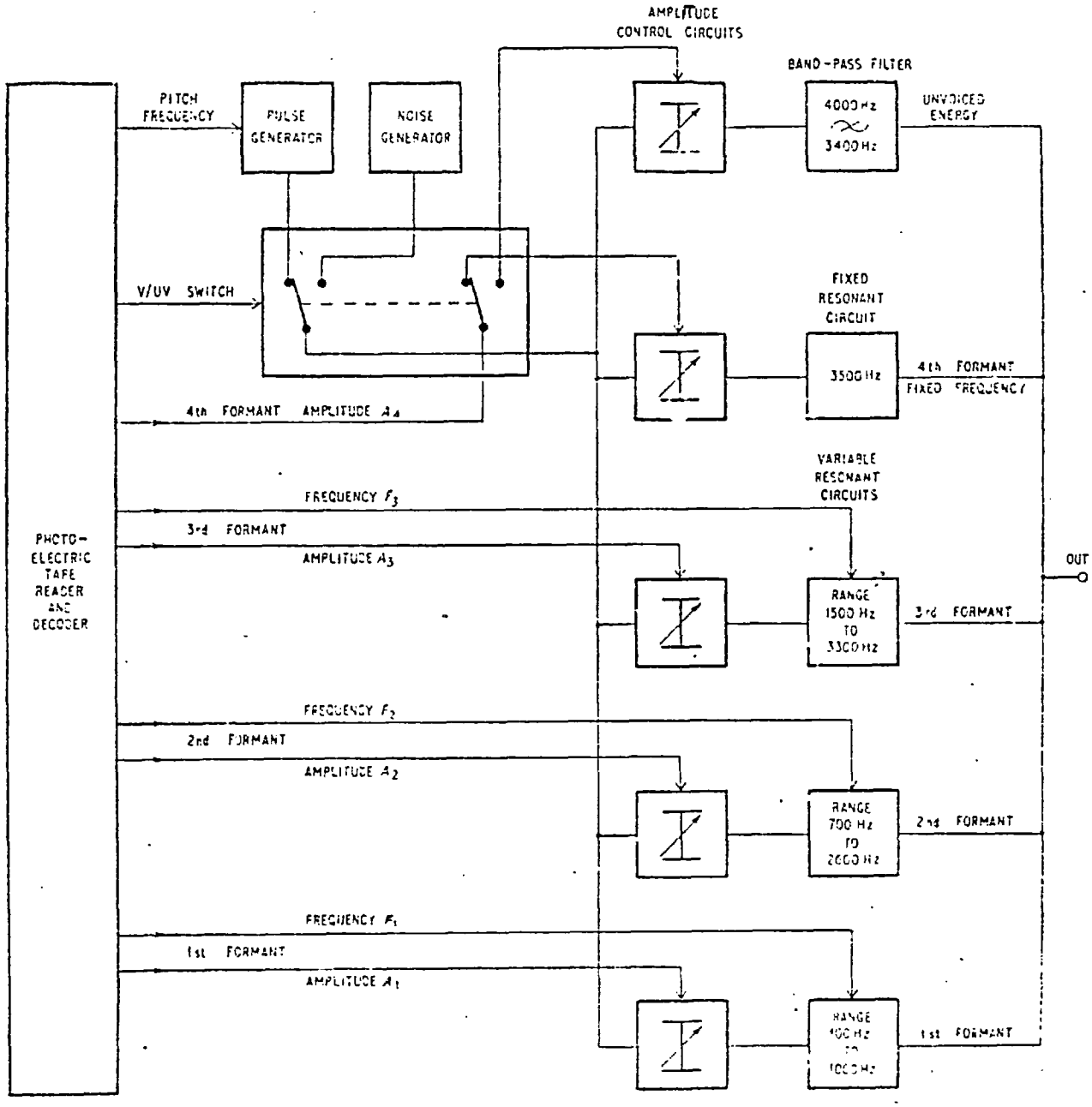


Fig. 2.6: Block diagram of a Formant Synthesizer

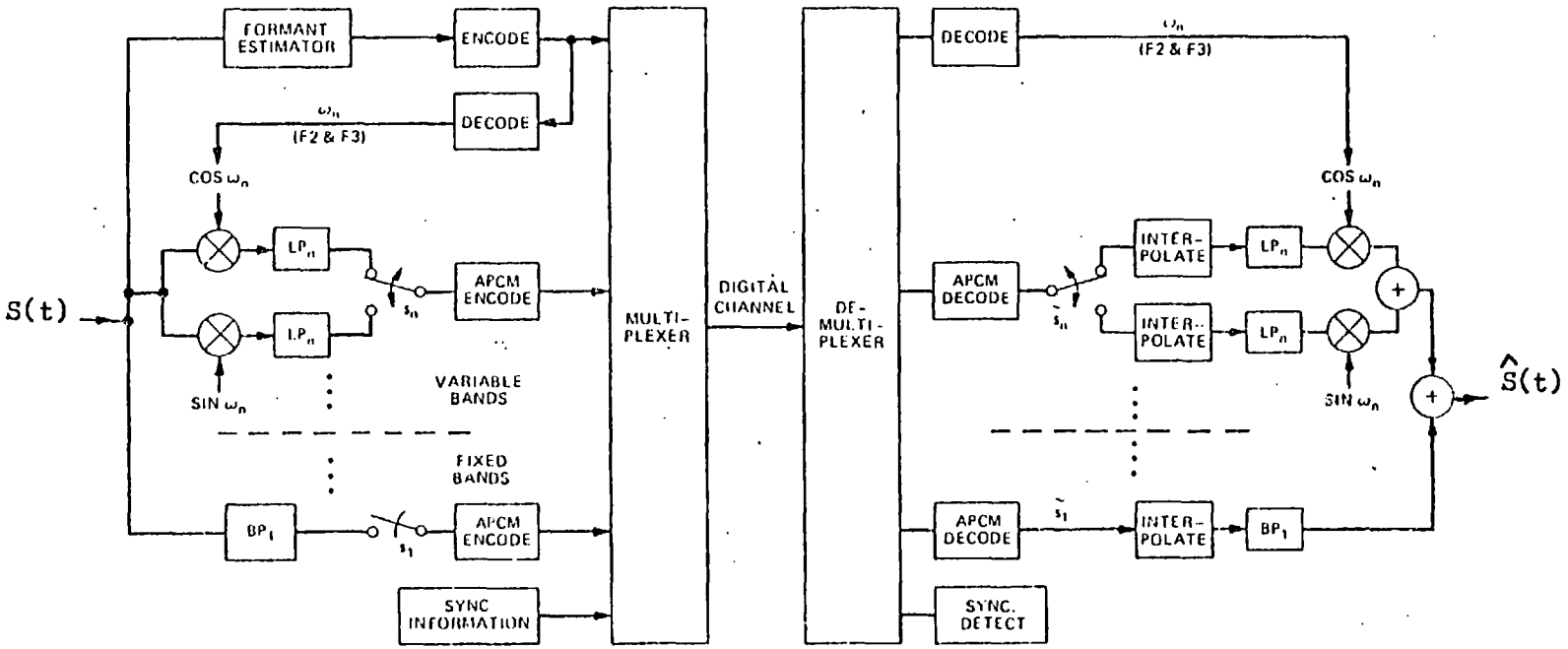
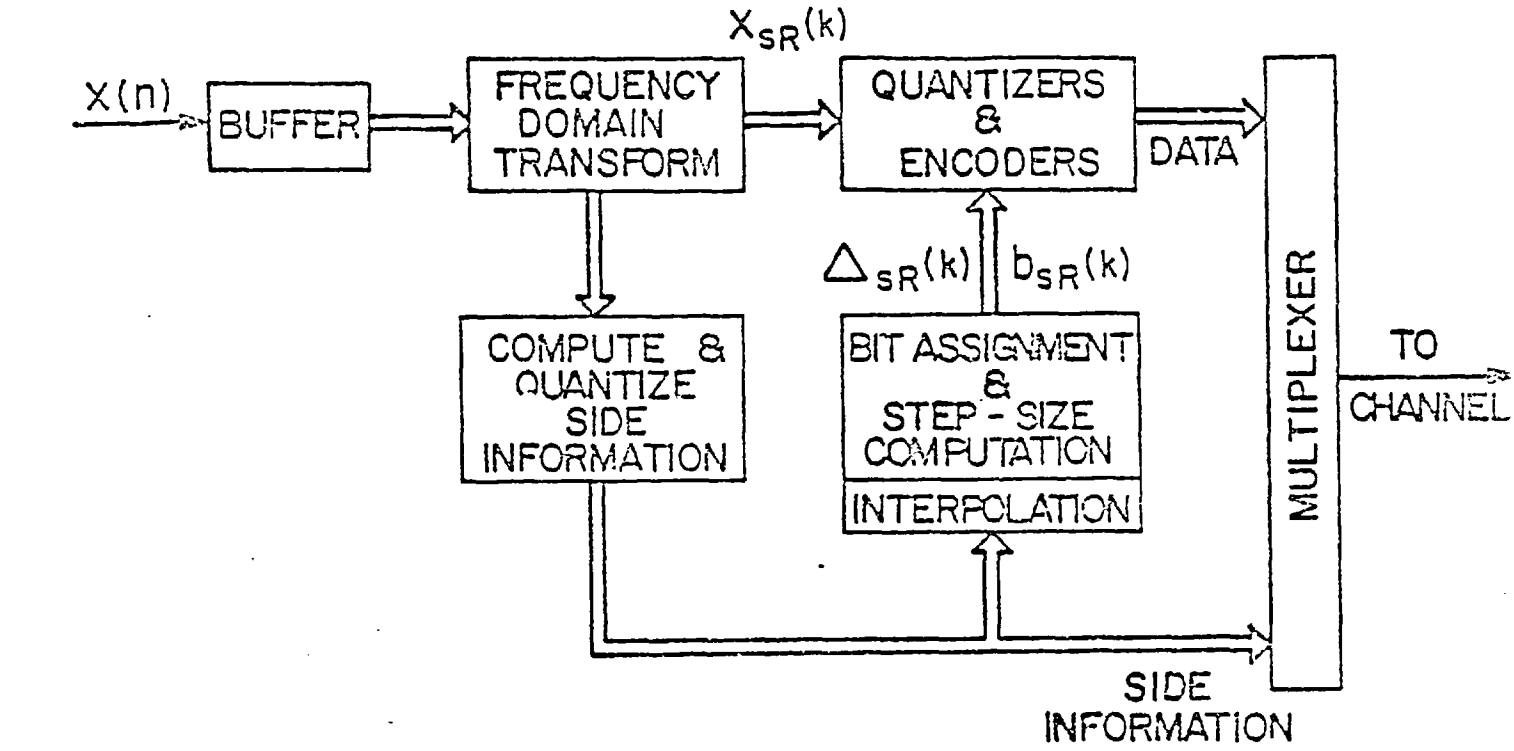
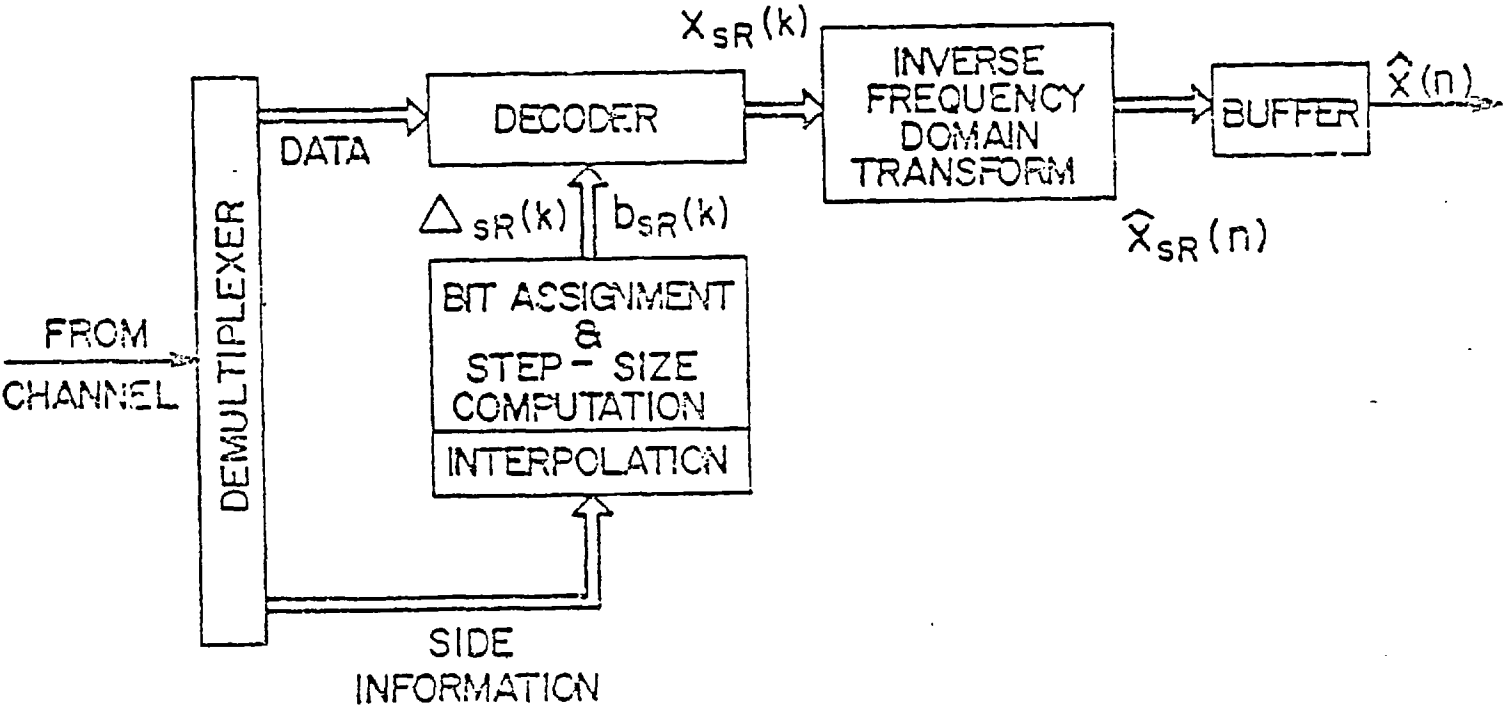


Fig. 2.7: Block diagram of a Variable-Band Coder



(a) TRANSMITTER



(b)RECEIVER

Fig. 2.8: Structure of an Adaptive Transform Coder

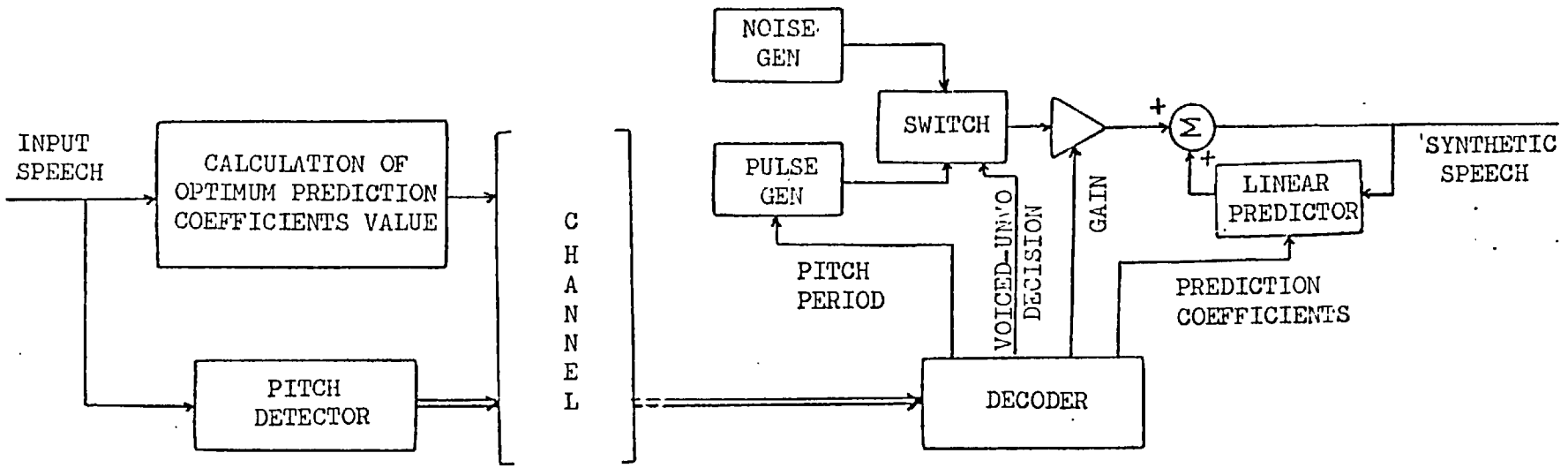


Fig. 2.9: Linear Prediction Vocoder System

CODER	Relative Complexxity	Bit Rate Kbits/s	Transmission Quality
1)Adaptive Delta Modulator (ADM)	1	40 24	Toll Communication
2)Adaptive Differential PCM (ADPCM)	1	32 16	Toll Communication
3)Sub-Band Coder (SUB-BAND)	5	24 9.6	Toll Communication
4)Pitch-Predictive ADPCM (PPADPCM)	5	24	Toll
5)Adaptive Predictive Coder (APC)	50	16 7.2	Toll Communication
6)Adaptive Transform Coder (ATC)	50	16 7.2	Toll Communication
7)Phase Vocoder (ΦV)	50	16 7.2	Toll Communication
8)Voice-Excited Vocoder (VEV)	50	16 7.2	Toll Communication
9)Linear Predictive Coder (LPC)	100	2.4	Synthetic
10)Channel Vocoder (CV)	100	2.4	Synthetic
11)LPC with Orthogonalized Coefficients (ORTHOG)	200	1.2	Synthetic
12)Formant Vocoder (FORMANT)	500	0.5	Synthetic

Table 2.1: Merits of different Speech Coding Systems

CHAPTER THREE

SIGNAL PROCESSING IN THE TRANSFORM DOMAIN

3.1 TRANSFORM DOMAIN ANALYSIS OF THE SIGNALS

The main advantage of transforming a signal into domains other than time is to obtain a non-uniform energy distribution with respect to the signal samples in that domain. Coding schemes based on this non-uniformity can be designed for signal processing applications with less bandwidth requirements and higher coding efficiency. To achieve this two main categories of transforms are available. Orthogonal and Non-Orthogonal.

Generally for signal processing applications the orthogonal transformation is used due to the fact that it converges to the original signal rapidly. A system $\{f(j,x)\}$ of real and almost everywhere non-vanishing functions $f(o,x), f(1,x), \ldots$ is called orthogonal in the interval $x_0 < x < x_1$, if the following conditions hold true:

$$\int_{x_0}^{x_1} f(j,x)f(k,x)dx = X_j \, \delta_{jk}$$

where

$$\delta_{jk} = \begin{cases} 1 & \text{for } j = k \\ 0 & \text{for } j \neq k \end{cases}$$

If the constant X_j is "1" then the functions are called orthonormal. Orthonormal expansions can be defined and used efficiently in signal processing.

Given the statistical properties of signals and their noise attributes, one can usually find a basis of orthonormal functions in which signal processing tasks, such as filtering and coding, can be implemented in an optimal fashion.

In most cases the spectral representation which is optimal in the statistical sense (such as Karhunen-Lòeve expansion) is not the most efficient computationally.

It is due to this reason that one is tempted to search for suboptimal representations that possess fast computational algorithms. The increase in the computational speed allows processing of larger size data blocks at higher sampling rates, which results in an overall performance higher than that obtained with statistically optimal but computationaly inefficient representations.

An important application of the orthogonal transforms is to data compression.

A typical transform coder consists of the following major sections; first, the digitizer which converts the analogue data into discrete and quantized blocks of data each of length N. Secondly, the encoder which multiplies the input vector by a unitary matrix (different for each domain), then acts upon the obtained transform domain coefficients according to the technique which is to be used in each case, The result is transmitted using a binary code. The receiver consists of a decoder which recovers the spectrum in the transform domain and transforms the spectrum back into the real (time or space) domain using an inverse transformer. The digital vector is then displayed in a continuous fashion with the help of a Digital to Analogue converter.

A block diagram of a transform coder is given in Figure 3.1.

As the speech signals are highly correlated, a linear transformation can be applied to take advantage of this fact and reduce the information rate necessary for the original speech transmission.

3.2 ORTHOGONAL TRANSFORM DOMAINS

The Orthogonal transformation of signals is a possible means of reducing the necessary information-rate for their transmission (25) (26). In this respect transform domains such as Karhunen-Loève, Discrete Cosine, Fourier, Walsh-Hadamard, Haar, Slant and Number Theoretic are of special interest since they are orthogonal and have been efficiently used for signal processing in different applications.

In the following section a brief discussion on each of the above mentioned transforms is given.

3.2.1. THE KARHUNEN-LOÈVE TRANSFORM (KLT)

The K-L transform is obtained from the eigenvectors of the signal covariance matrix. Generally it can be said that these eigenvectors are the functions that best describe the signal, in other words for the same number of coefficients, one could have a more precise reconstruction of the signal with these functions than with the same number of coefficients in another domain. (56). The transform operation can be given in general as: $Y = Tx$ where x represents samples of the signal to be transforme d. T is an orthogonal transform operation and Y is the signal representation in the transform domain. Then the transform

domain covariance matrix Σy can be given as,

$$\Sigma_y = T\Sigma_x T^{-1} = T\Sigma_x T'$$

where Σ_x is the covariance matrix of X, given by

$$\Sigma_x = E\{(x-\bar{x})(x-\bar{x})'\}$$

and E is the expectation value of X, and $\bar{X}$ is the mean value of the set x.

However, since T comprises the eigenvectors of Σ_x, it follows that,

$$\Sigma_y = \text{diag}(\lambda_1, \lambda_2, \ldots \lambda_N)$$

where λ_i, $i = 1, 2, \ldots\ldots N$ are the eigenvalues of Σx and "diag" stands for the diagonal elements.

Since Σy is a diagonal matrix, it can therefore be concluded that the transform vector components Y_i are uncorrelated. So the Karhuren-Loève transform is a signal dependent transform and has the unique property that the transform coefficients are uncorrelated, which means it is the optimum transform for signal representation with respect to the mean square error criteria.

The main disadvantage of the K L Transform is the lack of an algorithm for its fast computation (56).

The K L T is generally used as a basis for comparing the performances of other transforms (56).

3.2.2. THE DISCRETE COSINE TRANSFORM (DCT)

The performance of the Discrete Cosine transform compares more closely to the KL optimum transform than any other known transform. (28). The DCT coefficients of a

sequence X(n) (for n = 0, 1, ..., N-1) can be obtained from the following equations

$$C(0) = \frac{\sqrt{2}}{N} \sum_{n=0}^{N-1} X(n)$$

$$C(K) = \frac{2}{N} \sum_{n=0}^{N-1} X(n) \cos \frac{2(n+1)K\pi}{2N} \qquad K=1,2,\ldots N-1$$

where C(K) is the Kth DCT coefficient.

The inverse DCT is defined as

$$x(n) = \frac{1}{\sqrt{2}} C(0) + \sum_{K=1}^{N-1} C(K) \cos \frac{(2n+1)K\pi}{2N} \qquad n=0,1,\ldots N-1$$

It can be demonstrated that the DCT basis set provides a good approximation to the eigenvectors of the class of Toeplitz matrices as defined in reference (28).

This means that the DCT is only slightly sub-optimum compared to the optimum KLT.

It is interesting to note that as the long term correlation matrix of the speech signals is a Toeplitz matrix, then the DCT can be used efficiently for speech processing. For this case of the DCT a fast algorithm is available for the computation of its coefficients which gives it an advantage over the KLT.

The DCT has been used in speech processing systems such as the ADAPTIVE transform Coder (40, 41) and the Non-uniform Linear Predictive Coder (57), and in image processing systems for feature selection and pattern recognition.

3.2.3. THE DISCRETE FOURIER TRANSFORM (DFT)

The DFT is the most commonly used transform technique employed in signal processing.

In the communications field in general we shall encounter two distinct types of signals, and so two distinct and different corresponding types of spectra result. The first type of signal is the periodic function of time, which can be written in the form of a Fourier series. Let f(t) be a periodic function of time. Then f(t) may be represented in the frequency domain by an infinite number of sinusoidal components which are harmonically related to one another. The magnitude and phase of these components are given by a Fourier series expansion of f(t), as;

$$f(t) = \frac{a_o}{2} + \sum_{n=1}^{\infty} (a_n \cos n W_o t + b_n \sin n W_o t)$$

where

$$a_n = \frac{2}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) \cos n W_o t \, dt$$

and

$$b_n = \frac{2}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) \sin n W_o t \, dt$$

and

$W_o = \frac{2\pi}{T}$, the fundamental frequency;

and

$$a_o = \frac{2}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) \, dt$$

Where $\frac{a_0}{2}$ represents the average value of the D.C. term of f(t) over one complete cycle.. The second type of signal encountered in communications can be described as the non-periodic type. This can include transient and random time functions.

Non-periodic functions can best be analyzed by the use of the Fourier integral which is an extension of the Fourier series. It is a very powerful tool in the analysis of all types of linear systems and provides a clear picture of the physical situation in a practical problem.

Consider a non-periodic function

$$f(t) = f_1(t) + jf_2(t)$$

of a real variable "t" which satisfied the condition

$$\int_{-\infty}^{\infty} |f(t)|dt < \infty$$

Such a function may be transformed from the time domain to the frequency domain and vise versa by the following transform pair;

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(jw)e^{jwt}dw$$

$$F(jw) = \int_{-\infty}^{\infty} f(t)e^{-jwt}dt$$

The function F (jw) is called the Fourier Transform of f(t).

For the case of a discrete time periodic signal, the frequency domain representation can be given by the discrete Fourier series as;

$$x(n) = \frac{1}{N} \sum_{K=0}^{N-1} X(K) \exp(j\frac{2\pi Kn}{N})$$

and

$$X(K) = \sum_{n=0}^{N-1} x(n) \exp(-j\frac{2\pi Kn}{N})$$

where $x(n)$ are the discrete time samples in the time domain and $X(K)$ are their frequency domain counterparts (58).

A speech signal, can be regarded as a quasi-stationary (slowly time varying) signal and can be locally modeled with a short-time spectrum, as it is assumed to be block periodic locally.

To achieve this a windowing operation is applied on the continuous speech signal to obtain short time segments for performing these operations (59).

The Discrete Fourier transform is a sub-optimal transform compared with the Karhunen-Loève and Discrete Cosine transforms when applied to speech signals(40) but a very important advantage of the transform is the existence of a fast algorithm for its computation.

3.2.4. THE WALSH-HADAMARD TRANSFORM (WHT)

Another set of orthogonal functions are the Walsh functions (25 ,26,60). They consist of rectangular functions with different marks to space ratios. These functions have two levels, +1 and -1 only which make them attractive for digital implementation.

Any periodic function f(t) can be represented by these orthogonal rectangular functions, in a similar manner to the Fourier series, as follows:

$$f(t) = a_o + \sum_{n=1}^{\infty} \left(a_n \, Cal(n,\theta) + b_n \, Sal(n,\theta)\right)$$

where a_n and b_n are the Walsh coefficients. θ is the interval of orthogonality and n the sequency. The notation Cal (n,θ) and Sal (n,θ) is used for Walsh functions. The letters "C" and "S" remind one of the cosine and sine functions to which the respective Walsh functions are closely related, the letters "al" are derived from the name Walsh. The functions Cal(n,θ) are even functions similar to $\sqrt{2}\cos 2\pi n\theta$, whereas the functions Sal (n,θ) are odd functions similar to $\sqrt{2}\sin 2\pi n\theta$.

The parameter "n" for the Walsh functions equals one half of the average number of zero crossings in the interval $-\frac{1}{2} < \theta < \frac{1}{2}$. The Walsh functions of order 8 is shown in Figure 3.2.

The properties of Walsh functions can be given as;

(1) Instead of Sal (n, θ) and Cal (n,θ) the functions Wal (j,θ) can be defined as

$$\begin{aligned} Wal\ (2i,\theta) &= Cal\ (i,\theta) \\ Wal\ (2i-1,\theta) &= Sal\ (i,\theta) \end{aligned} \qquad \text{for } i = 1, 2, \ldots$$

(2) The product of two Walsh functions yield another Walsh function, $Wal(h,\theta)\ Wal(K,\theta) = Wal(r,\theta)$

where $r = h \oplus K$ and $\oplus$ is an addition modulo 2.

(3) The product of a Walsh function with itself yields Wal $(0, \theta)$

$$\text{Wal}(j,\theta)\ \text{Wal}(j,\theta) = \text{Wal}(0,\theta)$$

Since $j \oplus j = 0$

(4) The multiplication of Walsh functions is associative

$$(\text{Wal}(h,\theta)\ \text{Wal}(j,\theta))\text{Wal}(K,\theta) = \text{Wal}(h,\theta)(\text{Wal}(j,\theta)\text{Wal}(K,\theta))$$

(5) Dyadic correlation can be defined in the Walsh domain as

$$D_{FF}(\theta_\nu) = \int_{-1/2}^{1/2} F(\theta) F(\theta \oplus \theta_\nu)\, d\theta \qquad (1)$$

which can be considered for a finite interval of θ i.e. $-1/2 < \theta < 1/2$ which is more useful in communications. Then by using $F(\theta)$ and $F(\theta \oplus \theta_\nu)$

$$F(\theta) = \sum_{i}^{2^n - 1} a(i)\ \text{Wal}(i, \theta) \qquad (2)$$

and

$$F(\theta \oplus \theta_\nu) = \sum_{j}^{2^n - 1} a(j)\ \text{Wal}(j, \theta \oplus \theta_\nu) \qquad (3)$$

where

$$a(i) = \int_{-1/2}^{1/2} F(\theta)\ \text{Wal}(i, \theta)\, d\theta$$

By substituting (2) and (3) into (1) one gets

$$D_{FF}(\theta_\nu)=\int_{-1/2}^{1/2}\left[\sum_{i=0}^{2^n-1}a(i)\mathrm{Wal}(i,\theta)\right]\left[\sum_{j=0}^{2^n-1}a(j)\mathrm{Wal}(j,\theta\oplus\theta_\nu)\right]d\theta$$

by using the relation $\mathrm{Wal}(j,\theta\oplus\theta_\nu)=\mathrm{Wal}(j,\theta)\mathrm{Wal}(j,\theta_\nu)$ and the orthogonality of the Walsh functions one obtains:

$$D_{FF}(\theta_\nu)=a^2(0)\mathrm{Wal}(0,\theta_\nu)+\sum_{i=1}^{2^n-1}a_c^2(i)\mathrm{Cal}(i,\theta_\nu)+\sum_{i=1}^{2^n-1}a_s^2(i)\mathrm{Sal}(i,\theta_\nu)$$

where $a_c(i)=a(2i)$ and $a_s(i)=a(2i-1)$

Since addition and subtraction modulo 2 are identical operations, then there is no difference between dyadic correlation and dyadic convolution.

The Walsh functions are orthogonal in a finite interval, but can also be extended to functions orthogonal in an infinite interval (25, 26).

The Walsh functions can be classified into 3 groups These groups differ from each other in that the order in which individual functions appear is different ,

These groups are; (25, 26).

1) Sequency or Walsh ordering.

2) Dyadic or Paley ordering.

3) Natural or Hadamard Ordering.

In 1893, Hadamard defined an orthogonal matrix with elements +1 and -1, called the Hadamard matrix (26). The Hadamard matrices of rank 1 and 2 are

$$H_1=+1 \qquad \text{and} \qquad H_2=\begin{bmatrix}+1 & +1\\ +1 & -1\end{bmatrix}$$

Matrices of higher ranks can be obtained by Kronecker products. Using this the Kronecker product of H_2 with itself one gets;

$$H_4=H_2*H_2=\begin{bmatrix} H_2 & H_2 \\ H_2 & -H_2 \end{bmatrix}$$

The rank of any Hadamard matrix greater than 2 must be an integer multiple of 4.Certain Hadamard matrices of rank 2^n are related to Walsh functions. The Kronecker product also permits the generation of new functions.

When one deals ,in digital operations,with discrete values the Walsh functions must be presented in a sampled (i.e. matrix)form. Therefore because the general orthogonal matrices are called Hadamard matrices and Walsh functions are sequency ordered similar to the frequency domain(Fourier),the result of representing the Walsh domain in matrix (discrete) form is called the Walsh-Hadamard transform function.

There are three types of Walsh-Hadamard transforms similar to the ones mentioned before,i.e. Walsh-ordered,Paley-ordered and Hadamard-ordered.

As this transformation requires no multiplications or complex operations,then it can be performed much faster than the Fourier transformation.

Since the Walsh matrix is symmetric and orthogonal, the inverse transform is given by:

$$\left[W_n\right]^{-1} = \frac{1}{N}\left[W_n\right]$$

The factorization of the Hadamard matrix into "n" sparse matrices is the basis for the development of an efficient algorithm for the fast computation of the Hadamard transform (61,62). An "in place" computation algorithm is given in figure 3.3.

The total number of arithmetic operations (additions and subtractions only) required for the computation of the fast Hadamard transform is $N \text{ Log}_2 N$;where the FFT requires $(N/2)\text{Log}_2 N$ multiplications and a large number of additions.

The fast computation speed and its simplicity are the main reasons for the Walsh-Hadamard applications in the digital signal Processing fields(23,24,25,26,27,57,64).

3.2.5. The HAAR TRANSFORM (HT)

In 1909 Haar introduced a set of functions which are related to Walsh functions.The Haar functions are a three valued orthonormal basis of $L^2(0,1)$,the space of function f(x) that defined over (0,1) with $f^2(x)$ integrable in the Lebesque sense(26).

Any such function can be expressed as an infinite series in terms of Haar functions as given;

$$f(x)=C_0+\sum_{n=1}^{\infty}\sum_{m=1}^{2^{n-1}} C_n^m \, \phi_n^m(x)$$

where $\phi_n^m(x)$ is the orthonormal sequence defined on the closed interval (0,1) and can generally be given as;

$$\phi_n^m(x)=\begin{cases} 2^{(n-1)/2} & \text{for } \dfrac{m-1}{2^{n-1}} < x < \dfrac{m-\frac{1}{2}}{2^{n-1}} \\ -2^{(n-1)/2} & \text{for } \dfrac{m-\frac{1}{2}}{2^{n-1}} < x < \dfrac{m}{2^{n-1}} \\ 0 & \text{for } 0 < x < \dfrac{m-1}{2^{n-1}} \text{ and } \dfrac{m}{2^{n-1}} < x < 1 \end{cases}$$

and $$C_n^m = \int_0^1 f(x)\,\Phi_n^m(x)\,dx$$

This can be expressed by means of the partial sum;

$$S_N(x)=C_0+\sum_{n=1}^{N}\sum_{m=1}^{2^{n-1}} C_n^m\,\Phi_n^m(x)$$

which contains 2^N terms.

Several aspects of the potential utility of Haar functions derive from an important property of the partial sum. Following the Haar functions, an orthogonal Haar transform can be defined. Similar to other transforms, a discrete version of the transform can also be obtained which is represented in matrix form. The matrix operations are similar to those mentioned in previous sections. A fast algorithm for its computations also exists similar to the DFT and others.

The Haar transform has the advantage of an extremely efficient computational algorithm, but results in a large coding error.

It is important to note that the average number of operations per sample is independent of the transform size in this transform ,where as in the fast Fourier and fast Walsh transforms,the average number of operations per sample increases as Log_2N.

In general Haar functions have been used in data coding, edge detections, multiplexing and in speech processing.(63,64).

3.2.6. THE SLANT TRANSFORM (ST)

Another orthogonal transform which can be used in signal processing is the Slant transform based on the Slant basis vector (29). The slant vector is a discrete sawtooth waveform decreasing in uniform steps over its length. The orthogonal slant transform has a constant function for its first row and has a second row which is a linear slant function of the column index.

The discrete valued transform, in matrix form, can be formed by an iterative construction that exhibits the matrices as products of sparse matrices, which in turn leads to a fast transform algorithm for its implementation. The slant matrix of order N ($N=2^n$) can be given in terms of the slant matrix of order $N/2$ as

$$S_N = \frac{1}{2^{1/2}} \begin{bmatrix} 1 & 0 & & 1 & 0 & \\ a_N & b_N & 0 & -a_N & b_N & 0 \\ 0 & & I_{(N/2)-2} & 0 & & I_{(N/2)-2} \\ 0 & 1 & & 0 & -1 & \\ -b_N & a_N & 0 & b_N & a_N & 0 \\ 0 & & I_{(N/2)-2} & 0 & & -I_{(N/2)-2} \end{bmatrix} \begin{bmatrix} S_{N/2} & 0 \\ 0 & S_{N/2} \end{bmatrix}$$

where the matrix I $(N/_2)-2$ is the identity matrix of dimension $(N/_2)-2$ and the constants a_N and b_N are given by

$$a_2=1 \qquad b_N=(1+4(a_{N/2})^2)^{-1/2}$$

and $$a_N=2b_N\, a_{N/2}$$

$$s_2=\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

A typical 16 order slant basis vector is shown in Figure 3.4.

The orthogonal slant transform has the property of variable size, a fast computational algorithm, sequency property and high energy compaction (29).

The Fourier, Hadamard, and Haar transforms possess a constant valued basis vector but the slant transform possesses a sawtooth waveform decreasing in uniform steps over its length which is more adaptive ,to the nature of grey level image lines and therefore is used more efficiently in image processing (29).

3.2.7. NUMBER THEORETIC TRANSFORM (NTT)

In the early seventies (1972) Rader introduced transform operations, which by using a finite field (or finite rings) of integers resulted in the circular convolutions without quantization error. This class of transforms is called the Number Theoretic Transform (NTT) (30). The general form of the NTT can be defined as,

$$X(K)=\sum_{n=0}^{N-1} x(n)\,\alpha^{nK} \qquad K=0,\ldots,N-1$$

where $X(K)$ is the transformed sequence of the original $x(n)$ sequence, α is a general variable and N is the sequence length

The inverse NTT is given by

$$x(n)=\frac{1}{N}\sum_{K=0}^{N-1} X(K)\,\alpha^{-nK}$$

As it can be seen, it is similar to the Discrete Fourier Transform except that $\exp(\frac{-j2\pi}{N})$ is replaced by α. The value of α is taken as the primitive root of unity of order N:

$$\alpha^{N}=1$$

and when modular arithmetic is used, then the above mentioned equations are evaluated in modulo M, an integer.

Due to the nature of modular arithmetic, numbers do not have size or magnitude.

To obtain F F T -type algorithm, values of M, N and α have to be chosen carefully.

One of the best choices of the value M is when $M = 2^K + 1$. For K even M is called a Ferma t number. Ferma t numbers appear to be a good compromise with respect to practical transform lengths and moderate word size.

The NTT has been used mainly in convolution operations. It has also been used in Digital filter design (65) and appears to be a useful domain for digital signal processing.

3.3. COMPARISON OF TRANSFORM DOMAINS

Transform domains can be evaluated for a particular application, each with certain advantages and disadvantages. In the case of speech signals, extensive studies have been carried out, from which the behaviour of each transform domain can be stated as shown in Figure 3.5.

However, for signal processing applications the advantages and disadvantages of each domain is tabulated in Table 3.1.

3.4. SPEECH SIGNALS IN THE TRANSFORM DOMAIN

Speech consists of two types of sounds, voiced and unvoiced. Voiced sounds are periodically produced waveforms from the vocal cords. The unvoiced sounds have non-periodic, burst type waveforms. The transform domain representations of voiced and unvoiced sounds are also different. In the time domain case, they can be identified from their average number of zero crossings. The number of zero-crossings is higher in the unvoiced case compared to the voiced case. This is reflected in the transform domain in the way that

lower regions have more energy in the case of voiced sounnds and higher regions are more pronounced in the unvoiced sound case. This provides a good basis for separating these sounds in the transform domain and processing them separately in order to achieve more efficient systems.

Speech is considered to be in the silence mode where no sound is produced by vocal cords.

Typical spectrum shapes for voiced and unvoiced sounds in frequency and Walsh domains are shown in Figures 3.6 and 3.7.

It can be seen from Figure 3.6 that a high degree of similarity exists between voiced spectra in the frequency and Walsh domains. This fact is basically the same in other orthogonal domains.

This thesis uses the frequency and Walsh domains for digital speech processing but of course in general other orthogonal domains can also be used.

3.5 SPECIAL FEATURES OF THE VOICED SOUNDS

The voiced sounds are periodic waveforms which are shaped by the vocal tract movements. Their spectral representation contains distinct peaks corresponding to the resonances of the vocal tract, called Formants (11), (67), (69).

The voiced sounds can be classified into vowels and voiced consonants. A convenient way of classifying the speech sounds is in terms of the articulating configurations that produce them. In this respect the vowels and consonants differ, since the former result from relatively stable configurations whereas the latter

are produced by changing ones (67), All vowels are voiced sounds, but consonants are a mixture of voiced and unvoiced sounds. In the case of vowels, the resonant frequencies (Formants) are almost stable during the whole vowel period, which makes it possible to classify them in terms of the first two formants.

In the case of voiced consonants, due to the changing of the articulation system during the sound production, the formants are not stable and have rapid variations during their articulation. It is due to this reason that consonants can not be classified in a similar manner to vowels (i.e. on a basis of their formant positions.) Consonants can be divided into different groups; Nasals, Semi-vowels and Liquids are voiced sounds and Fricatives and Stops can be either voiced or unvoiced sounds (14) (67).

A List of vowels and consonants are given in Table 3.2.

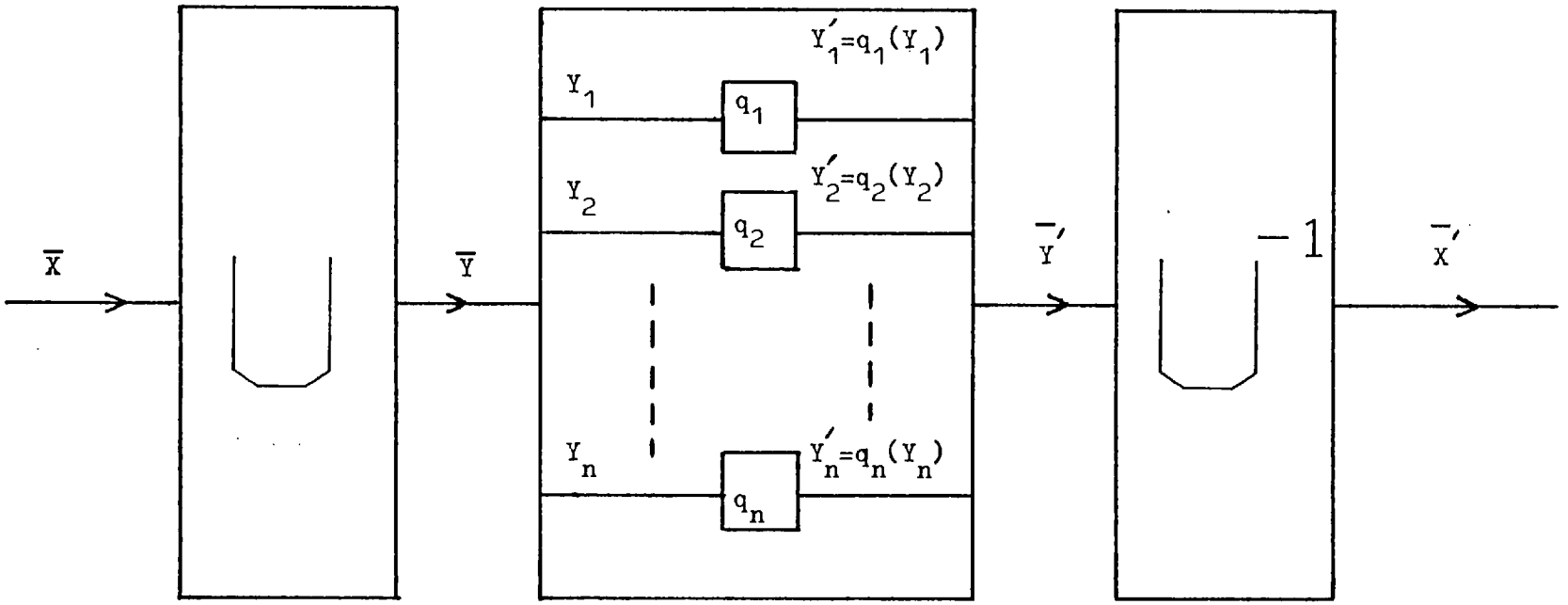


Fig. 3.1: Block Diagram of a Transform Coder

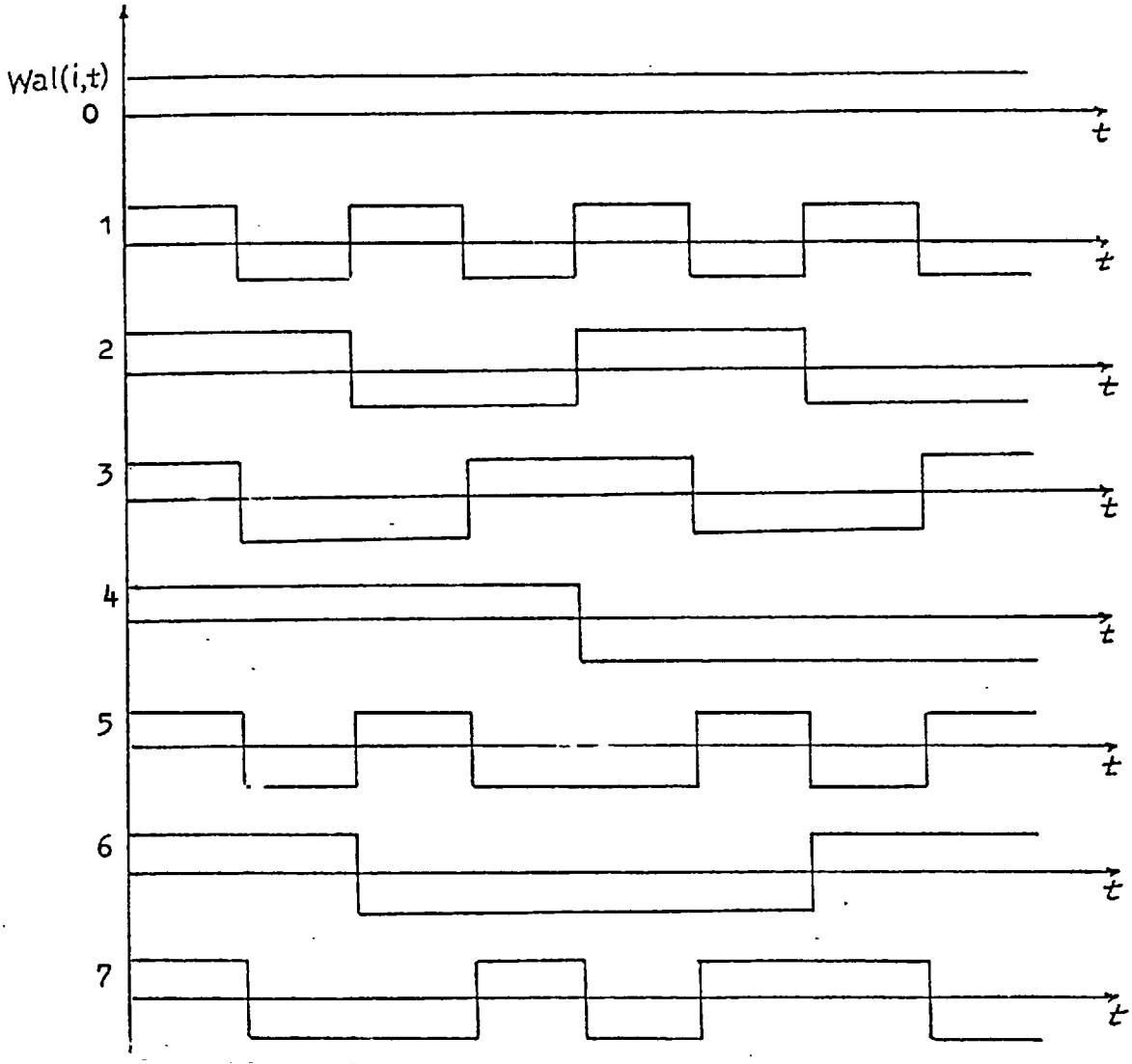


Fig. 3.2: Walsh functions of order 8

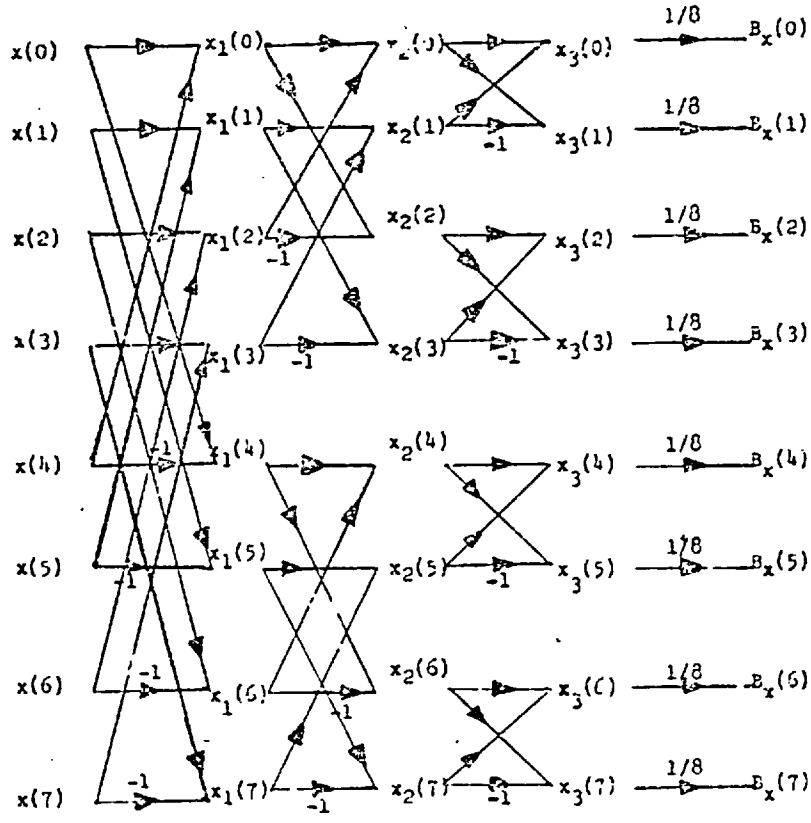


Fig. 3.3: Signal Flow Graph of Fast Walsh Computation

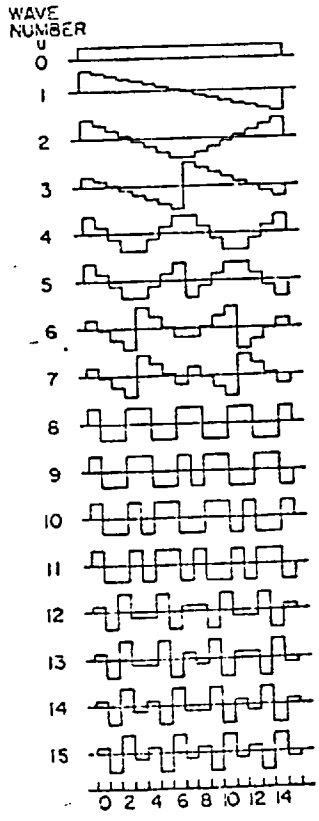
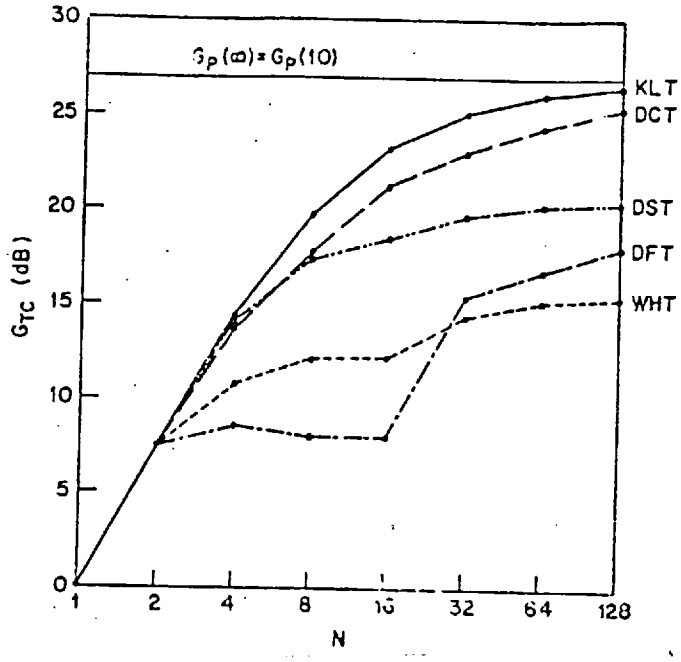
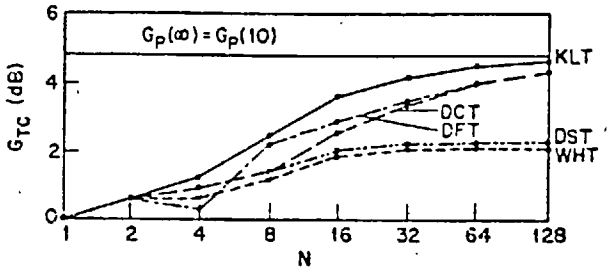


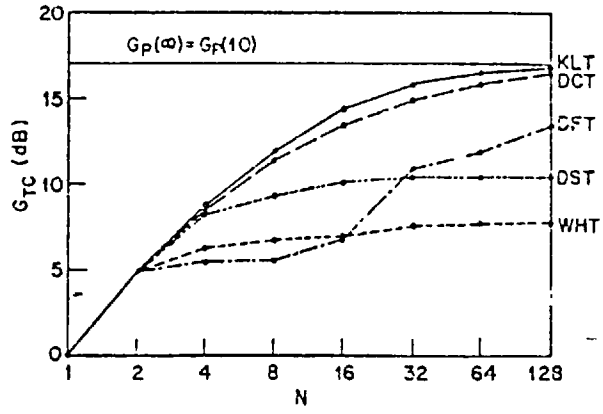
Fig. 3.4: Slant Transform basis vector



For Speech Sound "m"



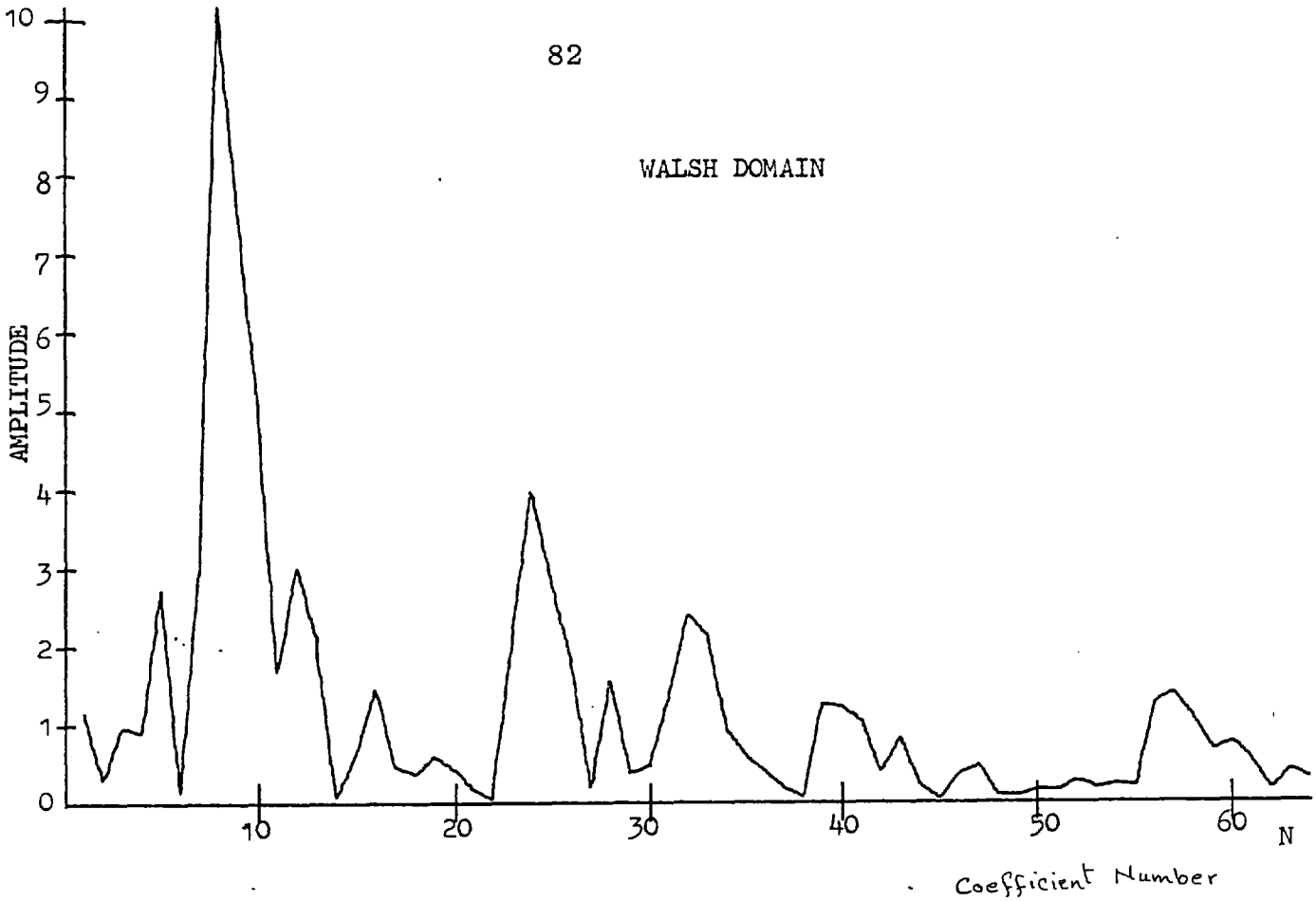
For Speech Sound "sch"



For Speech Sound "ü"

(G_{TC}=Theoretical Gain)

Fig. 3.5: Transform domain behaviour for different Speech Sounds



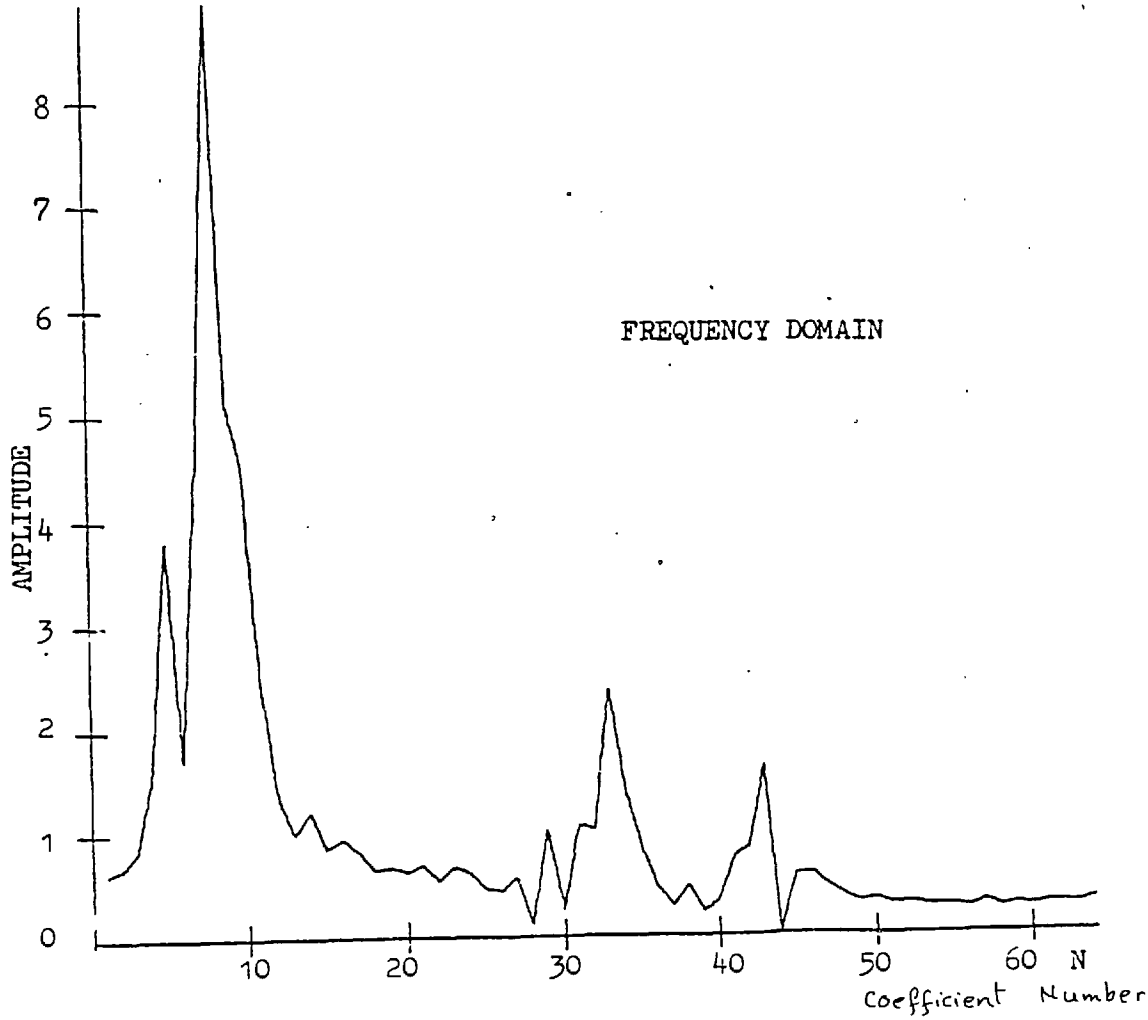


Fig. 3.6: Typical Spectrum Shapes for voiced sounds in Frequency and Walsh domains

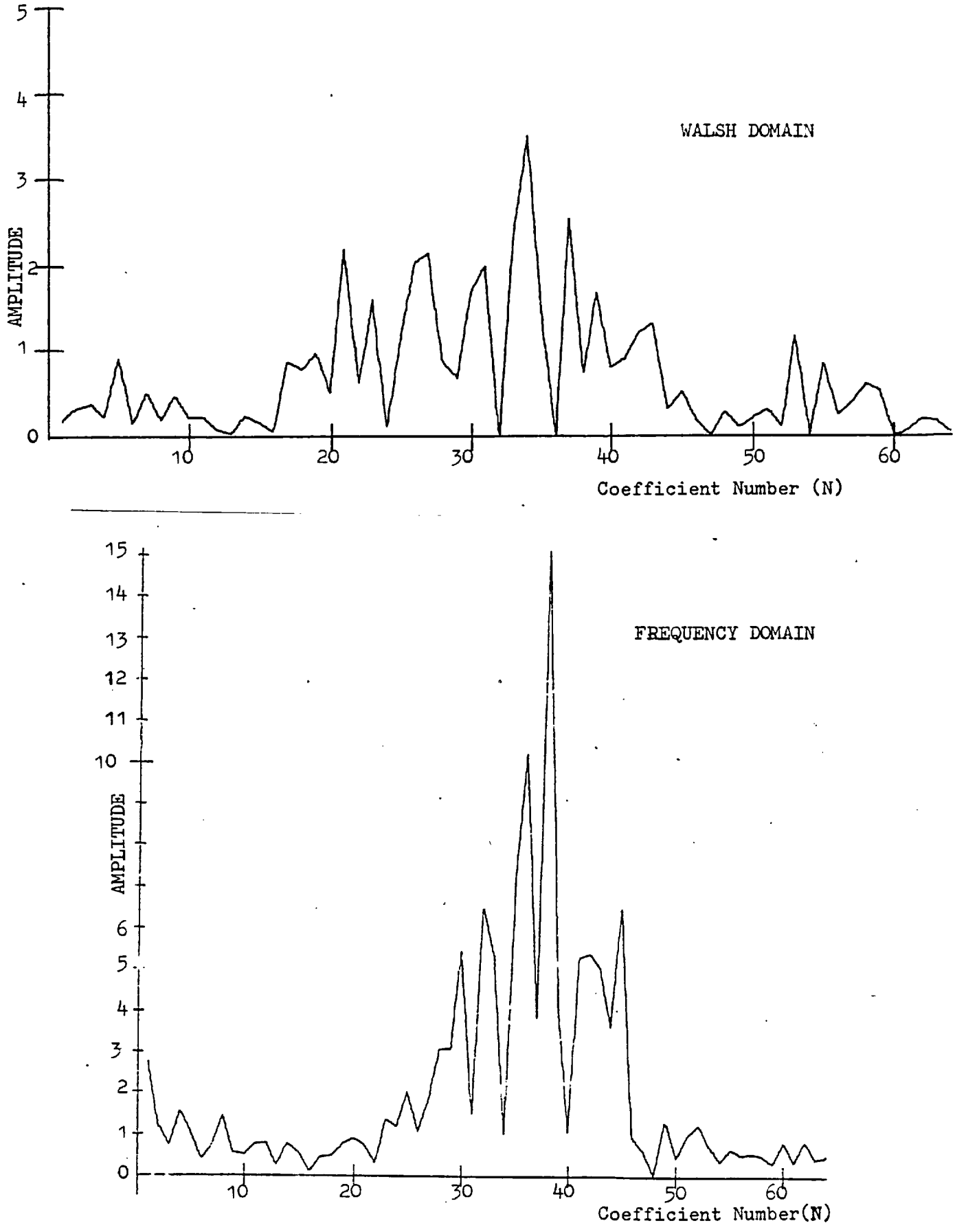


Fig. 3.7: Typical Spectrum shapes for unvoiced sounds in Frequency and Walsh domains

Transform Domain	Advantages	Disadvantages
1)Karhunen-Loeve (KL)	a) Provides minimum mean square error coding performance. b) Optimum transform,adaptive to input signal variations.	a) Does not have a fast algorithm for its computation. b) Transform dependent on the input signal.
2)Discrete Cosine (DC)	a) Sub-optimum transform;(nearest to KLT when compared with other transforms). b) Approaches KLT for large transform sizes. c) The basis vectors of the transform can be sorted according to their sequency. d) Its basis vector closely approximates to the eigenvectors of the Toeplitz matrices (speech type). e) A fast algorithm is available.	a) Not signal adaptive.
3)Discrete Fourier (DF)	a) Approaches KLT for large transform sizes. b) The basis vector can be sorted according to their sequency. c) A fast algorithm is available for its fast computation.	a) Diverts from optimum performance for small transform sizes. b) Needs complex operations.

Table 3.1: Comparison of Transform Domains (cont.)

Transform Domain	Advantages	Disadvantages
4)Walsh-Hadamard (WH)	a) Very simple to implement as it requires only additions and subtractions. b) Very fast computation.	a) Far from optimum for small and medium size transforms.
5)Haar (H)	a) The average number of operations is independent of transform size. b) Efficient computational algorithms available.	a) Results in large coding errors. b) Difficulties arise if transmitted over noisy channels.
6)Slant (S)	a) High energy compaction. b) Fast algorithm for its computation. c) Suitable for image coding (it has a staircase type basis vector).	a) Not suitable for constant valued or quasi-constant vectors.
7)Number-Theoretic (NT)	a) Perform convolutions as exact operations. b) Fast algorithms available	a) Has limited levels. b) Has finite wordlengh.

Table 3.1: Comparison of Transform Domains

VOWELS			
Degree of constriction	Tongue hump position		
	Front	Central	Back
High	/i/ eve	/ɝ/ bird	/u/ boot
Medium	/I/ it	/ɐ/ over(unstr.)	/U/ foot
	/e/ hate	/ʌ/ up	/o/ obey
	/ɛ/ met	/ə/ ado (unstr.)	/ɔ/ all
Low	/æ/ at		/a/ father

FRICATIVE CONSONANTS		
Place of articulation	Voiced	Unvoiced
Labio-Dental	/v/ vote	/f/ for
Dental	/ð/ then	/θ/ thin
Alveolar	/z/ zoo	/s/ see
Palatal	/ʒ/ azure	/ʃ/ she
Glottal		/h/ he

STOP CONSONANTS			NASALS
Place of articulation	Voiced	Unvoiced	Voiced
Labial	/b/ be	/p/ pay	/m/ me
Alveolar	/d/ day	/t/ to	/n/ no
Palatal/velar	/g/ go	/k/ key	/ŋ/ sing

GLIDES		SEMI-VOWELS or LIQUIDS	
Place of articulation	Voiced	Place of articulation	Voiced
Palatal	/y/ you	Palatal	/r/ read
Labial	/w/ we	Alveolar	/l/ let

DIPHTHONGS	
Voiced	
/eI/ say	/aU/ out
/Iu/ new	/aI/ I
/ɔI/ boy	/oU/ go

AFFRICATES	
Voiced	Unvoiced
/dʒ/ jar	/tʃ/ chew

Table 3.2: List of Vowels and Consonants.

CHAPTER FOUR

DIGITAL SPEECH PROCESSING BASED ON FORMANTS

4.1. FORMANTS IN THE TRANSFORM DOMAIN

Formants are the resonances of the vocal tract present during the production of voiced sounds. Their presence can be detected by sharp peaks in certain regions of the transform domain, revealing the fact that the formants have high energies. It has been shown that they make a large contribution to the perception of the relevant sounds (49,50).

Most investigations in speech processing have been carried out in the frequency domain, but in principle due to the similarity of other orthogonal domains to frequency similar speech characteristics exist in these domains also. The system which is discussed in this chapter for achieving low bit rate digital speech transmission in fact employs the Walsh-Hadamard transform. This transform does exhibit properties similar to the frequency domain with many attributes carrying over e.g. frequency being equivalent to average zero crossing etc.

4.2 DIFFERENT FORMANT DETECTION METHODS IN THE TRANSFORM DOMAIN

It is a well-known fact that an important role is played by the formants in the perception of voiced sounds in speech (49,50) and due to this importance, systems based on formants and the regions surrounding them have been designed and used for speech processing (47, 48, 49, 50).

To achieve a simple and efficient system, the presence of formants must be then detected by effective methods with

relatively simple algorithms. Moreover, the effect of small errors in the detection of formants does not have a high perceptual significance. This is evident from the vowel categorization studies, represented by $F_1 - F_2$ (a graph of first formant location against second formant location), as shown in Figure 4.1.

In the vocal classification it has been shown that so long as the $F_1 - F_2$ point is within a specified region in that plane, the vowel sounds almost the same, no matter which part of the region it is placed in.(75) The boundaries between the defined regions overlap in some cases as seen from Figure 4.1.

Similar error insensitivities exist for the formant amplitudes also, in as much as the effect on perception is negligible, if their amplitudes are not coded exactly. The methods used for formant detection in the transform domain do not need to be exact, in as much as they give the information content of the formants which can be used to recover the speech with a minimum perceptual distortion.

There are many different methods employed for detecting the formants in the transform domain.

Among these the following are more effective for the purposes of this system.

1) Peak Picking method
2) Spectral matching (analysis - synthesis) approach.
3) Spectral envelope method.
4) Moments approach.
5) Linear Predictive Coding approach.

These approaches have employed frequency as their analyzing domain but due to similarity between the orthogonal domains other domains could also be used.

It should be noted at this stage that the formants have a variable position depending on the speaker and in addition it is not possible to define distinct regions for different formants as they tend to overlap. One convenient method for formant detection is to consider separately regions for male and female speakers and base formant detection methods on these formant regions. Indeed the average regions of the first four formants for male speakers in the frequency domain are (67):

First formant Region (F_1): 200 to 900 Hz

Second formant Region (F_2):550 to 2500 Hz

Third formant Region (F_3): 1500 to 32000 Hz

Fourth formant Region (F_4):2500-4500 Hz

The fundamental pitch of the female voice is on average one octave higher than that of the male voice but formant frequencies are only on average 17 per cent higher (67). Children about 10 years of age have still higher formants, which on average are 25 per cent higher than the corresponding formants for adult males and their fundamental pitch is approximately 300 Hz; the variance of this result however is high.

4.2.1. PEAK PICKING APPROACH

One of the relatively simple and effective formant tracking systems in the transform domain is the peak picking approach (11,72). The method is used to detect simply the peaks in the short-time speech signal amplitude spectrum. Flanagan suggested two methods for this

approach (11, 14). One is based on locating points of zero slope in the spectral envelope, and the other is the detection of local spectral maxima by magnitude comparison.

In the first approach, a short-time amplitude spectrum is first produced by a set of bandpass filters. The outputs of the filter channels are scanned rapidly, of the order of 100 times per second. This produces a time function which is a step-wise representation of the short-time spectrum at a number of frequency values. For each scan, the time function is differentiated and binary-scaled to produce pulses marking the maxima of the spectrum.

The marking pulses are directed into separate channels by a counter, where they sample a sweep voltage produced at the scanning rate. The sampled voltage are proportional to the frequencies of the respective spectral maxima and are held constant during the remainder of the scan. The resulting stepwise voltages are subsequently smoothed by low-pass filtering.

The second method segments the short-time spectrum into frequency ranges that ideally contain a single formant. The frequency of the spectral maximum within each segment is then measured. In the simplest form the segment boundaries are fixed. However, additional controls can be used to adjust the frequency range of a given segment. The maxima of each segment are selected at a rapid rate and a voltage proportional to the frequency of the selected channel is delivered to the output. These two methods can be simulated and different algorithms can be used for th formant detections. Non-iterative approaches are appealing due to their comparative computational efficiency. These approaches

often depend on detecting spectral peaks and identifying them as possible formants.

Cepstral methods have also been used to obtain smoothed spectra from which the formants are obtained by peak picking via human intervention or by computer (16).

4.2.2. ANALYSIS BY SYNTHESIS APPROACH

This is a frequency domain iterative approach in which adjustments are made on the parameters of a speech synthesis model until some desired degree of matching is obtained between the actual speech spectrum and the spectrum resulting from the model (68). The best spectral match is normally achieved using a minimum mean square error criterion. The success of this technique depends on the accuracy of the speech model. The performance of the system is best for the vowels during which the formants (and spectra) are almost stationary.

Using formant continuity constraints a formal searching achieves better results.

Analysis by synthesis permits great flexibility in achieving spectral matches but requires extensive processing in its iterations.

4.2.3. THE SPECTRAL ENVELOPE METHOD

Another method of obtaining vocal tract resonances is by the spectral envelope approach (69). To do this the spectral envelope is obtained by lowpass filtering the log-magnitude of the discrete Fourier transform. Then, in the cases where the first three formants (F_1, F_2 and F_3) are separated by more than about 300 H_z, there is no difficulty in resolving the corresponding peaks in the smoothed spectrum. However, when F_1 and F_2 or F_2 and F_3 are closer than 300Hz, cepstral

smoothing (the Cepstrum of a segment of a sampled speech waveform is defined as the inverse transform of the logarithm of the Z-transform of that segment) results in the non-resolution of the peaks. In these cases a new spectral analysis algorithm called the chirp Z-Transform (CZT) can be used to advantage (70). In similar cases, another method has been suggested using the spectral envelope approach (69).

In this approach, once the spectrum is obtained by filtering the relevant cepstrum, then all the peaks are located, and the location and amplitude level of each peak is recorded. This information of the peak locations and peak levels is then used for the estimation of the formants, the algorithm for which can be found in reference (69).

4.2.4. MOMENTS APPROACH

The method of formant detection by moment calculations (71) is a compromise between the analysis by synthesis method and the peak picking approach. This is in fact a compromise between obtaining the formants from the whole spectrum pattern through a feedback process of active analysis which requires relatively long processing time and detecting the formants from a limited and local information of the spectrum which takes less time.

The nth moment, M_n of an amplitude spectrum A(w) is given by:

$$M_n = \int W^n A(W) \, d W$$

where W is the radian frequency.

If a suitable pre-filtering or partioning of the spectrum can be achieved, then a formant frequency can be approximated by

$$\bar{W} = \frac{M_1}{M_0} = \frac{\sum_i W_i \, A(W_i)}{\sum_i A(W_i)}$$

Difficulties in partitioning the spectrum accurately and in the asymmetry or skewness of the spectral resonances due to which the measured formant may be weighted towards the heavier side of the spectrum rather than placed at the spectral peak.

A pair of parameters are used to define the frequency range of each formant in the approach of ref. (71) where the formants are estimated from first order moment calculations. Two main sources of error in this method are the harmonic structure of the spectrum and the coexistence of antiformant frequencies in the spectrum. The former is the more serious for the female voice.

4.2.5 THE LINEAR PREDICTIVE METHOD (LPC)

The LPC is one of the most effective techniques for formant analysis (48). It is based on the general linear prediction method used for speech processing (Section2.5) which leads to relatively accurate formant estimation. A block diagram of the procedure of formant analysis by the LPC technique is shown in Figure 4.2. The preprocessing section includes low-pass filtering, analogue to Digital

Conversion, preemphasis (for better second and third formant detection) and speech detection. The prepocessed speech is then windowed by an appropriate window e.g. a Hamming or Hanning window (59). A sampled speech signal S (nT), at discrete times t = nT, is then linearly predicted by the past P samples using the formula:

$$\hat{S}_n = \sum_{K=1}^{P} a_K S_{n-K}$$

where $\hat{S}_n$ is the predicted value of S_n, and a_K are the Prediction Coefficients. The speech spectra can then be obtained by a Fourier transformation of the LPC coefficients, using the pruned fast Fourier transform (73), and the relevant formants can be detected by picking the maxima from the obtained spectrum. If the formants are too near to each other, methods such as the one suggested in Section 4.2.3 or the chirp Z-transform method (70), or the double difference peak picking method (72) can be used.

The approaches discussed in this Section are not the only ones available. Moreover, they illustrate many of the basic principles involved in the estimation of formants, and greater accuracy can be achieved if different techniques are combined together, at the disadvantage of having more complex structures.

4.3. SPEECH PROCESSING SYSTEMS BASED ON FORMANTS

There are many speech processing systems, both analogue and digital, which use the vocal tract resonances (Formants) as their principal aspect in the processing of speech. The main advantage of such systems is their low-information rate (bandwidth) requirements for speech transmission.

Generally these systems can be considered as having two major parts, the transmitter and the receiver. The transmitter is the analyzying section of the system which detects the formants and other necessary information for speech recovery. The receiver combines the received information and recovers the speech waveform. This is refereed to as the synthesizing section.

Of the formant based speech processing systems, the following are shown to represent the total existing systems.

1) Formant Vocoder;

This analogue system uses formants as its main parameter for processing the speech waveforms. Historically, it is one of the first systems of this group (14). A brief discussion on this system is presented in Section 2.4.2.

2) Sub-band System for Speech Coding.

In this digital system regions around the formants are extracted from the short-time spectrum and transmitted since they represent the highest information bearing regions of the spectrum, and an adaptive strategy needs to be followed to back the changes in formant positions (42). More details on this system is given in Section 2.4.3.

3) Digital Formant Vocoder.

This is similar, in principle, to the formant vocoder mentioned above. The LPC technique is used for the detection of the formants which results in a low transmission rate system. (48).

4) Speech Recognition Systems.

The formant analysis has also been used in the recognition of sounds (75) (83), and have also been used as a

method to enable the deaf to recognize different sounds (76). Although the systems mentioned above are frequency based, other orthogonal transforms can conceivably be employed.

4.4 A NEW FORMANT BASED SYSTEM FOR SPEECH PROCESSING

Our aim here is to design a simple digital system leading to low information-rate for the processing of speech signals. For this purpose the speech waveform is transformed into domains other than time as the sample-to-sample correlations can be exploited more easily. Among the orthogonal transforms, the simplest and the most suitable for digital systems is the Walsh, or Sequency, domain

To ensure simplicity, the non-pitch synchronous approach is taken as the pitch detection adds complexity to the system. The Speech Signal is segmented into blocks of constant length and the choice of length for the appropriate transform is discussed in the following section.

The system discussed in this chapter, uses the formants as the most important information bearing parameters, and is referred to as The Formant Sequency Speech Coding System. The rest of this Chapter is devoted to discussions on different parts of the proposed system.

4.4.1 THE CHOICE OF TRANSFORM DOMAIN

In the previous Chapter discussion was devoted to different transform domains available for digital signal processing. For speech processing, different orthogonal transforms can be employed. Among these, the Karhunen-Loève transform (KLT) is the optimum for any given

second-order statistics and is obtained from the eigenvalues of the input data (56). As speech is a non-stationary signal the K-L transform coefficients have to change, in accordance with the changes in the speech waveform. This results in a rather complex system with relatively high transmission rate requirements (Section 3.2.1).

The search is therefore directed towards finding sub-optimum transforms which are not signal dependent and result in simple structures.

A very good sub-optimum transform for speech signals is the Discrete Cosine transform (28). The performance of DCT is compared with the KLT and other orthogonal transforms and is presented in Figure 3.5 (Chapter 3) (40, 41).

The DCT can be implemented using fast algorithms, the drawback of this system being its relative complexity. Another transform is the Discrete Fourier Transform which is a well known analysis domain in the signal processing field. It can be computed through a fast algorithm (Fast Fourier Transform), but as it requires complex operations for its computation, it makes the system fairly complex. (Section 3.2.3).

The next interesting orthogonal transform used in the signal processing field is the Discrete Slant transform and is best suited for Image Processing (29) (Section 3.2.5.).

Another interesting transform domain is the Walsh domain. It is formedby square waves, for digital applications is very attractive as no multiplications are involved in the computation and only shift register and accumulators

are used which makes it quite simple compared with other transform domains.

Although its Signal to Noise ratio performance is not as good as other already mentioned orthogonal transforms the fact that it is quite simple compared to others gives it an advantage for applications where simplicity is of prime importance. Due to this reason it was felt that systems based on Walsh-Hadamard transform (WHT) (discrete version of Walsh transform) ought to be designed and their performance be studied. Previously some work has been carried out with the WH domain for the processing of speech (23, 24, 27).

A new system for digital speech processing has been designed, employing the Walsh Hadamard Transform, the details of which are given in the rest of this section.

4.4.2 SIMILARITY OF WALSH AND FOURIER TRANSFORMS

The Fourier-transform is a well known transform to any researcher in the field of signal processing.

In order to familiarize ourselves with the Walsh-transform a mathematical comparison between the Fourier and Walsh transforms is given below.

It is well known that in Fourier analysis the signal is represented as a weighted sum of sine and cosine waveforms. In the discrete case the best rerpesentation of a periodic function f(t) can be obtained using the orthogonal sine and cosine series:

$$f(t)=a_0 f(0,\Theta)+\sqrt{2}\sum_{i=1}^{\infty}\left(a_c(i)\cos 2\pi i\Theta + a_s(i)\sin 2\pi i\Theta\right) \quad (1)$$

where $\Theta = \frac{t}{T}$, T is periocity of the waveform and coeff-

icients a_c (i) and a_s(i) are known as the Fourier Coefficients.

Similarly, using Walsh analysis a periodic function can be represented as sum of square waves:

$$f(t)=a_0\ \mathrm{Wal}(o,\Theta)+\sum_{i=1}^{\infty}(a_c(i)\mathrm{Cal}(i,\Theta)+a_s(i)\mathrm{Sal}(i,\Theta)) \qquad (2)$$

where Θ is the orthogonality interval, i is the sequency and Cal and Sal are odd and even periodicity. Here a_c(i) and a_s(i) are not the same as those of equation (1) they have been chosen to indicate the analogy. It can therefore be stated that the relation between the Walsh and Fourier coefficients is a one to one linear transformation (77).

4.4.3. BLOCK SIZE DECISION

The aim of proposing the Formant Sequency Speech Coding system is to achieve a simple and effective system. To achieve this aim Walsh-Hadamard Transform was chosen as the analyzing domain. Furthermore, a fixed transform length is used instead of a pitch adaptive one, as the latter results in a more complex system than the former.

The decision would have to be taken on the length of the transform which would stay unchanged throughout the analysis period. The transform length should be such as to ensure availability of sufficient spectral resolution to capture the fine details of the speech spectrum.

Also, as formant detection is the major element in the proposed system, the transform length should be short enough such that the effect of formant transitions especially in the consonant sounds, is a minimum.

As the voiced-unvoiced decision is made for each transform length (Section 4.4.4), and each type is processed separately, then the transform length should be short enough such that it minimizes the effect of misjudgement in the V-UV decision.

Finally if a small transform size is used the system will have to update the information very often which results in a higher information rate.

The compromise seems to be a transform length of 64, this is equivalent to 8 msec segments of the speech waveform for low pass filtered speech to 4 KHz and sampled at 8 KHz. This results in a sample resolution of 125 μ sec. This transform length is good enough from the formant transition point of view(see also Chapter 6), and a reasonable for the voiced-unvoiced decision. It also produces reasonable resolution for obtaining fine spectral details.

This transform length has been used by other researchers as well and is likewise claimed to be a reasonable one (57, 64).

More discussion on the transform length regarding the formant transitions is given in Chapter 6.

4.4.4. VOICED-UNVOICED DECISION

For achieving an efficient system of speech processing different speech sounds, voiced and unvoiced segments ought to be processed differently. To separate voiced-unvoiced

sound segments many different methods have been used. Most of these methods are based on the detection of the pitch, the fundamental frequency of the vocal cords in the case of voiced sounds. The existence of the pitch indicates the presence of the voiced sound, whereas the lack of pitch indicates an unvoiced sound segment. From the voiced-unvoiced decision making methods, the following are noteworthy.

1) Buzz-hiss detection. 2) Energy Ratio method. 3) Autocorrelation Approach (Autoc). (4) Cepstral method (CEP). 5) Simplified inverse filtering technique (SIFT). 6) The parallel processing time domain method (PPROC). 7) The data Reduction approach (DARD). 8) Spectral flattering linear Prediction Coding (LPC). 9) The average magnitude difference function (AMDF). 10) Maximum Likelihood Estimation Approach. 11) Pattern Recognition Approach.

Most of these approaches are based on the detection of pitch. More detail on these systems can be obtained from references (78) and (79).

As the proposed system here is a non-pitch synchronous approach, the voiced/unvoiced decision is made based on other features of speech rather than pitch period.

It is generally difficult to decide on the voiced or unvoiced sound exactly as the transition between voiced and unvoiced sound segments and vice-versa takes approximately 2 msec.

The method which has been employed here for the voiced/unvoiced (V/UV) decision is based on the energy distribution of each sound in the analyzing domain. The decision is made, based on the ratio of the energy

available at the lower and higher frequency (sequency) regions of the short time speech spectrum (see Section 3.4 also).

In the case of the orthogonal transforms, the ratio of the energy at the lower part of the spectrum to the total spectrum energy would indicate, to a high degree of reliability, the type of the speech sound. This is due to the different energy distributions of each sound segment as mentioned in Section 3.4.

The difference in the variance distribution, in the Walsh Hadamard domain, of the unvoiced sounds, to the voiced sound was first noticed by Gethoffer (80) for German speech. This is shown to be similar for English speech (57) and is illustrated in Fig. 3.7.

The energy distribution for each sound depends on the correlation between the speech samples in the transform domain. As a fixed transform strategy is employed in the proposed system to separate the two speech sounds, the energy of different sections of the spectrum is considered for each transform block, in this case 64 samples.

The proposed ratio is given as (57)

$$R = \frac{\text{Energy of the lower spectrum region}}{\text{Energy of the total spectrum}}$$

where R is the ratio calculated for every block of 8 msec segments of speech (64 samples). To obtain a proper ratio for each transform domain, statistical investigation is carried out for each case. For the proposed system the statistical investigation was carried out for Walsh Hadamard transform and later on for the Fourier transform

in Chapter 6.

The sentences which were used for the investigations here, and for testing the systems later in the work, were obtained from the recommended Harvard list of balanced sentences which had been approved by the Institute of Electrical and Electronic Engineers (IEEE). (81).

The region considered for the calculation of the value of R, in the 64 point WHT was from samples 1 to 12, considered as the low sequency region. So the formula for the Ratio would become

$$R = \frac{\sum_{i=1}^{12} C_i^2}{\sum_{i=1}^{64} C_i^2}$$

where C_i is the ith coefficient of each block of the 64 transform. The value of 0.3 for R was decided to be a good threshold level below which the sound is considered unvoiced and above which it is considered a voiced sound. A graph of energy distributions for different threshold values is given in Figure 4.3.

For other orthogonal domains, a similar approach can be taken where the threshold would be different and can be obtained through statistical evaluation.

4.4.5. FORMANT DETECTION TECHNIQUE

Following the voiced-unvoiced decision discussed in the previous section, each type of sound segment is processed separately. In the case of voiced sounds, the formants are detected as the information bearing parameters, and transmitted to the receiver. In the unvoiced case,

several coefficients are transmitted as discussed in the next section. In this present section the voiced sound processing is discussed.

Different formant detection methods were mentioned in Section 4.2. Among the methods mentioned, those which do not require pitch detection are useful for the proposed system here for simplicity and the constant block transformation approach taken in this system. The method chosen for formant detection is the peak picking method (11, 72) discussed in Section 4.2.1. Other non-pitch dependent methods of formant detection could also be used.

In the peak picking method, the appropriate maxima are searched for in the short-time spectrum of the voiced sounds. in the relevant regions for formant existence. These regions are given as 200 to 900 Hz for the first formant, 550 to 2500 Hz for the second formant and 1500 to 3200 Hz for the third formant which correspond to coefficients 3 to 12 for the first formant, coefficients 10 to 35 for the second formant and 27 to 50 for the third formant in the 64 point transform considered here. This is obtained from statistical investigation using a number of sentences.

The Double Difference peak picking method can also be used in the proposed system (72).

4.4.6 UNVOICED SOUND PROCESSING METHOD

The approach taken in the case of voiced sounds was mentioned above. Here the method of unvoiced speech sounds processing is discussed.

It is a well-known fact that the unvoiced sounds have energy at high regions of the frequency spectrum, above

2 kHz normally. This is also similar in other orthogonal domains. In the Walsh Hadamard domain particularly the energy is on average concentrated in the mid-sequency region, as can be seen from Figure 4.3.

This has been obtained from statistical analyses over several balanced sentences from the Harvard list (81). The high energy region in the WH domain on average is from coefficients 25 to 40 for a 64 point transform.

This suggests that for transmitting the unvoiced sounds, the mid-sequency region is to be considered as the most important region to be transmitted. This has therefore been used for the unvoiced sound processing in the proposed system.

4.4.7. COEFFICIENT QUANTIZATION

There exist many different methods of parameter quantization for the digital transmission of signals. The basic ones are Pulse Code Modulation (PCM), Adaptive PCM (APCM), Differential PCM (DPCM), Delta Modulation (DM), Adaptive Predictive Coding, (APC), and others which use the same general principles (18, 21, 22, 34, 35, 36, 38, 43, 44, 45). A detailed discussion on each of these systems is given in Chapter 2.

As speech is a non-stationary signal with a large dynamic range, and has an exponential (second order) probability density function (PDF) it follows that to achieve minimum quantization noise and at the same time an efficient system, an adaptive quantization technique should be used.

The general idea is to increase the step size during overload and decrease it during granularity. Two forms of the step size adaptation may be employed (i) instantaneous (ii) syllabic methods. Logarithmic quantization methods can also be used for the exponential PDF class of signals (13).

The two adaptive quantization techniques which have widely been used are, (a) Syllabically adaptive PCM having the advantage of a better tailoring to the signal and a high resistance to errors. (b) Instantaneously adaptive quantization having the advantage of minimal structure and good noise protection.

Among the quantizers discussed above, the adaptive ones are the most suitable for speech signals. The adaptation technique which was employed in the work presented here is adaptive quantization with a one-word Memory. (36).

In this adaptation technique, the quantizer adapts its step size for every new input sample by a factor depending only on the knowledge of the previously used step. If the outputs of a uniform B-bit quantizer (B > 1) are of the form:

$$Y_u = P_u \frac{\Delta_u}{2}$$

where $\pm P_u = 1,3,.....,2^B-1$ and $\Delta_u > 0$

The step-size Δ_r is given by the previous step-size multiplied by a time-invariant function of the code-word magnitude $|P_{r-1}|$

$$\Delta_r = \Delta_{r-1} M(|P_{r-1}|)$$

where the step size multiplier M is a function only of the latest quantizer output. Since adaptations follow quantizer output rather than input, step size information in this system does not have to be explicitly sent out but can be recreated exactly by the receiver, provided that error-free transmission is assumed.

"M" can have a maximum of $2^{(b-1)}$ values, where b is the total number of bits allocated to each sample. Thus each slot from a number of 2^{b-1} slots will correspond to the multiplication factor.

More details on the quantizer can be found in the reference (36).

4.4.8 OVERALL SYSTEM DESCRIPTION

The proposed system for low bit rate digital speech transmission consist of transformation of time domain signals into the Walsh-Hadamard domain where the formants are detected as the most important parameters in the case of voiced sounds, and high energy coefficients in the case of unvoiced sounds. These parameters are selected and transmitted to the receiver end, the speech signal is then recovered by replacing the appropriate coefficients and transforming the speech back into the time domain.

The details of the proposed system are as follows:

First, the analogue speech is low pass filtered to 3.04 KHz. Then a sampling rate of 8 KHz is employed and digitization of the waveform in time is obtained by using a 10 bit Analogue to Digital Convertor. The obtained time domain samples are then taken in segments of 8 msec., that is 64 samples, and transformed into the Walsh-Hadamard

domain using a fast Walsh-Hadamard Transform (FWHT) algorithm as mentioned in Section 3.2.4. A decision is made on the obtained segment spectrum for its classification into a voiced or unvoiced sound.

The decision strategy is discussed in Section 4.4.4. Silent sections of the speech can be treated as unvoiced sounds with very low energy in the spectrum (they may be considered as very low level noise).

In the case of the voiced sounds, the first three formants are detected from the relevant regions using a peak picking method as discussed in the Section 4.4.5. For unvoiced sounds, some high energy content coefficients are chosen as shown to be perceptually the most important parameters (Section 4.4.6). Then the so obtained coefficients are digitzed in amplitude using an adaptive strategy mentioned in Section 4.4.7 employing different step size levels for the voiced and unvoiced case. After the coefficient quantization, the relevant parameters are transmitted to the receiver.

At the receiver, the received information is put together in the appropriate manner to reconstruct the speech spectrum for each voiced or unvoiced block. Time domain speech is then recovered by performing an Inverse Fast Walsh Hadamard Transform (IFWHT) on each block of the transform domain speech.

The proposed system was simulated on a PDP-15 computer, the results of which are discussed in the next section.

4.4.9 COMPUTER SIMULATION RESULTS

A computer program was written to simulate,on a PDP-15 computer,the digital speech processing system proposed in this chapter.

As mentioned above, the 3.04 KHz low pass filter analogue speech is sampled at 8 KHz where digital speech is obtained using an A/D converter. Then, a 64 point Walsh-Hadamard transform was performed on 8 msec intervals of speech signals.

The voiced/unvoiced decision is them made subject to different energy distributions of these sounds in WH Spectrum domain (Section 4.4.4.).

For voiced sounds the regions of formant existance are not fixed and are approximately 200 to 900 Hz for the first formant, 550 to 2500 Hz for the second formant region and around 1500 to 3000 Hz for the third formant. In the Walsh Hadamard domain these regions are similarly obtained where for a 64 point transform used in this system the formant regions are coefficients 3 to 12 for the first formant, and 10 to 35 for the second formant and 27 to 50 for the third formant. Then the peak picking method (Section 4.4.5) can be used to detect the first three formants from the 64 point WHT. These three parameters can be transmitted after quantization, as perceptually the most important parameters of the voiced segment spectrum. Because the low sequency region of the WH spectrum represents the primary speech energy for voiced sounds and is very important perceptually, then part of this region can always be transmitted to obtain better speech quality.

This is based on the sub-band principle which has been applied to the frequency domain with successful results (49, 50). For quantization of the coefficients the adaptive quantization technique with one word memory is used (Section 4.4.7) with the multiplication factors as $M_1 = 0.9$, $M_2 = 1$, $M_3 = 1$ and $M_4 = 1.75$, 3 bits per coefficient being used in this case. For better quantization, 4 or more bits per coefficient can be used which results in a higher overall transmission rate but not much difference in the speech quality.

In the case of unvoiced sounds, high energy coefficients are chosen for transmission. These are coefficients 25 to 40 in the 64 point WHT spectrum, (Section 4.4.6). The same quantization technique as in the voiced sound case was used here but 2 bits per coefficient proved to be sufficient in this case, where the multiplication factors are $M_1 = 0.845$ and $M_2 = 1.96$. If low bit rate speech transmission is required, the unvoiced coefficients can be reduced and only the highest valued ones transmitted.

For the case where only three formants are transmitted for voiced sounds, the bit rate can be given as follows:

Bits needed for informing the receiver of the position of the coefficients:

First formant	3 bits
Second formant	5 bits
Third formant	5 bits
Total bit for Addressing	13 bits

Coefficient quantization:

For each coefficient 3 bits

Therefore for total quantization 9 bits

For indicating whether the segment is voiced or unvoiced one bit is needed, so the total bits needed for voiced sound transmission is

Address	13 bits
Quantization	9 bits
Voiced/unvoiced	1 bit
Total	23 bits per segment

If 11 coefficients are always transmitted in the case of unvoiced sounds, those on average carrying the highest energy, there would be no need for addressing information and only the quantization bits need to be transmitted together with the voiced/unvoiced decision information which results in the following:

Coefficient quantization:

For each coefficient	2 bits
∴ For total quantization	22 bits
Also, For Voiced/unvoiced decision	1 bit
Total:	23 bits per segment

which is the same as for the voiced sound case.

As each segment is 8 msec long, the above is equivalent to

$$\frac{23}{8 \times 10^{-3}} = 2875 \text{ bits/sec}$$

If a fixed low sequency region, for example coefficients 5 to 10, is transmitted in the case of voiced sounds instead of the first formant for achieving better quality, then the

system transmission rate is increased as follows:

Only the second and third formant positions are to be transmitted as the first formant is assumed to be included in the fixed region transmitted always.

Bits required for addressing the second and third formant positions (as shown before) = 2 X 5 bits.

Now there are 8 coefficients to be transmitted, as six coefficients are in the fixed region and the second and third formants, in the voiced case; therefore the quantization needs,

8 x 3 = 24 bits per segment

One bit is necessary for voiced /unvoiced decision, so total bits per segment becomes:

Address	10 bits
Quantization	24 bits
Voiced/unvoiced	1 bit
Total	35 bits/segment

In the unvoiced case 17 coefficients are transmitted, coefficients 25 - 41, as they carry on average most of the energy.

The quantization is done by 2 bits using Adaptive strategy mentioned before. Thus the number of bits necessary for each segment is

Quantization	17 x 2 = 34 bits
Voiced/unvoiced decision	1 bit
Total =	35 bits per segment

which is the same as the voiced case. As the length of each

segment is 8 msec. the bit rate

$$\frac{35}{8\text{X}10^{-3}} = 4375 \text{ bits/sec.}$$

The simulation results of the systems discussed above showed intelligible speech recovery with reasonable quality in the first case and improved quality in the second case. These qualities are similar to the 6.8 K bits/sec. digital speech processing systems which are based on dominant coefficient principals in the Walsh-Hadamard domain (27, 64), but here the systems are simpler in structure for hardware implementations.

A block diagram of the proposed system is given in Figure 4.4.

To compare the system, a section of the recovered speech waveform is presented together with the original speech waveform in Figure 4.5, and the transform domain spectrum of the recovered and original speech is given in Figure 4.6.

The results discussed in this chapter have been published in (89).

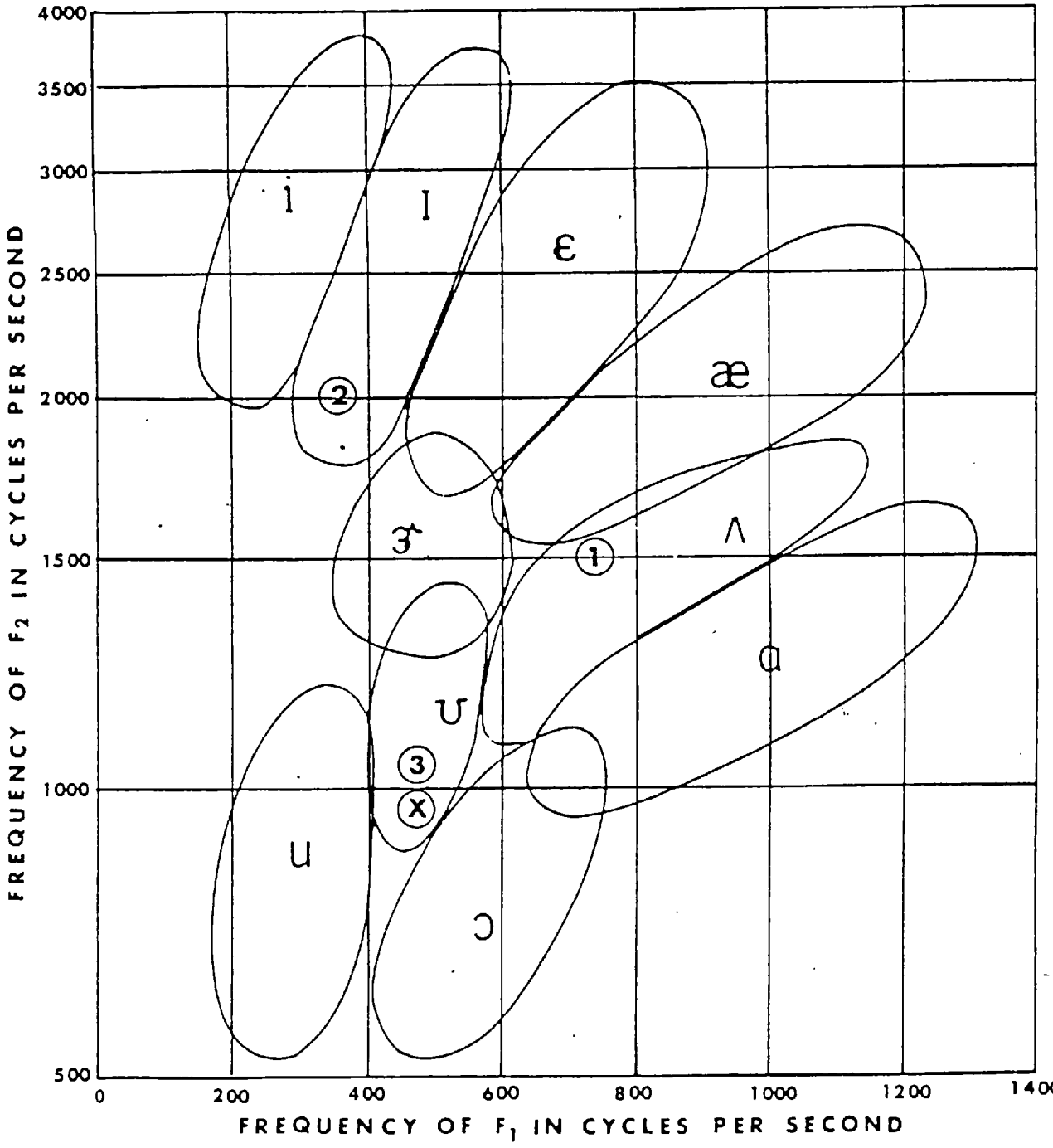


Fig. 4.1: F_1-F_2 space for different Vowels

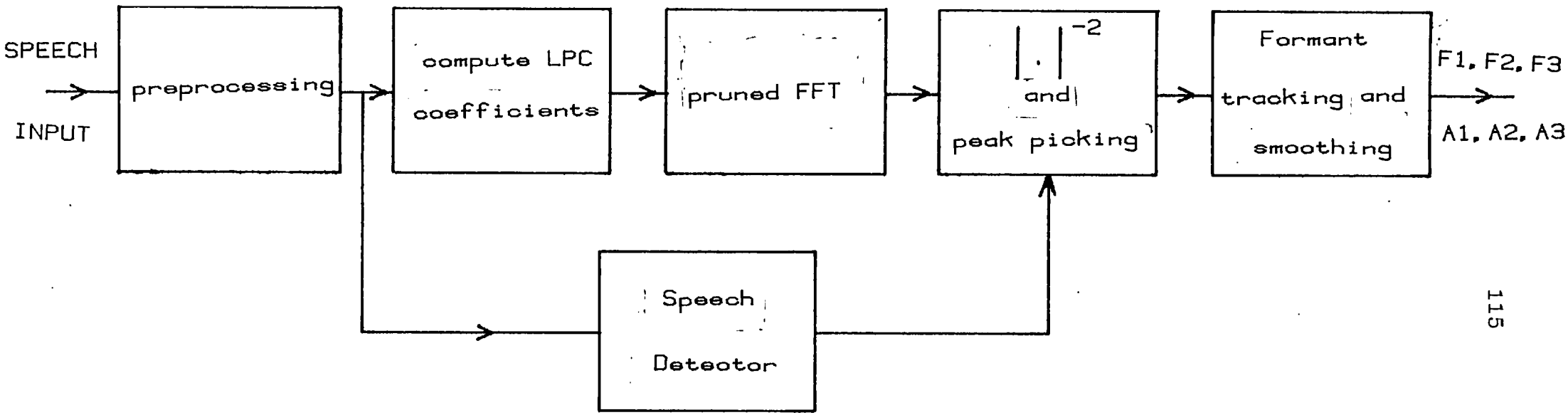


Fig. 4.2: Block diagram of a LPC dependent method for determining Formant Frequencies and Amplitudes

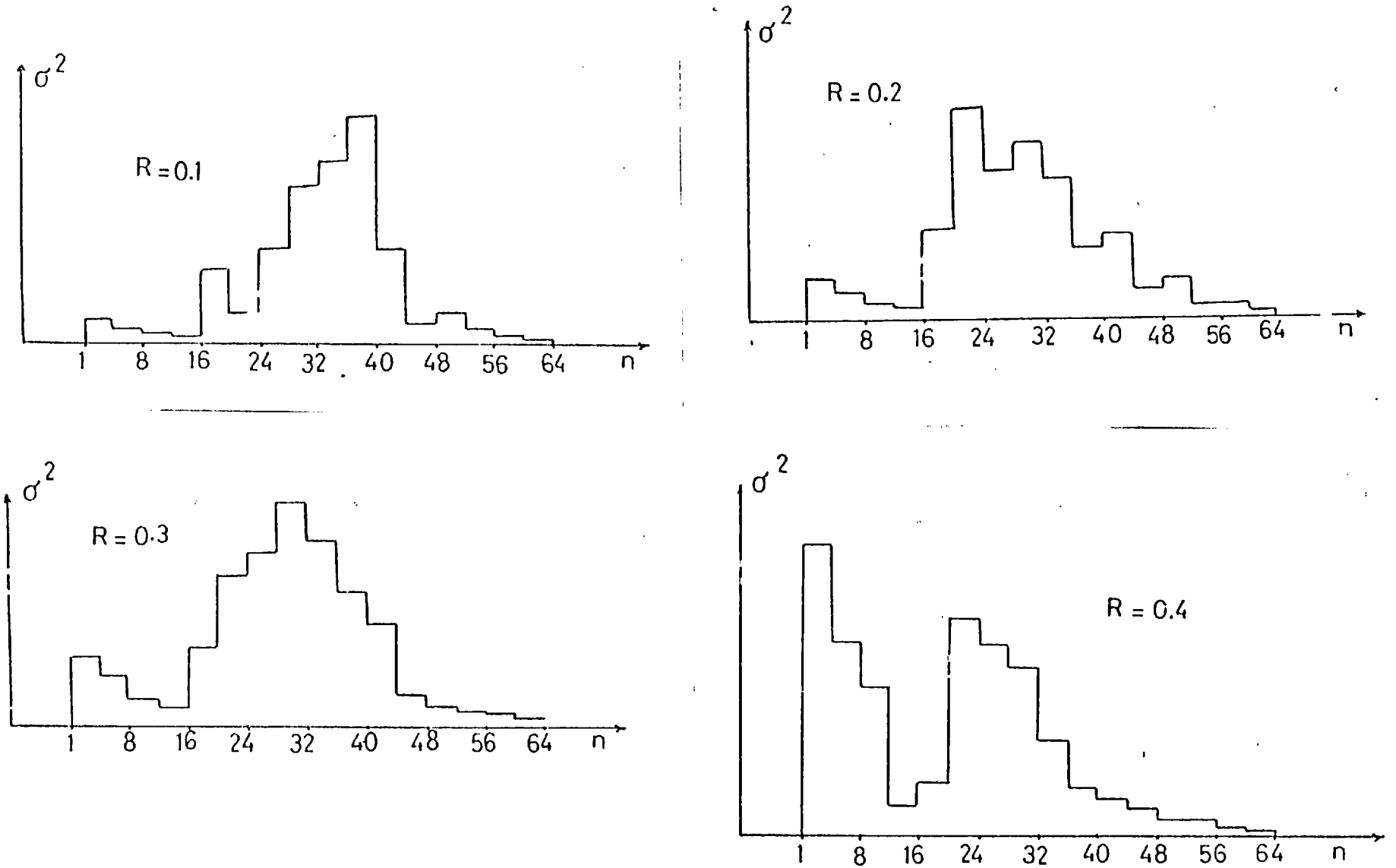


Fig. 4.3: Energy distribution of Walsh coefficients with different threshold levels

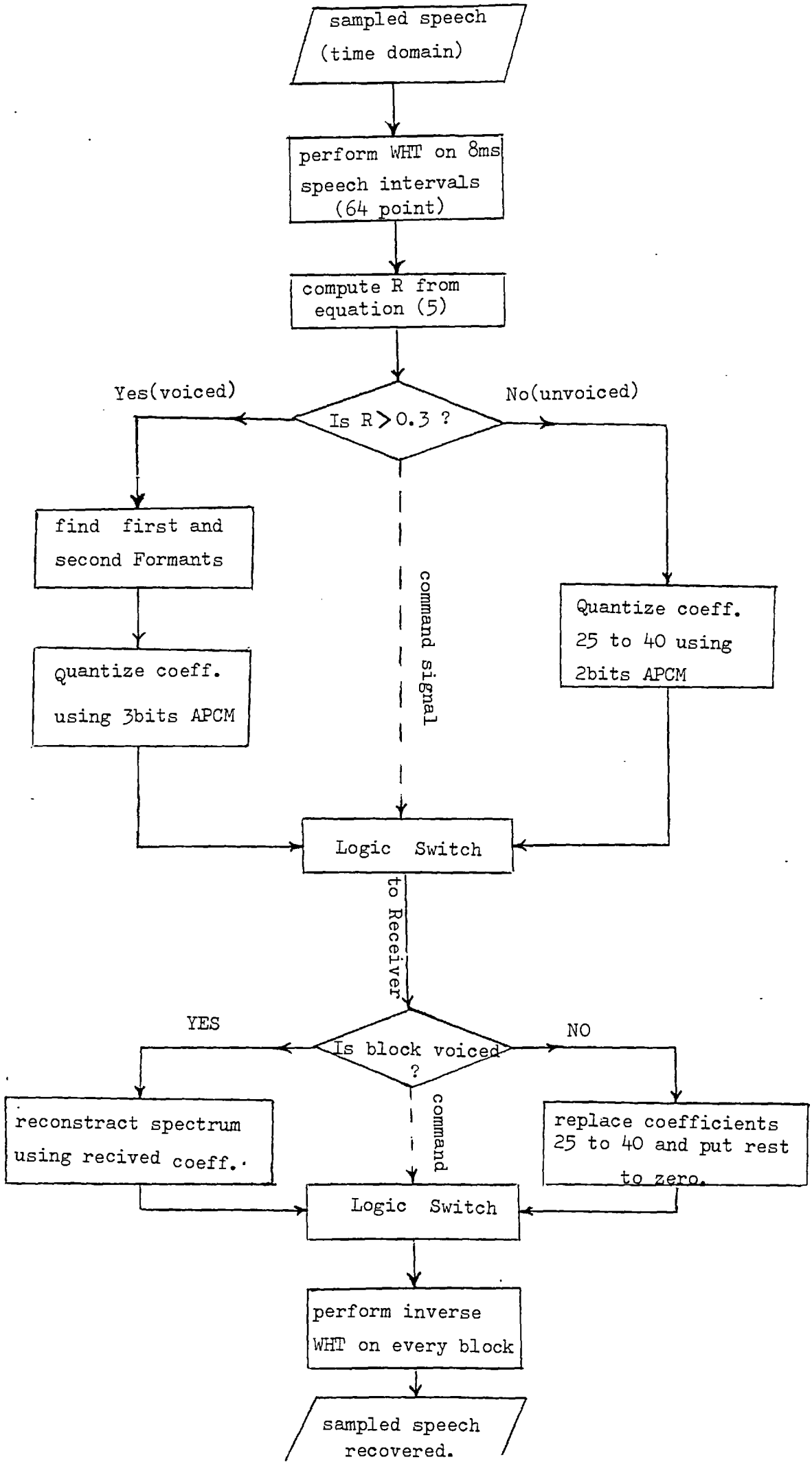


Fig. 4.4: Block diagram of the Formant Sequency System

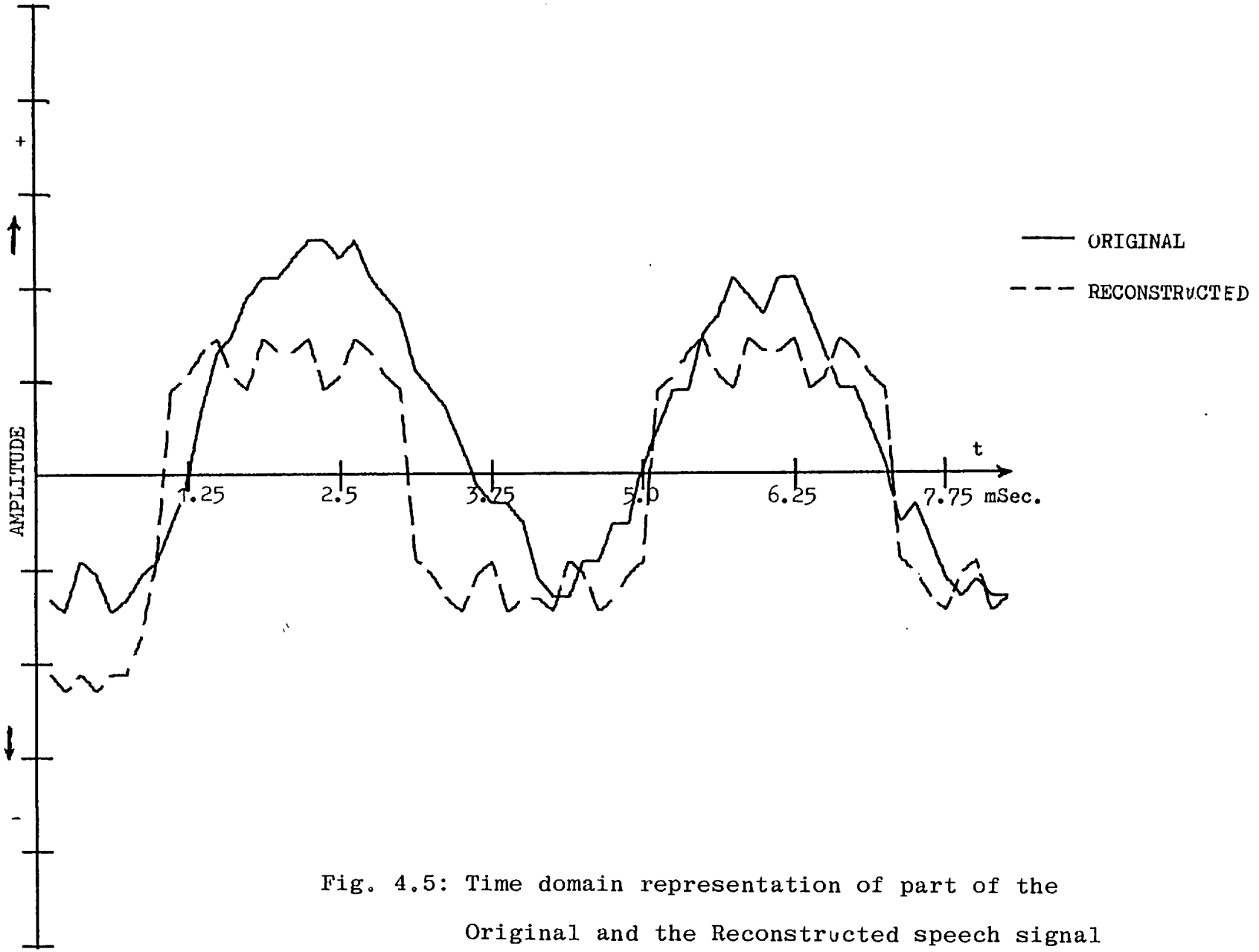


Fig. 4.5: Time domain representation of part of the Original and the Reconstructed speech signal using the Formant Sequency approach

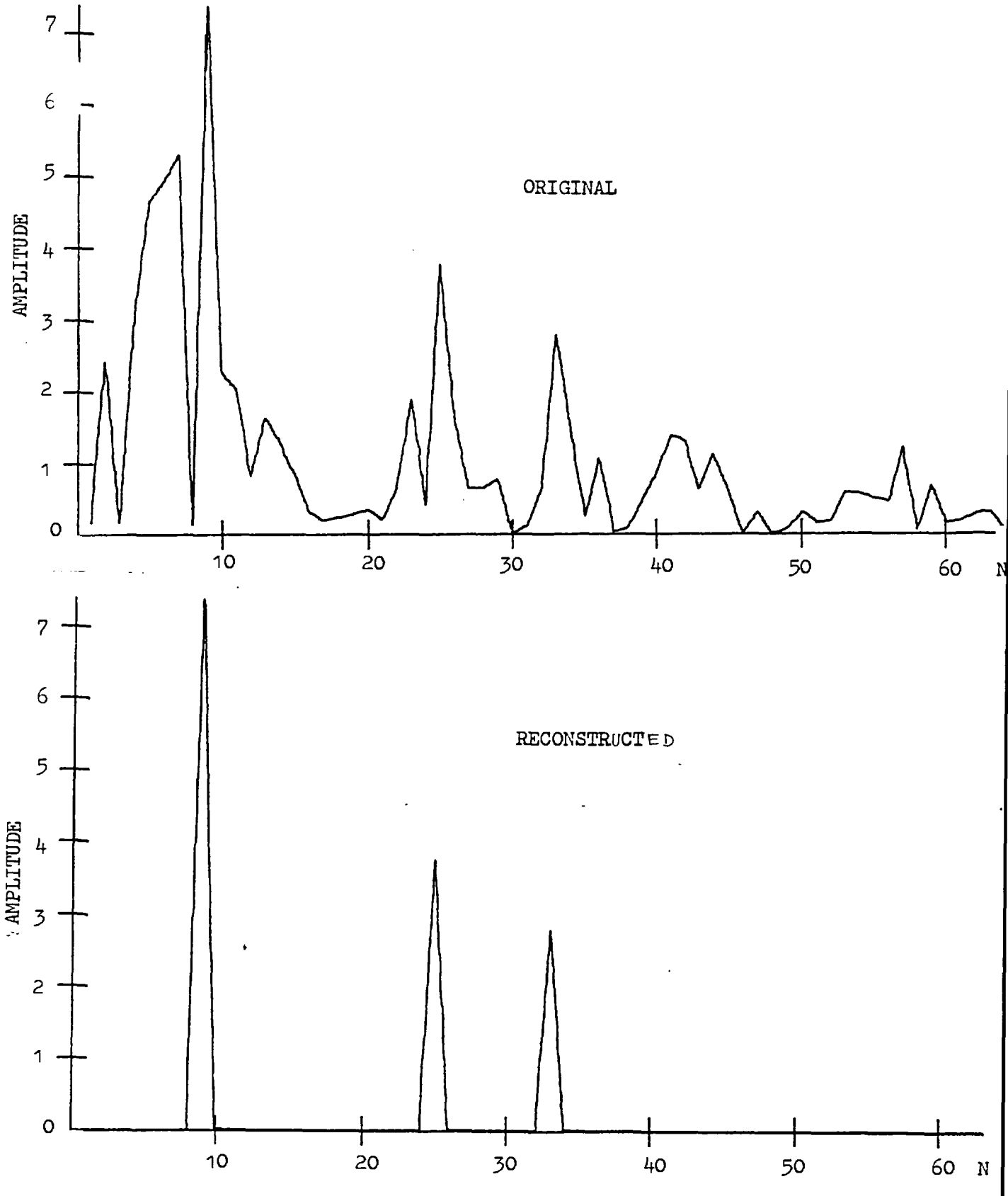


Fig. 4.6: The Walsh domain representation of part of the Original and the Reconstructed speech signal using the Formant Sequency approach

CHAPTER FIVE

LINEAR PREDICTION OF FORMANTS FOR SPEECH PROCESSING

In the previous Chapter the information bearing parameters of the speech signals were highlighted and the transform domain analysis was considered more efficient than the time domain in the fact that inherent redundancies in speech can be removed there more easily. The vocal tract resonances, the Formants, are considered as perceptually the most important parameters of the voiced sounds and are used in the design of a simple and efficient system (14), (48).

In this Chapter the application of the well known linear prediction method in the transform domain is discussed with respect to the formants and it emerges as a useful approach for the removal of speech redundancies in an efficient manner. A system based on the principles of linear formant prediction is proposed and examined where intelligible digital speech transmission is achieved at rates of around 2 Kbits/sec.

5.1 CORRELATIONS BETWEEN SPEECH SAMPLES IN THE TRANSFORM DOMAIN.

It is well known that speech samples are highly correlated in the time domain with their immediately preceeding neighbours due to the fact that the vocal tract configuration changes slowly during speech production (14). For achieving an efficient speech processing

system these redundancies must be removed in the time or any other convenient domain. The correlation between the time domain samples for speech signals still remains if the speech signals are transformed into domains other than time. But such correlations of course may not be with the immediate neighbours but they are dependent upon the orthogonal domain employed for transformation. (17)(90).

In order to verify the existence of correlations between different samples a so-called "co variance matrix" is formed (57). The elements of such a matrix are obtained from the relation

$$C_{IJ} = A(I) \cdot A(I-J) \qquad (5.1)$$

where C_{IJ} are the covariance matrix elements and A is the sample value in a transform domain at a specific location.

The study of the covariance matrices for speech samples in different orthogonal domains such as Walsh-Hadamard, Fourier and others shows the existence of a high degree of correlation between different regions of the speech spectrum for that particular domain. An average symmetrical covariance matrix of 64 by 64 elements is formed from transformed speech samples employing a 64 point Walsh Hadamard transformation on the time domain speech samples over a few balanced sentences taken from the Harvard list (81). A high degree of correlation is observed between certain regions of the covariance matrix which indicates the existence of correlations between speech samples in the Walsh-Hadamard domain. These regions can be verified from the appropriate section of the average covariance matrix as shown in Figure 5.1 where the high amplitude areas indicate the high degree of correlation

between samples of the matrix. This correlation can be explored and used to design efficient speech processing systems as discussed in more detail in the following sections of this Chapter.

5.2. FORMANT CORRELATION IN THE TRANSFORM DOMAIN

One way of indicating the presence of correlations between different parts of a vector is to use the covariance matrix approach as discussed earlier in this Chapter. Another method, independent of the one mentioned above, is to explore such correlation relations through statistical investigation over a high number of vectors of similar characteristics. In the case of speech signals, in order to investigate the possibility of correlations between the formants a statistical approach was employed. To achieve this the time domain speech signal is transformed into the desired transform domain, in this case the Walsh Hadamard Domain. Then the voiced/unvoiced decision is made on small segments of the transformed samples based on different energy distributions of each sound.

The segment length was taken to be 64 samples of the transformed signal which is equivalent to 8 msec duration of the time domain speech for 4 KHz low passed and 8 KHz sampled speech waveform. The formants only appear in the voiced segments of the speech signal hence only these sections of the speech are considered for the investigation. The next step is to note the positions of the first, second and third formants using a peak piking method as discussed in Section 4.2.1. These three values when chosen from the 64 available positions are noted in a symmetrical 64 by 64 matrix.

The investigation is carried out over a few balanced sentences taken from the Harvard list (81) including both male and female speakers.

From the so obtained matrix an interesting result is observed that is the positions of the first, second and third formants are correlated. Thus for a particular first formant position, the possible positions of the second and third formants are quite limited and can be predicted with some degree of accuracy. The useful part of the symmetrical 64 x 64 matrix, that is the part that includes first, second and third formants, is plotted as shown in Figure 5.2. It can be seen that the positions of the first three formants are shown clearly on the figure. In this plot the high amplitude areas indicate the highly probable formant location. If the position of the first formant is known then a vector can be extracted from the above formed matrix which indicates , on average, the possible positions of the second and third formants. Then for all possible first formant positions a series of such vectors exist, which is useful in predicting the second and third formant positions for a known first formant. A plot of the so-called conditional probability of the second and third formant positions given the first formant position is given in Figure 5.3.

This is the so-called conditional matrix which is later used in designing a system for efficient low bit-rate digital speech transmission based on predicting the higher formant positions from the first formant. An interesting confirmation of the existence of correlations between the

formants is the existence of high degree of similarity between the figures 5.1 and 5.2 which have been obtained from two different approaches.

5.3. LINEAR PREDICTION TECHNIQUES IN THE TRANSFORM DOMAIN

The linear prediction technique is a correlation type of analysis which can be approached either from the time or indeed from the transform domain.

Uusally the aim in linear prediction is to minimize the error between a predicted form of the signal and its original version. The normal way to ensure this is through the least squared error criterion. Using this criterion a set of linear equations is obtainable from which the prediction coefficients can be derived. The time domain version of this technique for speech signals has proved to be highly successful and efficient sytems (53, 54, 55). Recently linear prediction techniques have been applied to domains other than time (57).

In the time domain, each speech sample is correlated to its immediately preceeding samples whereas in the transform domain different parts of the spectrum are correlated depending upon the type of transformation employed. All pole linear prediction (LP) modelling is designed for time domain speech signals (53,54) with their specific correlation characteristics. However, by using it without modification the linear prediction of speech samples in the transform domain results in an inefficient method.

The time domain LP approach must therefore be modified to suit the correlation characteristics and the

variance distribution associated with the employed transform domain (57). Full details of such modifications is given in Reference (57) with a brief discussion on the method given below.

In this scheme a baseband region is defined and the rest of the spectrum is predicted linearly from it thus we have

$$C_n = a_n C_j \tag{5.1}$$

where C_n is the predicted coefficient from a baseband coefficient C_j using the prediction coefficient a_n which can be transmitted to the decoder instead of C_n. For appreciable data compression achievement, the number of prediction coefficients should be significantly lower than the number of the so called sub-band coefficients C_n. The aim is to obtain a prediction coefficient, a_n, common for a number of C_n's in such a way that the reproduction of C_n results in minimum error, that is to say equ.(5.1) becomes

$$C_n' = a C_j \tag{5.2}$$

with

$$\text{Error} = \sum_n (C_n - C_n')^2 \tag{5.3}$$

Due to the non-uniform variance distribution of the transform domain speech samples, a single prediction coefficient would not result in an optimum approach which means more than one prediction coefficient must be used depending on the position of C_n. An efficient approach for this purpose is achieved when the prediction coefficients a_n are considered differentonly by a constant factor W (n)

which depends on the statistical behaviour of their corresponding coefficients C_n

$$a_n = W(n)\ \bar{a} \tag{5.4}$$

where W(n) are referred to as the weighting coefficients. Then on combining equations 5.2 and 5.4 we have

$$C_n' = W(n)\ \bar{a}\ C_j \tag{5.5}$$

The constant W(n) is independent of the input speech signal but rather relates to the long term statistics of the speech samples in the employed transform domain and therefore it is necessary to calculate it only once for each domain which can then be stored in the coder and decoder. For a more generalized case where the non-uniform linear prediction is of order P, equation 5.5 becomes

$$C_n' = \sum_{K=1}^{p} W_K(n)\ \bar{a}_K\ C_{jK} \tag{5.6}$$

The squared error over a length n for the first order NULP (Non-Uniform Linear Prediction) technique is obtained by subtracting the original from the predicted coefficients as given below

$$E = \sum_n (C(n)-C'(n))^2 = \sum_n (C(n)+W(n).\bar{a}.C_j)^2 \tag{5.7}$$

from which the optimum values of W(n),, $\bar{a}$ and j can be determined by minimizing the squared error E. These values can be obtained by setting the derivative of the equation 5.7 with respect to each variable to zero; the speech signal for each case being assumed to be stationary during each segment of measurement.

The optimum values for the weighting coefficients can be obtained from;

$$\frac{\partial E}{\partial W(n)} = \sum_n 2(W(n).C_j^2.\bar{a}^2 + C(n).C_j.\bar{a}) = 0 \qquad (5.8)$$

Thus the solution of this equation gives the optimum values for W(N) as given below

$$W(n) = \frac{1}{\bar{a}} \cdot \frac{\sum_n C(n).C_j}{\sum_n C_j^2} \qquad (5.9)$$

where $\bar{a}$ is constant due to the assumed stationarity of speech signals during the measurement period.

The summation is over long term speech signals for different speakers so that the weighting vector can be used usefully for different circumstances.

Similarly, the prediction coefficients $\bar{a}$ can be calculated by setting the derivative of equation 5.7 with respect to $\bar{a}$ to zero.

$$\frac{\partial E}{\partial \bar{a}} = \sum_n \left(W^2(n).C_j^2.\bar{a} + W(n).C(n).C_j \right) = 0 \qquad (5.10)$$

where $\bar{a}$ is obtained as

$$\bar{a} = \frac{-\sum_n W(n).C(n).C_j}{\sum_n W^2(n).C_j^2} \qquad (5.11)$$

where the summation is over the chosen block size. Then

by substituting equation (5.11) into the equation (5.7) the optimum value of the total squared-error can be obtained as,

$$E_{op} = \sum_n C^2(n) - \frac{\left(\sum_n W(n).C(n).C_j\right)^2}{\sum_n W(n).C_j^2} \qquad (5.12)$$

which can be used in calculating the error for different circumstances.

In the following sections the use of linear prediction techniques in the transform domain to achieve an efficient digital speech processing system is discussed.

5.4. LINEAR PREDICTION OF FORMANTS

The correlation between the formants in the Walsh-Hadamard domain has been examined through two independent approaches; firstly the covariance approach and secondly the statistical investigation method both of which showed that the positions of the first three formants are highly correlated as can be seen from Figures 5.1, 5.2 and 5.3.

More specifically for a given first formant position the second and third formant positions can be predicted with a certain accuracy.

Similar correlation relations also exist in other orthogonal domains such as Fourier, Discrete Cosine and others but due to reasons of simplicity and efficiency the Walsh-Hadamard domain is used as the analyzing transform domain in the work presented here,

The use of the linear prediction technique in the transform domain has been discussed in the previous section. In the Non-Uniform Linear Prediction (NULP) technique applied to the transform domain reported in reference (57), a baseband region has been defined from which the rest of the spectrum can be predicted by using weighting and prediction coefficients. The 8 msec. analysis length is used in the NULP approach where the parameters are updated during that time in a similar manner to the case here. An interesting phonemon is that the baseband region used in the NULP technique is on average the possible region of the first formant occurance. This would mean that for each analysis frame only one of the baseband samples, the first formant, is important perceptually and the rest play a minor role in the perception of the speech and have relatively low amplitude. Similarly the rest of the spectrum would not participate equally in the speech perception but only those parts which carry higher energy are significant. Indeed these high energy coefficients are the second and third formants.

Therefore instead of sending all the baseband region for each frame, as is the case in the NULP technique, the first formant is chosen from the appropriate region which happens to be similar to the baseband and sent to the receiver together with weighting coefficients (the factors which are used to emphasise the perceptually more important coefficients in the spectrum).

At the receiver the position of the second and third formants are linearly predicted to some accuracy from the knowledge available to the receiver on a similar

manner to the NULP case. A system based on this principle is discussed in the next section where by using weighting and prediction coefficients low bit rate digital speech transmission is achieved.

5.5. COMPUTER SIMULATION RESULTS OF THE FORMANT PREDICTION SYSTEM.

A system for low bit rate digital transmission of speech signals has been designed based on the Linear Prediction of formants in the transform domain. The system has been simulated on a computer and results tabulated. The description of the simulated system is given as follows:

A lowpass filter is used to limit the analogue speech to 3.04 KHz. Then a sampling circuit is employed to sample the speech signals at 8 KHz and a digital version of the speech signal was obtained using a 10 bit analogue to digital convertor. Following digitization of the speech, transformation of the signals from time to other domains is carried out through appropriate fast algorithm s. In the work presented here the Walsh-Hadamard domain has been employed for simplicity but in general any other orthogonal transform can be applied.

A fixed length transformation is also used here to keep the system as simple as possible. On the other hand a short enough analysis length is used to be able to follow the change in the speech behaviour. The block transformati : length of 8 msec was chosen for the work presented here. This is equivalent to a 64 point transform in this case where the speech signal is low passed to 4KHz

and sampled at 8KHz. So the transformation is carried out every 64 samples through a fast Walsh-Hadamard algorithm as mentioned in Section 3.2.4. The next step after performing the transformation is to decide on the nature of the block, that is to say to decide whether it represents voiced or unvoiced speech, and separate the different types.

To decide on the voiced or unvoiced sounds, the approach based on the energy distribution of these different sounds is adopted as discussed in Section 4.4.4. The same threshold level of 0.3 was chosen as appropriate. Also in this case a value below 0.3 is taken as unvoiced and above as a voiced sound. Each of the sound type is treated separately for transmission over the channel. In the case of the voiced sounds the first formant is chosen from the appropriate region in the Walsh-Hadamard domain,which for the 64 point transform considered, this region is represented by coefficients 3, to 10. This is similar to the baseband region chosen in the method used for speech processing in reference (57) employing the linear prediction technique in the transform domain. The peak-picking method is used to choose the first formant as discussed in Section 4.2.1. The baseband region has on average the largest variance and perceptually is the most important region, but at any given instant by considering a typical transform block the first formant in that region has the most energy and therefore has the most perceptual importance as well. Therefore it is reasonable to use the first formant for each block instead

of the whole baseband.

The prediction coefficient is calculated from equation 5.11. In order to achieve better results and obtain a more adaptive system, it would be more appropriate to use the different prediction coefficients for the prediction of the formants depending on whether the first formant is among the first four coefficients (i.e. coefficients 3 to 6) or the last four coefficients (i.e. 7 to 10). The chosen baseband coefficients are such as to maximize the values of the weighting vector $W(n)$.

Such values are given in Table 5.1. The prediction table for speech in the Walsh-Hadamard domain is given in Table 5.2. The high probability positions for the second and third formants for a given first formant position are given in this table. The results are obtained from the statistical investigation mentioned in Section 5.2 over a few balanced sentences.(81)

The first formants are quantized by an adaptive quantization technique with a one-word memory as given in Section 4.4.7 and the proposed system 3 bits per sample with multiplication factors of M_1=0.9, M_2=1, M_3=1 and M_4=1.75 proved to be good enough. The prediction coefficients are quantized to 4 bits by non-linear Pulse Code Modulation.

The first formant and the prediction coefficient are then transmitted to the receiver for the voiced sound segment of the speech.

In the case of unvoiced sounds the mid sequency region proved to be the most important region perceptually as it also carries the highest energy of the spectrum. The relevant graph for the behaviour of the unvoiced sound in the Walsh-Hadamard domain is shown in Figures 3.7 and 4.3. It is evident in this graph that choice made for the decision is just one.. Therefore in the case of unvoiced sounds a few high energy coefficients from the mid sequency region of the unvoiced spectrum had been transmitted. These coefficients are qunatized by a 2 bit ADAPTIVE PCM with one word memory with multiplication factors of M_1=0.85 and M_2=1.9.

At the receiver end the segments are separated into voiced or unvoiced sections. In the case of the voiced sounds the second and third formants are recovered to some accuracy using:Tables 5.1.and 5.2, the prediction coefficients together with the first formant and the use of equation 5.5..

The rest of the spectrum in this case can be set to zero, but for better performance it can be filled by using a normalized average covariance vector related to the position of the detected first formant. Such a vector can be obtained statistically by considering all the Covariance vectors for each specific first formant position over a few balanced sentences. Then by averaging over these sentences to obtain one vector for each position of the first formant and normalizing the vector with respect to that first formant, only a few vectors would be available for the possible first formant positions. Then in each case the appropriate vector can be used to fill-in

the unpredicted coefficients by multiplying the first formant amplitude by the relevant vector elements.

In the case of the unvoiced sounds the high energy coefficients are replaced in their appropriate positions. The rest of the spectrum in this case can be filled in by using an average normalized vector obtained over different unvoiced sounds. As the unvoiced sound behaviour does not differ much in a transform domain then an average vector can be obtained by considering a series of unvoiced spectra. Then the so obtained average vector can be normalized with respect to the highest coefficient. In the Walsh-Hadamard domain the five highest energy carrying coefficients are 32 to 36 on average, with the highest one on average being the 34th coefficient where the average vector for unvoiced sounds is normalized with respect to this coefficient.

So to recover the unvoiced spectrum at the receiver the normalized average vector is multiplied by 34th coefficient and other high energy coefficients which have been transmitted are inserted in their appropriate positions in the vector.

Finally the voiced and unvoiced segments are placed in their correct order as in the transmitter and the inverse Walsh-Hadamard transformation is performed using a fast algorithm. Thus the time domain speech samples are recovered and the analogue speech signal can be obtained when these samples are passed through a Digital to Analogue Convertor. The bit rates necessary for this transmission can be obtained as follows:

For the voiced sound, as the first formant is chosen from 8 possible positions i.e. coefficients 3 to 10, 3 bits are necessary to address its position. Then, as it was mentioned before, 3 bits are used for the quantization of the first formant. The prediction coefficients require 4-bit non-linear PCM and one bit is needed to indicate whether the segment is voiced or unvoiced and hence we have:

For Voiced Sounds:	Address	3 bits
	Quantization:	
	first formant	3 bits
	of prediction	
	coefficient	4 bits
	Voiced/Unvoiced	
	Decision	1 bit
	Total	11 bits per segment.

For the unvoiced sounds 5 high energy coefficients (32 to 36) are transmitted then there is no need for addressing and as 2 bits APCM is used for quantization of each sample and one bit is necessary for the voiced/unvoiced decision so the total bits per segment in this case is :

For unvoiced sounds:	Quantization of five high energy coefficients	10 bits
	Voiced/unvoiced decision	1 bit
	Total bits per segment:	11 bits

Since each segment is of 8 msec duration the bit rate becomes $\frac{11}{8X10^{-3}}$ = 1375. bits/sec. which is quite low compared to speech processing systems with similar complexity (57) (64).

The simulation results of the system discussed above produced intelligible speech recovery with reasonable quality. The quality can be improved if the bit rate is increased by detecting two maxima instead of one in the voiced sound case and more high energy coefficients in the case of unvoiced sounds. The quality in this case is similar to the system developed in the previous chapter, with bit rates around 4 Kbits/sec.

A block diagram of the system discussed here is given in Figure 5.4.

For the system visual evaluation, the original and processed time domain waveforms of the speech are presented in Figure 5.5. and their Walsh-Hadamard domain representation of the original and processed speech is shown in Figure 5.6.

The results of this chapter have been published as given in References (90) and (91).

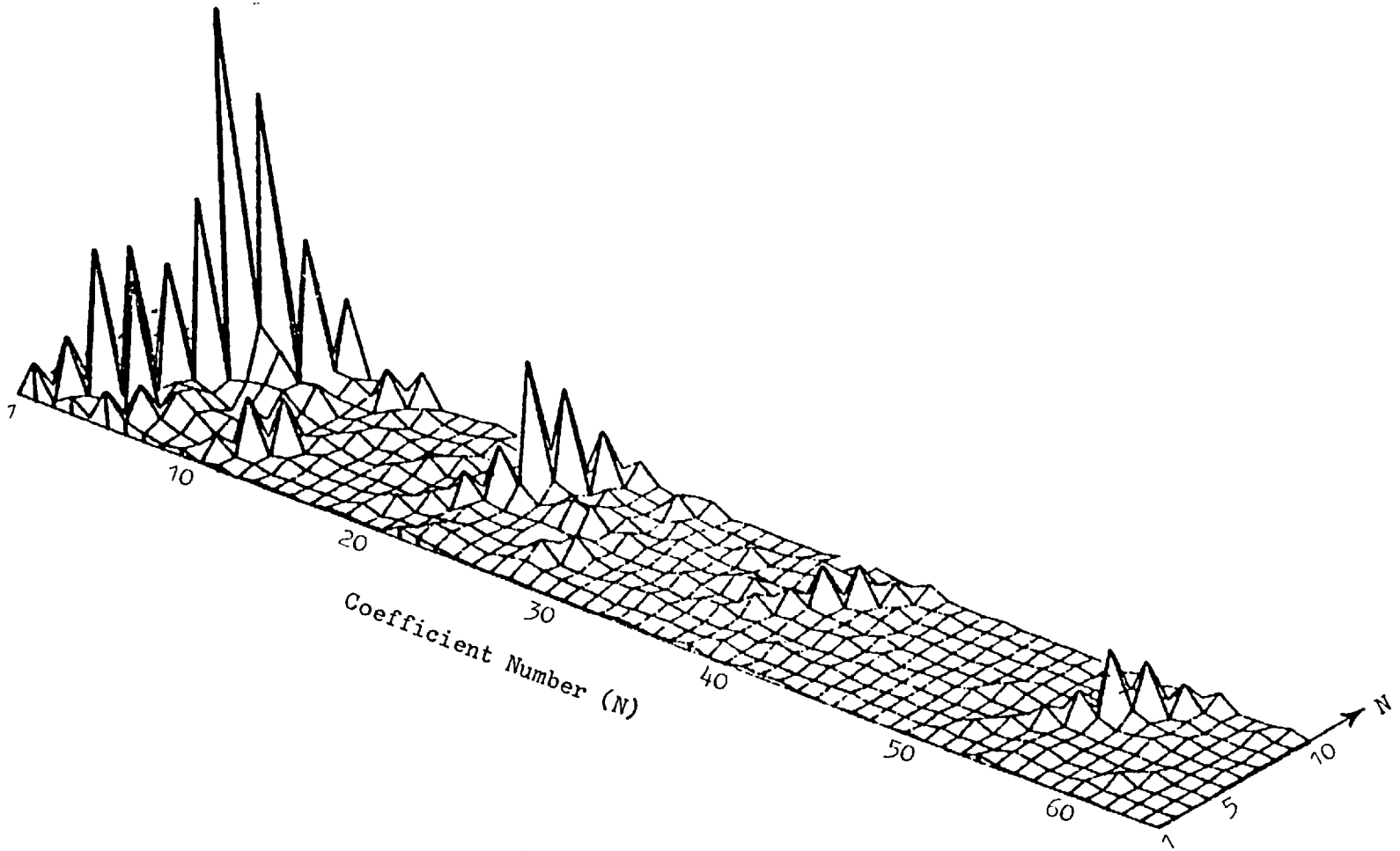


Fig. 5.1: Representation of part of the Average Covariance Matrix obtained from the Walsh-Hadamard domain

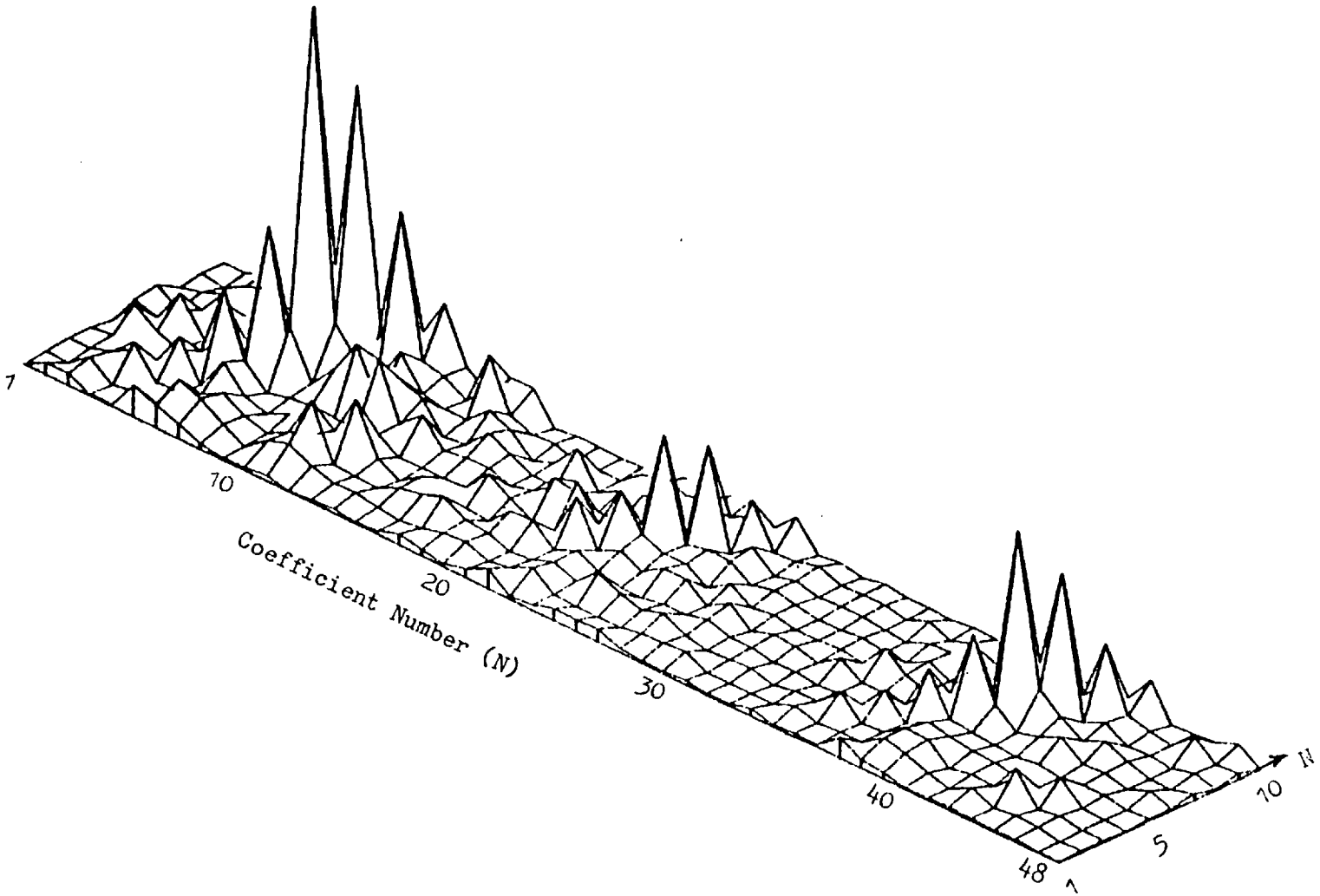


Fig. 5.2: Probability Distribution graph of the first,second and third formant locations

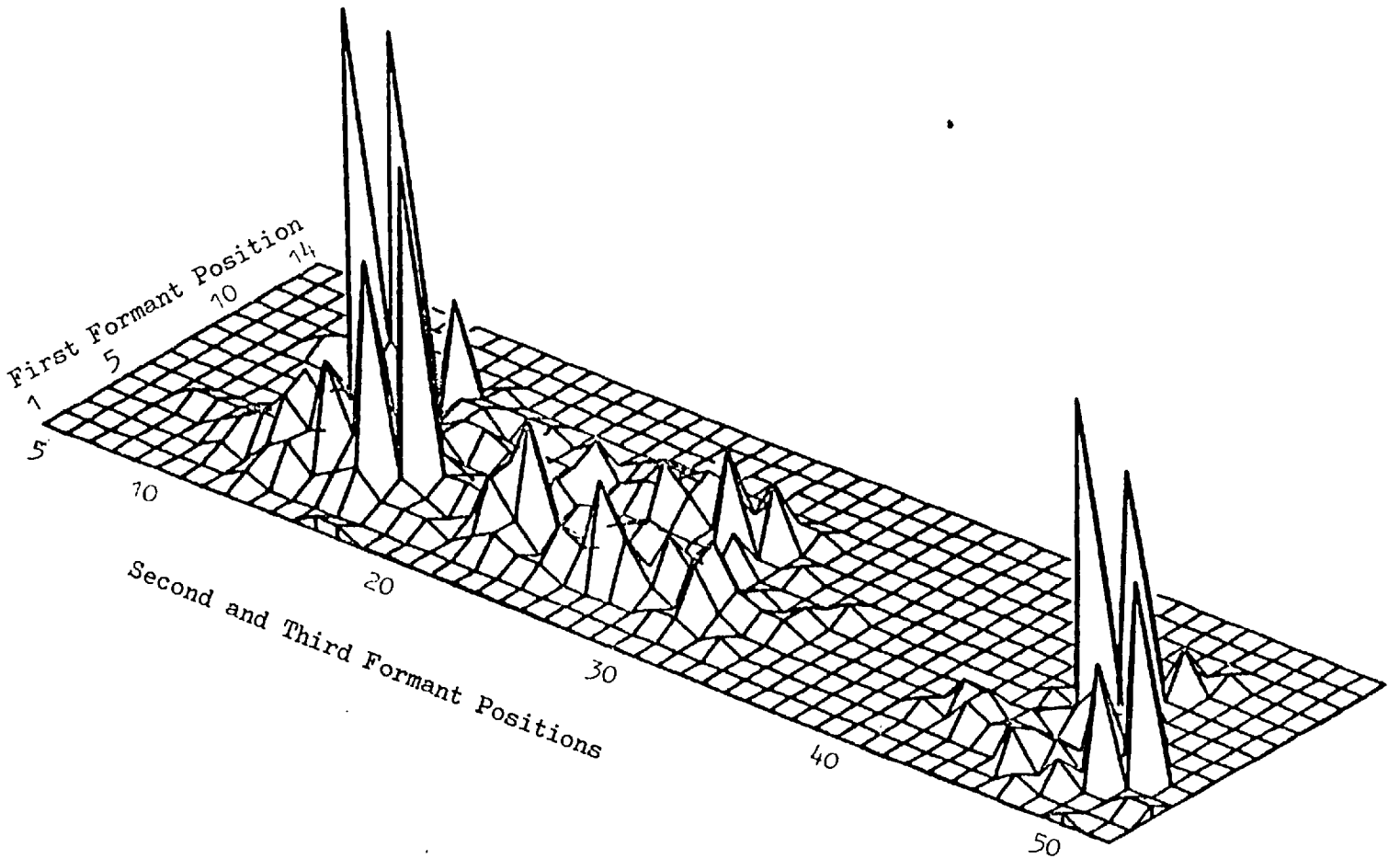


Fig. 5.3: Conditional Probability Distribution graph of the Second and Third Formant Locations for given First Formant locations

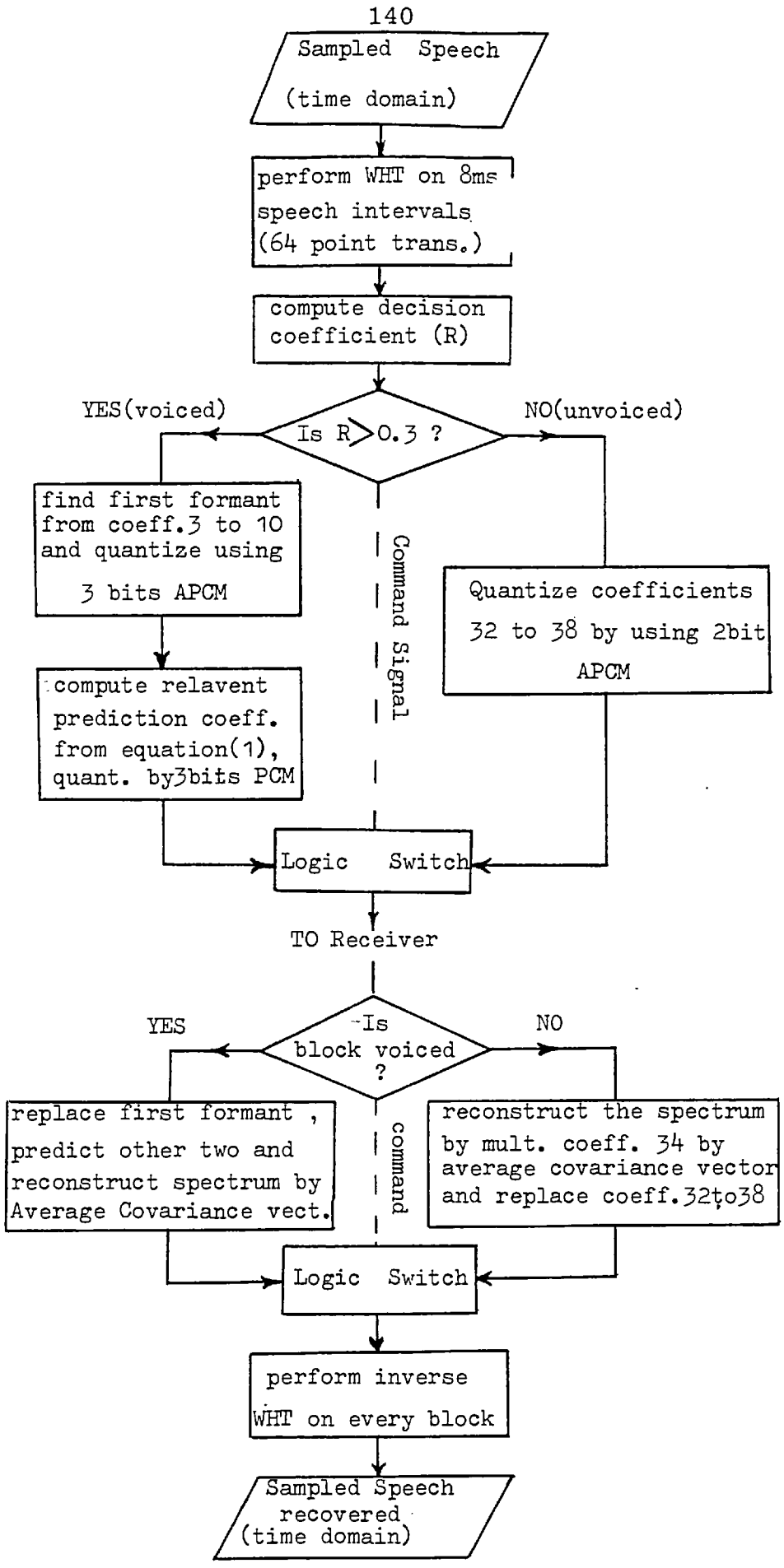


Fig. 5.4: Block diagram of the "Linear Prediction of Formants" system

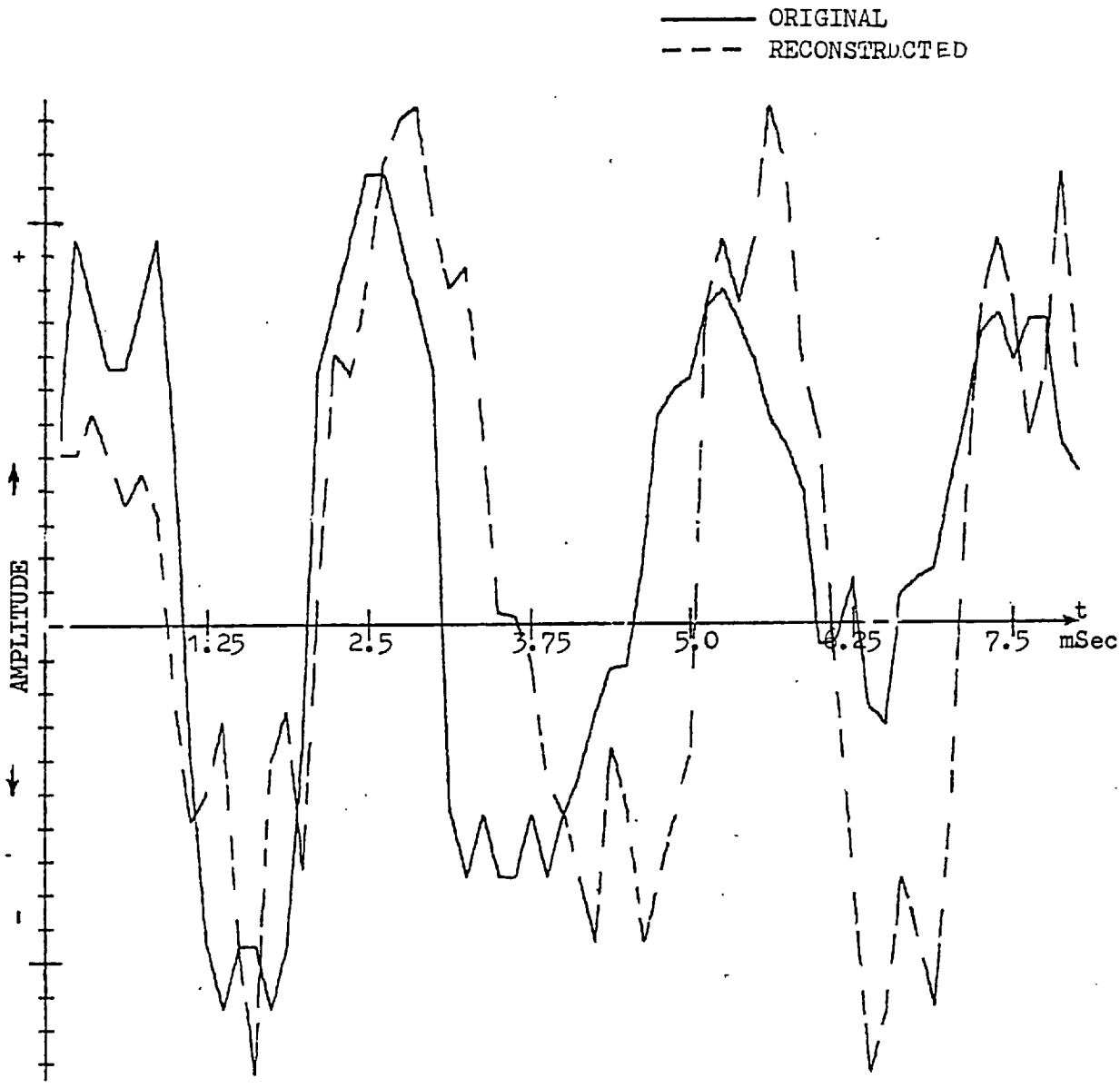


Fig. 5.5: Time domain representation of part of the Original and the Reconstructed speech signal using the "Linear Prediction of Formants" approach

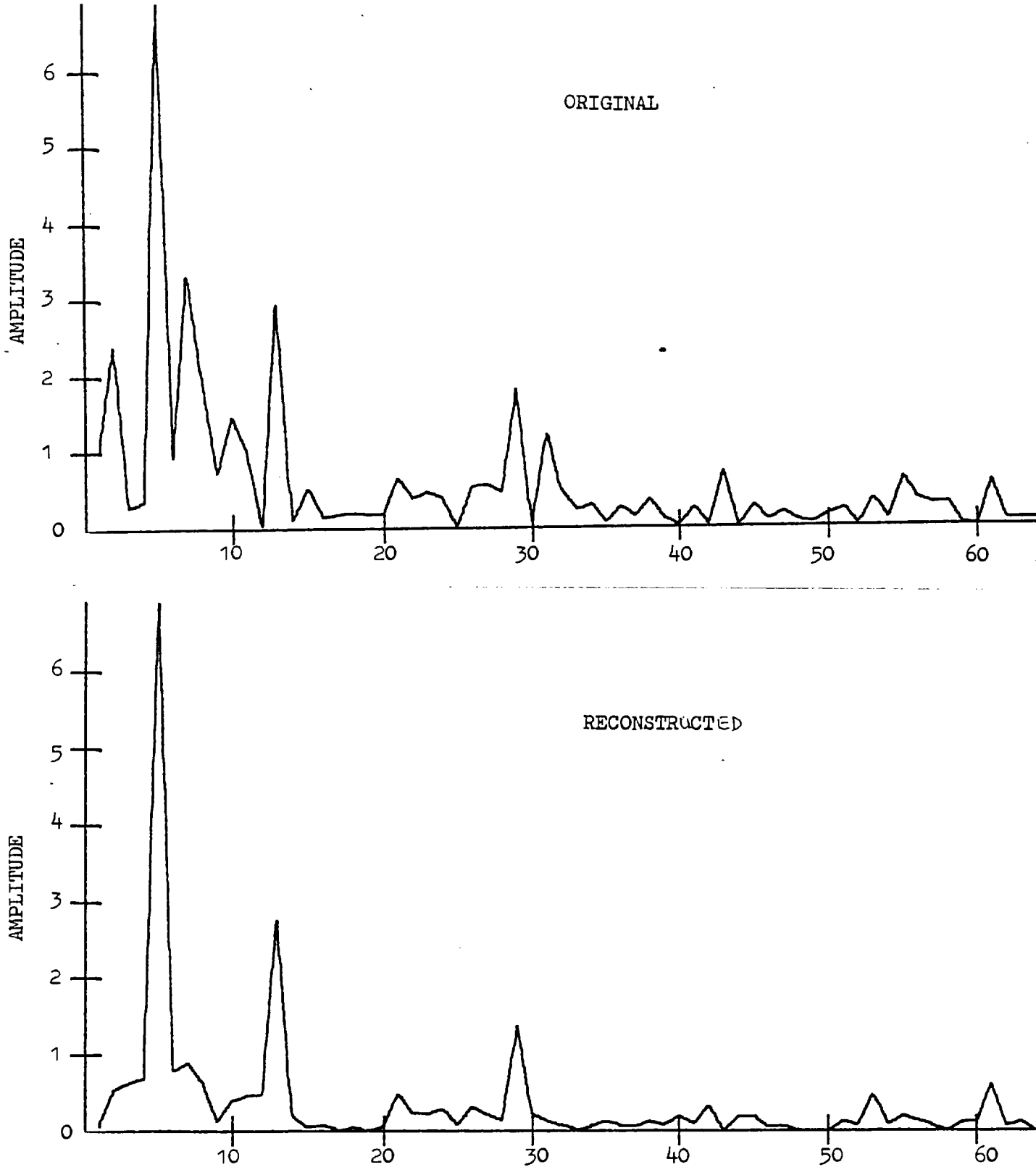


Fig. 5.6: The Walsh domain representation of part of the Original and the Reconstructed speech signal using the "Linear Prediction of Formants" approach

n	W(n)	n	W(n)	n	W(n)	n	W(n)
D.C.	.102	16	.028	32	-.05	48	-.01
1	.341	17	-.09	33	-.12	49	-.04
2	-.23	18	-.13	34	..069	50	.05
3	1	19	-.06	35	.081	51	-.06
4	1	20	-.07	36	-.09	52	-.06
5	1	21	-.28	37	-.16	53	-.16
6	1	22	-.26	38	-.12	54	-.15
7	1	23	-.02	39	-.11	55	-.15
8	1	24	.35	40	-.13	56	.17
9	1	25	.45	41	-.12	57	.16
10	1	26	.29	42	-.14	58	.16
11	-.38	27	-.14	43	.06	59	-.07
12	.32	28	.14	44	-.07	60	.06
13	-.27	29	-.18	45	.03	61	-.06
14	.38	30	.17	46	-.06	62	.07
15	.06	31	.08	47	.06	63	.01

Table 5.1: Weighting Coefficients for the 64 point Walsh-Hadamard Transform

FIRST FORMANT CHOSEN FROM COEFFICIENTS 3 TO 10	HIGH PROBABLE SECOND AND THIRD FORMANT POSITIONS FOR A GIVEN FIRST FORMANT (ON AVERAGE)
3	-11 , -27
4	12 , -20 , 28
5	-21 , 23 , -37 , -53
6	-22 , 26 , -27 , -38 ,40 , -54
7	-23 , 36 , -39 , -55
8	24 , -40 , 56
9	25 , -29 , -41 , 57
10	26 , 58

Table 5.2: Prediction table of Speech segments in the Walsh-Hadamard domain

CHAPTER SIX

CLUSTERING APPROACH FOR SPEECH PROCESSING SYSTEMS

6.1 INTRODUCTION

In Chapter 4 a system was discussed where the formants were used as the main parameters for speech processing, the rest of the spectrum remaining unused.

In Chapter 5 a more efficient use of the speech characteristics has been achieved by combining the well-known linear prediction techniques and the formant data which resulted in low bit rate digital speech processing but still some parts of the spectrum have not been taken into account. At this stage the desirability was made patent for designing a simple system which takes into account all the spectrum realised. The research was then concentrated on the design of relatively simple system which tends to recover the whole of the spectrum segment. After an extensive investigation the idea of comparing the incoming spectra with a set of fixed spectra and chosing the nearest one for transmission over the digital channel was found appealing. In this way the formants as well as other important features of the spectrum are preserved with some degree of accuracy. In the following sections the method through which this idea is put into practice is explained and a simple low bit rate digital speech processing system based on this novel approach is

designed and the results of the computer simulation presented.

6.2. VOWELS AND CONSONANTS

As already mentioned speech basically consists of voiced or unvoiced sounds. The voiced sounds have periodic time domain structures whereas the unvoiced sounds have non-periodic, noise like shapes, in the time domain. In the transform domain the voiced sounds are identified by the presence of peaks in their spectrum, whereas the lack of such peaks indicates the presence of the unvoiced sounds.

Speech sounds can also be classified into vowels and consonants. The vowels result from relatively stable vocal tract configurations whereas the consonants are produced by changing configurations.

It has therefore been convenient to treat these two types of sounds separately. Vowels are all voiced sounds, but consonants are a mixture of voiced and unvoiced sounds. A list of vowels and consonants is given in Table 3.2. In the case of the vowels the formants are almost stable during the whole vowel period which is of the order of 100 m sec. on average. The vowels have been classified and identified from their first two formants due to the stable structure of their formants (75). An example of such classification is given in Figure 6.1 where the $F_1 - F_2$ space specifies different vowels at each particular region. As can be seen from Figure 6.1 each vowel can almost be uniquely identified through its first two formants except for the overlapping areas which need more investigation for exact identification of the desired vowel, such as the direction of movement of the first and second

formants or other similar information. This classification has been utilized in the literature for speech recognition and synthesis (75).

In the case of the voiced consonants due to the changing configuration of the articulation system their formants are not stable and have rapid variations during the transitional periods of such sounds. It is due to this reason that consonants cannot be identified so easily on the basis of their first few formants in a similar way to the vowels.

This has been the basic design obstacle for the speech processing systems based on the classification of voiced sounds according to the positions of their first few formants. The attempt in the work presented here has been to somehow overcome this problem and to treat all the voiced sounds similarly in this respect. The method which is proposed to achieve this aim is novel and proved to be successful in practice as the simulated system resulted in reasonable speech recovery. This is described in Section 6.4.

6.3 FORMANT TRANSITIONS IN CONSONANT SOUNDS

As mentioned previously unlike the vowels, consonants are dynamic sounds which are produced by the changing vocal tract configurations. This difference is apparent in their short time spectra which typically exhibit rapid formant transitions. For convenience the consonants can be divided into the following classes based on their manner of production;

stops, fricatives, nasals, liquids and semi-vowels (14) (67). The formant transitions for the voiced consonants vary from one sound to the other. Their rates of change differ but that has been estimated about 30 msec, on average, for a frequency change of around 300 Hz and has been assumed to be, approximately, a linear change (83), as can be seen from Figure 6.2.

These transitions are of vital importance to speech perception. Due to the presence of the formant transitions in the voiced consonants, the use of formants for the identification of such sounds similar to the vowels has proved to be impossible. Attempts have been made to overcome this problem and the solution suggested here is one of practical significance. In order to be able to represent the voiced consonants by their formants in a similar manner to the vowels, segments of the signal are considered during which the formants can be assumed to be almost stationary. This is equivalent to approximating the sloped line referring to formant projectory by a staircase. The shorter the segments, the nearer the staircase approximation will be to the actual line. On the other hand the more complex and higher bandwidth systems result due to an increasing in number of the staircase steps. The segment length can be found experimentally by considering the worst case of the consonant formant transitions.

Different segment lengths and relevant staircase approximations to the formant transitions are shown in Figure 6.3. As the average transition length is around 30 mseec,

a fixed segment length of 8 msec seems to be a good compromise between the complexity and accuracy. During the chosen segment of 8 msec. the formants are assumed to be constant as shown in Figure 6.3.

Such a segment length then renders the treatment of vowels and voiced consonants the same, since both types can then be classified with respect to their first two formants.

6.4. VOICED SOUNDS SHORT TIME SPECTRUM

It has been shown in the previous section that the voiced consonants can be treated in a similar manner to the vowels. This is achieved by considering short enough segments of the voiced consonants during which the formants can be assumed to be constant. The segment length of 8 msec. was decided upon to be a good compromize between the accuracy and the complexity of the proposed system. In this way each voiced sound, represented by its short time spectrum, can be identified through its first two formants regardless of whether it is a vowel or voiced consonant.

However to recover a voiced sound it is best to reproduce its short time spectrum as close to the original spectrum as possible. As the first two formants are the essential parameters of the short time spectrum it follows that they should be preserved for the purpose of reconstruction of the sound.

It was mentioned in Section 4.2 that the formants exist within some specified regions in a given transform domain.

If the discrete transform domain such as the

Walsh-Hadamard is considered, then the formants can take a few specified discrete positions. In this respect the first formant in the Walsh-Hadamard domain of a 64 point transform can have about 8 different positions effectively. The second formant has an effective region of some 20 different positions in the Walsh Hadamard domain. This means that the first and second formants can have 160 different positions combined.

In the previous chapter it was shown that given the first formant position in the WH domain the second formant can take only few positions. This resulted from statistical studies over a number of sentences as presented in Figures 5.1, 5.2 and 5.3. This would mean that the number of different combinations of the first and second formant positions would reduce from the figure stated above.

Following this, a novel method is proposed to make use of the limited number of this combination. It uses a relatively small number of short time spectral shapes for all possible combinations of the first and second formant positions to represent the voiced sounds. A library is formed from these spectral shapes.

The library so formed is considered as the centre of the system. In this respect any voiced sound is first segmented into short intervals. Then its short time spectrum shape is compared with the library shapes considering the first two formants as the essential parameters. Following this the nearest library shape to that of the incoming voiced spectrum shape is chosen through an algorithm. Then the information regarding the library shape position is

transmitted. At the receiver an identical library of short time spectral shapes is available. The nearest spectrum shape is recovered from the received information which leads to the recovery of a similar voiced sound as at the transmitter. Similar procedure can also be employed for the unvoiced sounds.

The formation of such a library is given in more detail in the next section.

6.4.1 FORMATION OF THE LIBRARY FOR SPECTRUM SHAPES

Extensive statistical investigations have been carried out to obtain different short time spectra for the voiced sounds of 8 msec segmental length. A number of balanced sentences were used throughout the investigations (81).

The short time spectrum shapes were obtained for the voiced sounds and indexed according to their first and second formant positions. The spectrum shapes are normalized with respect to the first formant as in general this indicates the highest amplitude coefficient in the spectrum. A tree like structure for the library of the short time spectrum shapes has been formed where the first and second formant positions have been used for identification of each shape as shown in Figure 6.4. Although some spectrum shapes have identical first and second formants, they can differ in their overall shape due to formant bandwidths, fine spectral details and third and higher ormant positions. The library so formed can have different number of spectrum shapes depending on the degree of the complexity and sound quality required. For each orthogonal domain a separate

library of spectrum shapes is needed. Following the formation of such a library the short time spectrum of any voiced sound with the equal segmental length can be compared to the elements of this library from which the nearest shape is chosen for transmission to the receiver. Different algorithms can be employed for comparing the spectrum shapes. Two such algorithms are discussed in the next section which produce an acceptable performance from a perceptual point of view.

6.5. COMPARISON OF SPECTRUM SHAPES "THE BEST CHOICE"

There is a need for an efficient algorithm in order to choose the "nearest" library spectrum shape to the short time spectrum of an incoming voiced sound under consideration. Many different algorithms can be used for this purpose but those which resulted in better sound intelligibility and quality are discussed here.

As the spectrum energy distribution is vital to the perception of the desired sound this should be considered as the most important basis of comparison between the incoming spectrum and the library of spectra. On the other hand the sign changes (phase) in the analysis domain has some, if not significant, role in the perception of the sound. It is to improve the speech quality that the signs of the perceptually important coefficients from the original spectrum can be transmitted to the receiver for replacement in the nearest spectrum shape chosen from the library. In this case the bit rate would increase for achieving a better speech quality.

Two comparison algorithms have been considered namely the least Mean Square Error and the Moment approach which produce the perceptually nearest library spectrum shapes to the incoming voiced spectrum.

6.5.1 LEAST MEAN SQUARE ERROR APPROACH

A well-known method which has been used extensively in the field of communication for comparing sets of data is the Least Mean Square Error criterion. The method can be used in this particular case for obtaining the nearest library spectrum shape to any incoming voiced spectrum. This is formulated as;

$$E_m = \sum_{n=1}^{64} \left(a(n) - b_m(n) \right)^2 \qquad (6.1)$$

where E_m is the Mean Square Error between the incoming spectrum coefficients $a(n)$ and the m^{th} library entry spectrum coefficients $b_m(n)$. The library entries "m" are those spectrums with identical first and second formant positions as the incoming spectrum and are considered as the appropriate candidates for the choice of the nearest spectrum shape. Then among the appropriate library shapes the one which results in the least mean square error when compared to the incoming spectrum is chosen to be the nearest spectrum shape. An interesting point to note is that due to the non-uniform energy distribution of the short time spectrum the regions of the spectrum with higher energy which also have more perceptual importance, contribute

more in error calculations than the lower energy regions of the spectrum. This results in the fact that an error in the higher energy region would be more pronounced perceptually compared to the same error in the lower energy regions.

This can be considered as an automatic weighting of the spectrum accroding to the perceptual properties of different regions of the spectrum when the least mean square error method is used for deciding on the nearest spectrum shape to the incoming spectrum.

6.5.2. MOMENT APPROACH

Another method employed for choosing the nearest library spectrum to that of incoming short time spectrum is the Moment approach. Generally moments are used to represent the probability distribution curves (84) (85). and as the short time spectrum can be regarded as an energy distribution curve, therefore the moments can be used to represent it.

The n^{th} moment of a set of values can be given as;

$$M_n = E((X_i - E(X_i))^n) \qquad (6.2)$$

where $E(X_i)$ is the expectation and M_n is the n^{th} moment of the set X. This is generally referred to as the n^{th} moment about the mean of the set X.

It is easily verified that $M_o = 1$ and $M_1 = 0$.

Moments about the mean are of special importantce since they can be used to describe the shape of a probability distribution curve. The second moment about the mean indicates

the spread or dispersion of such a curve. The third moment describes the symmetry or skewness of the curve i.e. if it is zero the curve is symmetrical; if it is positive the curve is positively skewed and a negative third moment indicates the negative skewness of the curve. Higher moments reveal more details about the curve shape.

It can therefore be concluded that any distribution curve can uniquely be represented by a number of moments. In this case the library spectrums can be represented by their moments. The incoming short time spectrum is also represented by its moments. The number of moments to be used for each spectrum depends on the accuracy of the results required. Then the degree of similarity between two spectras can be obtained by comparing their moments. In this case also the library spectrums with identical first and second formant positions as the incoming spectrum are only considered for the comparison purposes. Then for each incoming spectrum the library spectrum with the least moment difference is considered to be the nearest .

This can be given as:

$$D_i = \sum_{n=1}^{k} (M_n - (M_n)_i) \qquad (6.3)$$

where M_n is the n^{th} moment of the incoming spectrum and $(M_n)_i$ is the n^{th} moment of the i-th library spectrum and K is the order of the moments which is decided upon depending on the accuracy required. Thus the library spectrum "S"

with the minimum moment difference is chosen to be the nearest to the incoming spectrum:

$$S = \text{Min } (D_i) \qquad (6.4)$$

The perceptual importance of the moments can be taken into account if a weighting scheme is introduced whereby those moments which contribute more to the perceptual properties of the sound recovered from the spectrum are weighted relative to the other moments. Weighting has also been employed in our case which is to be discussed in Section 6.6.1

6.6 SPEECH PROCESSING SYSTEM BASED ON CLUSTERING APPROACH

A novel approach is suggested in this Chapter through which all types of voiced sounds can be treated similarly as they are recognized through their first two formants; this being achieved provided that short segments of speech are considered for analysis. The formant positions are limited in the discrete transform domain as discussed in Chapter 5. Through an extensive statistical investigation a library of a limited number of typical short time spectra related to different combinations of the first two formants has been formed. In the proposed system the nearest spectrum shape to that of the incoming spectrum is chosen from the library and information regarding its location is transmitted to the receiver where an identical library is available. Then the relevant spectrum shape is chosen for the speech sound

recovery. In general the system can be designed for any orthogonal transform domain. In the investigation carried out the Frequency and Walsh-Hadamard domains have been used as the analysis domains.

A computer simulated system for each transform domain is discussed in the following sections.

6.6.1. WALSH-HADAMARD DOMAIN SYSTEM APPROACH

A system based on the clustering approach is designed in the Walsh-Hadamard domain and simulated on a digital PDP-15 Computer. Similar to the other two systems discussed in previous chapters the analogue speech is low pass filtered at 3.04KHz and then sampled at 8 KHz where digital samples are obtained using a 10 bit A/D convertor.

Then 8 msec segments of speech are considered for analysis in order to treat all voiced sounds similarly as discussed in previous sections. Each short-time segment of speech signal is transformed into the Walsh-Hadamard domain using a fast algorithm. This results in a 64 point transform length throughout the analysis, as a fixed transformation length is employed. Following the transformation of each block, the decision on the type of sound is made. The same principles are used for such decision making as in previous systems (Section 4.4.4). A threshold value of 0.3 was taken as the ratio of the low sequency region energy to the total spectrum energy below which the sound was taken to be unvoiced and above it as voiced. Different strategies are used for each sound.

In the case of voiced sounds, the first and second formants are detected from the short time spectrum using the peak-picking approach as discussed in Section 4.2.1. From the knowledge of the first two formant positions the relevant section of the short time spectrum library is considered as the region which contains the nearest spectrum shape to the incoming spectrum. (Section 6.4.1).

The spectrum shapes with identical first and second formants as the incoming spectrum are then taken from the relevant section of the library and compared with the incoming voiced spectrum using Least Mean Square Error method or the Moment approach. Perceptually the LMSE approach is slightly better than the moment approach for the Walsh-Hadamard domain. The nearest library shape to that of the incoming spectrum is then chosen. The information regarding the position of the nearest spectrum shape is transmitted to the receiver. The library spectra are normalized with respect to their first formant amplitude as it usually contains the highest energy. The first formant amplitude of each voiced sound spectrum is therefore transmitted to the receiver. Since the probability Density Function PDF of the first formant is the same in all the systems discussed here, the first formant is quantized using 3 bit APCM with one word memory (36) as it was employed in other previous systems and discussed in Section 4.4.7.

In the case of unvoiced sounds a similar library approach can be taken. Different short time spectra for typical unvoiced sounds can be obtained through

statistical investigation of such sounds. The library in this case contains much less elements than the voiced sound case. This is due to the fact that unvoiced sound has a noise-like spectrum and does not differ considerably from one unvoiced sound to the next. The Least Mean Square Error method or the Moment approach can be employed to compare the incoming unvoiced spectrum to the library spectrum shapes. The library spectra are normalized with respect to the highest amplitude sample. Such a sample has shown to be the 34th sample of the 64 point spectra ,on average, chosen through statistical investigations carried out over a number of balanced sentences (81).

PCM is employed as the quantization technique in this case for representing the maximum sample. The average covariance vector approach to recover the unvoiced spectrum can also be used in here as in the system described in the previous chapter.

The received information is then used to recover the speech. The voiced sounds are recovered by referring to the relevant spectrum shape from the identical library to that available at the transmitter. The spectrum so obtained is then multiplied by the first formant amplitude to regain its appropriate level. In the unvoiced case the relevant nearest spectrum shape is recovered from the appropriate library. The spectrum is then multiplied by the 34th sample amplitude to denormalize the spectrum. Then the spectra obtained in the Walsh-Hadamard domain are inverse transformed into the time domain in the same order as they were received.

The overall bit rate necessary for the transmission of information depends on the library size. In the simulated system described here a 32 element library was used. This was due to the limitations imposed by the computing facilities available. Increasing the library elements to 64 spectrum shapes would increase the overall bit rate by a small fraction whereas the quality of the recovered speech would improve in higher proportion as better spectrum shapes would be chosen to represent the incoming short time spectrum. In the system described here the bit rate is calculated as follows:

For the voiced sound; the bits needed per frame are;

Voiced Sounds:	Address (for library size of 32)	5 bits
	First formant Amplitude quantization using APCM	3 bits
	Voiced/unvoiced decision	1 bit
	Total bits per frame	9 bits

Since the analysis segment is 8 msec long then the frequency of each frame is $\frac{1}{8X10^{-3}}$ sec^{-1} therefore 9 bits per frame would result in overall bit rate of:

$$\frac{9}{8X10^{-3}} \text{ bits/sec} = 1.125 \text{ Kbits/sec.}$$

Increasing the library spectra from 32 to 64 shapes would add only 1 bit to the addres necessary for the transmission per frame. This results in 1.25 Kbits/sec. transmission rate.

And increasing it even more to 128 elements in the library results in:

$$\frac{11}{8X10^{-3}} = 1.375 \text{ bits/sec.}$$

It can be seen that the increase in library spectrum shape would result in little increase in the bit rate but would improve the speech quality drastically.

In the case of unvoiced sounds two different approaches can be taken. Firstly the average covariance vector approach as discussed in Chapter Five can be used where a variety number of high energy coefficients are transmitted depending on the voiced sound bit rate. This is due to the fact tht a fixed bit rate is aimed at rather than a variable one. If 9 bits per frame are used in the voiced case then four high energy coefficients would be transmitted for the unvoiced sound. If these high energy coefficients which are fixed on average, are quantized by 2 bit APCM, as discussed in Chapter Five, then the bits per frame become;

Quantization of Four high energy coefficients (coefficients 33 to 36)	8 bits
Voiced/unvoiced decision	1 bit
	9 bits per frame

For other bit rates more high energy coefficients are used or an alternative approach is taken similar to the voiced sound case. A library consisting of a number of typical unvoiced spectra can be formed. The library would contain much less of the spectrum shapes in this case as compared with the voiced library.

The number of library elements can vary according to the speech quality required.

The spectra are normalized with respect to the highest valued sample which on average has been shown to be the 34-th coefficient in the 64 point spectrum. This coefficient is quantized using 4 bits PCM or more depending on the speech quality required. For example the bit rate required for a 16 element library in the case of unvoiced sound is:

For Unvoiced:	Address of the nearest spectrum shape from the library	4 bits
	34-th coefficient quantization	4 bits
	voiced/unvoiced decision	1 bit
	Total bits per frame	9 bits

Which results in : 1.125 Kbits/sec.

For better speech quality more library spectra might be used and more bits can be devoted to the quantization of the highest energy coefficient.

The speech quality in the case of 1.125 Kbits/sec. digital speech transmission using the clustering approach

is similar to the 8 Kbits/sec transmission quality of the system proposed in Chapter 4 and the 2.4 Kbits/sec. speech quality of the digital speech transmission system proposed in Chapter 5. Bit rates of 1.25 Kbit/sec and higher would result in, of course, much better speech quality.

It was mentioned earlier in Section 6.5 that the sign of the spectrum coefficients in the analysis domain has some role to play in the perception of the desired sound. To improve the speech quality further the original spectrum coefficients signs are transmitted to the receiver where they are properly placed in the chosen library spectrum. This, naturally, increases the bit rate depending on the number of signs transmitted, resulting in an improvement in the speech quality.

A block diagram of the proposed clustering system is shown in Figure 6.5.

In order to evaluate the system performance the original and processed time domain waveforms of a part of speech are shown in Figure 6.6. The Walsh-Hadamard domain representation of the original and the processed speech are shown in Figure 6.7. The presence of a small deviation of the processed speech from the original show the degree of success of the processing system proposed here.

The results of this Chapter have been published as given in Reference (92).

6.6.2. FREQUENCY DOMAIN SYSTEM APPROACH

The frequency domain has also been used as the analysis domain for the clustering approach. A system based on this is simulated and studied on a PDP-15 computer. Similar to the systems described so far in this thesis the analogue speech is low pass filtered at 3.04KHz and sampled at 8 KHz and digitized through a 10 bit A/D convertor.

Then segments of 8 miliseconds length are taken for speech analysis due to the reasons stated in Section 6.4. The Fast Fourier Algorithm (86) is used to transform each segment into the Fourier domain. As the input samples to the FFT are real only (speech signals) then using a 64 point transform for each segment (64 sample) would result in a real, even and symmetric output (Fourier domain) which means that the resolution of the samples in the Fourier domain is reduced as information is concentrated in half of the samples only. One way to overcome this problem is to make the input signal appear to be symmetric. In this case the transform length is to be doubled, that is 128 samples to be transformed for each segment of 8 msec. Then the inverse transform would result in the time domain waveform we started with together with its mirror image. A special fast algorithm called O^2 - FFT can be used to reduce the processing time (87). The procedure is shown in Figure 6.8. This is in fact similar to a cosine transform analysis (28).

Then taking the first 64 samples of the transformed speech, a decision is to be made on each block to indicate whether it is a voiced or unvoiced sound. The ratio of

energies of different regions is the basis of such a decision, as in the Walsh-Hadamard domain analysis case.

The decision threshold is given as the ratio:

$$R = \frac{\sum_{i=3}^{10} C_i^2}{\sum_{i=0}^{63} C_i^2} = \frac{\text{low frequency energy}}{\text{total spectrum energy}}$$

where C_i's are the Fourier spectrum coefficients. The threshold value was chosen from perceptual studies. Different values of R were examined to find perceptually the best threshold.

Indeed as in the Walsh-Hadamard case the value of R = 0.3 proved to be a good threshold level for the Fourier domain. That is to say below this threshold the sound is taken to be unvoiced and above it is considered voiced. Each type of sound is treated differently. In the voiced sound case in a similar way to the Walsh-Hadamard transform approach, the first two formants are detected through the peak-picking method (Section 4.2.1). A library related to the frequency domain has been formed in accordance with the procedure explained in Section 6.4.1 The relevant section of the library is then referred to from the knowledge of the first two formant positions of the incoming voiced sound spectrum. On applying the comparison algorithms the nearest spectrum shape to the incoming short time spectrum is chosen from the library. The address of the chosen spectrum shape is transmitted to the receiver.

The first formant amplitude of the incoming short time spectrum is also transmitted as the library spectra are normalized with respect to their first formant amplitude.

For the unvoiced sounds a library can be formed in a similar way to the voiced sound case with less elements as the spectral variation of such sounds is much less than the voiced sounds. Such a library can be formed through statistical investigations over large samples of unvoiced sounds. Incoming short time spectra can be compared with the library spectrum shapes employing similar algorithms as for the voiced case and the nearest one chosen. The average covariance vector approach can also be employed for the frequency domain to recover the unvoiced sounds as described in the previous chapter for the Walsh-Hadamard domain analysis case.

The speech signal is recovered at the receiver from the available information. The voiced sound spectra are taken from the relevant library and denormalized by multiplying their coefficients by the first formant amplitude. The unvoiced sounds are recovered either from the unvoiced library spectrum shapes or from the average covariance vector approach depending on the method which has been employed at the transmitter. The so formed spectra at the receiver are then inverse transformed using an Inverse Fast Fourier Transformer (IFFT) in the received order of segments.

The bit rate necessary for the digital transmission of the information varies depending on the lib ry size. As in the Walsh-Hadamard domain case a library of 32 spectrum shapes was formed for the frequency domain approach for the voiced sounds. Expansion of the library size to the next level, 64 elements, would increase the bit rate by a small fraction with an improvement in speech quality. For the unvoiced sounds two different techniques can be applied namely the library approach and the average covariance vector approach

The bit rate can then be given as:

For voiced sounds:

Address (for library size of 32)	5 bits
First formant Amplitude quantization using APCM	3 bits
Voiced/unvoiced decision	1 bit
Total bits per frame	9 bits

which result in 1.125 Kbits/sec.

For the unvoiced sound, if the average covariance vector approach is employed then similar to the WHD case four high energy coefficients would be quantized and transmitted using 2 bits per sample APCm quantization technique with one word memory (36).

The four high energy coefficients in this case are the samples 33, 34, 35, and 36 on average.

Then the bit rate in this case is :

For unvoiced sound - Average Covariance vector approach :

Quantization of four high energy coefficients	8 bits
Voiced - unvoiced decision	1 bit
Total bits per frame	9 bits

And for the library approach for the unvoiced sound if we have 16 spectrum shapes and quantize the highest energy coefficient (on average) by 4 bits PCM then we have:

For unvoiced sound - Library approach :

Address of the spectrum shape from a 16 element library	4 bits
quantization of the highest energy coefficient	4 bits
Voiced-unvoiced decision	1 bit
Total	9 bits per frame

which results in 1.125 Kbits/sec in this case as well. In order to improve the speech quality more spectrum shapes can be included in the library. This would result in only a slight increase in the overall bit rate as shown in previous section. To improve the quality even further the signs of the original spectrum coefficients can be transmitted and replaced properly in the received spectrum. This would result in higher bit rates depending on the number of signs to be transmitted. Overlapping of the blocks and winding techniques (92) were employed to observe the speech quality improvement in the frequency domain approach.

The result showed little change in the quality of speech due to application of such techniques although it leads to higher overall bit rate.

A block diagram of the frequency domain system approach is similar to the Walsh-Hadamard case as shown in Figure 6.9. For system evaluation the original time domain speech waveform is shown together with the processd one in Figure 6.10. Similarly the frequency domain representation of the original and processed speech waveform is plotted in Figure 6.11. The quality of the processed speech using the frequency as the orthogonal transform is similar to the processed speech quality using the Walsh-Hadamard transform as discussed in the previous section. This was evaluated from informal listening sessions. This indicates that the proposed system does not depend on the transform domain used for the analysis. Different orthogonal domains can be employed for various applications when it is suitable. The proposed clustering method sheds new light on the problem of the digital speech processing systems with low bit rate requirements. Using the approach discussed in this section one should be able to achieve high quality speech through a reasonable library size. However more future work is needed to explore and improve different aspects of the method.

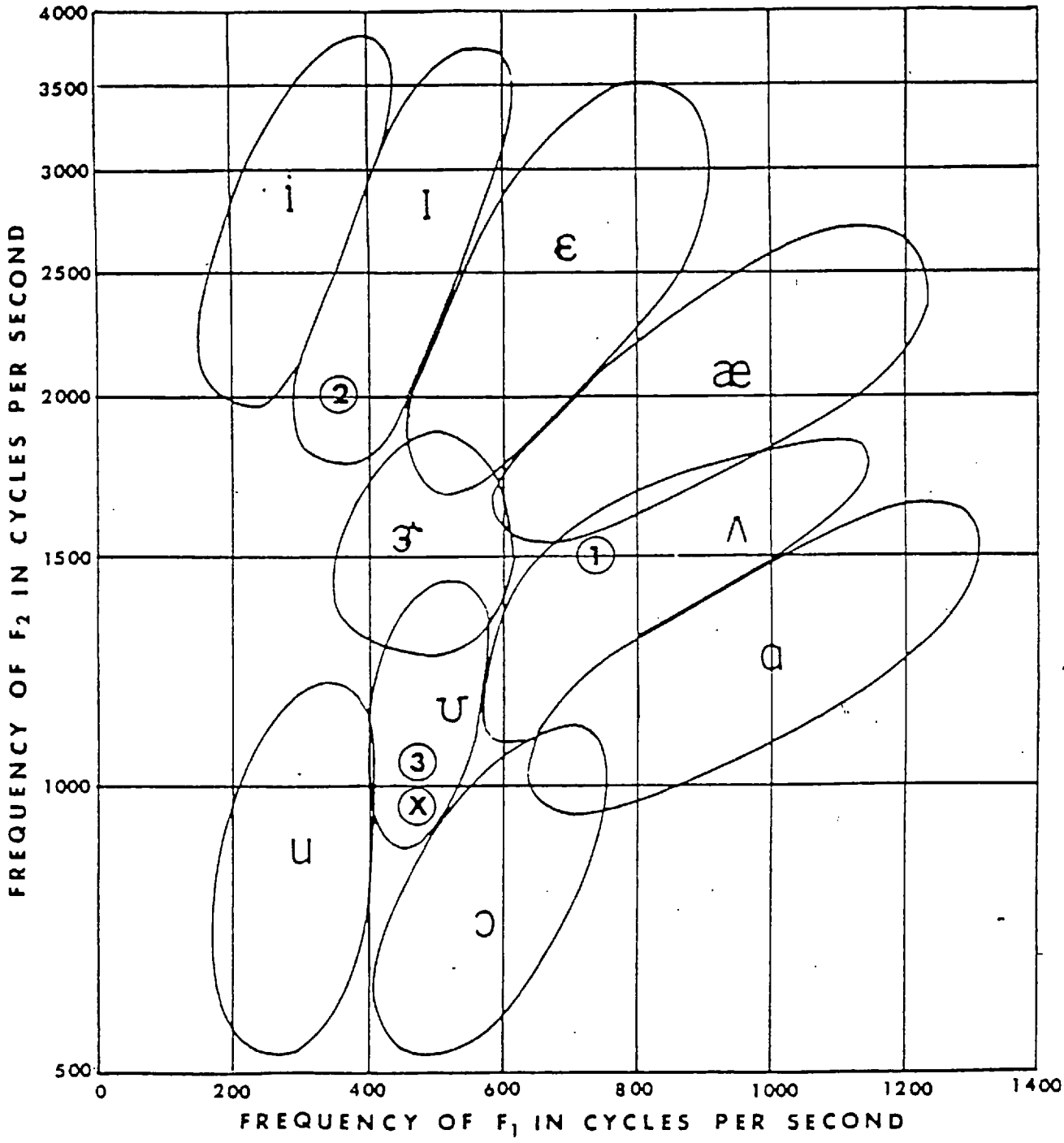


Fig. 6.1: F_1-F_2 graph for Vowel identification

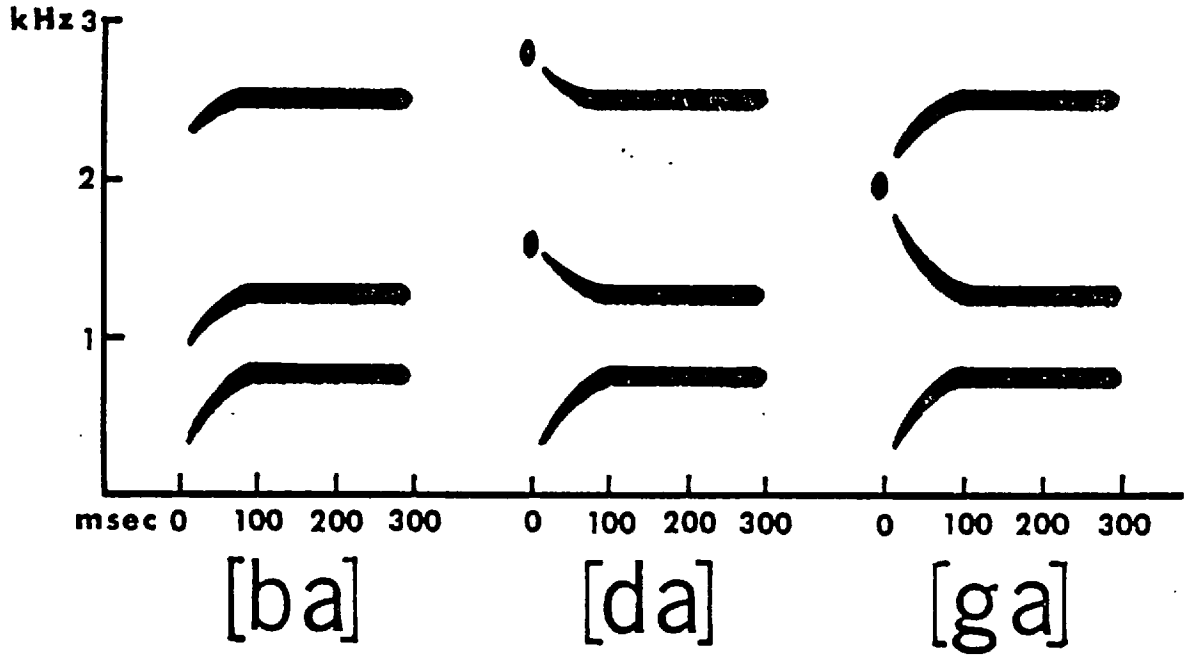


Fig. 6.2: Typical Formant Transitions in the case of the Consonant sounds

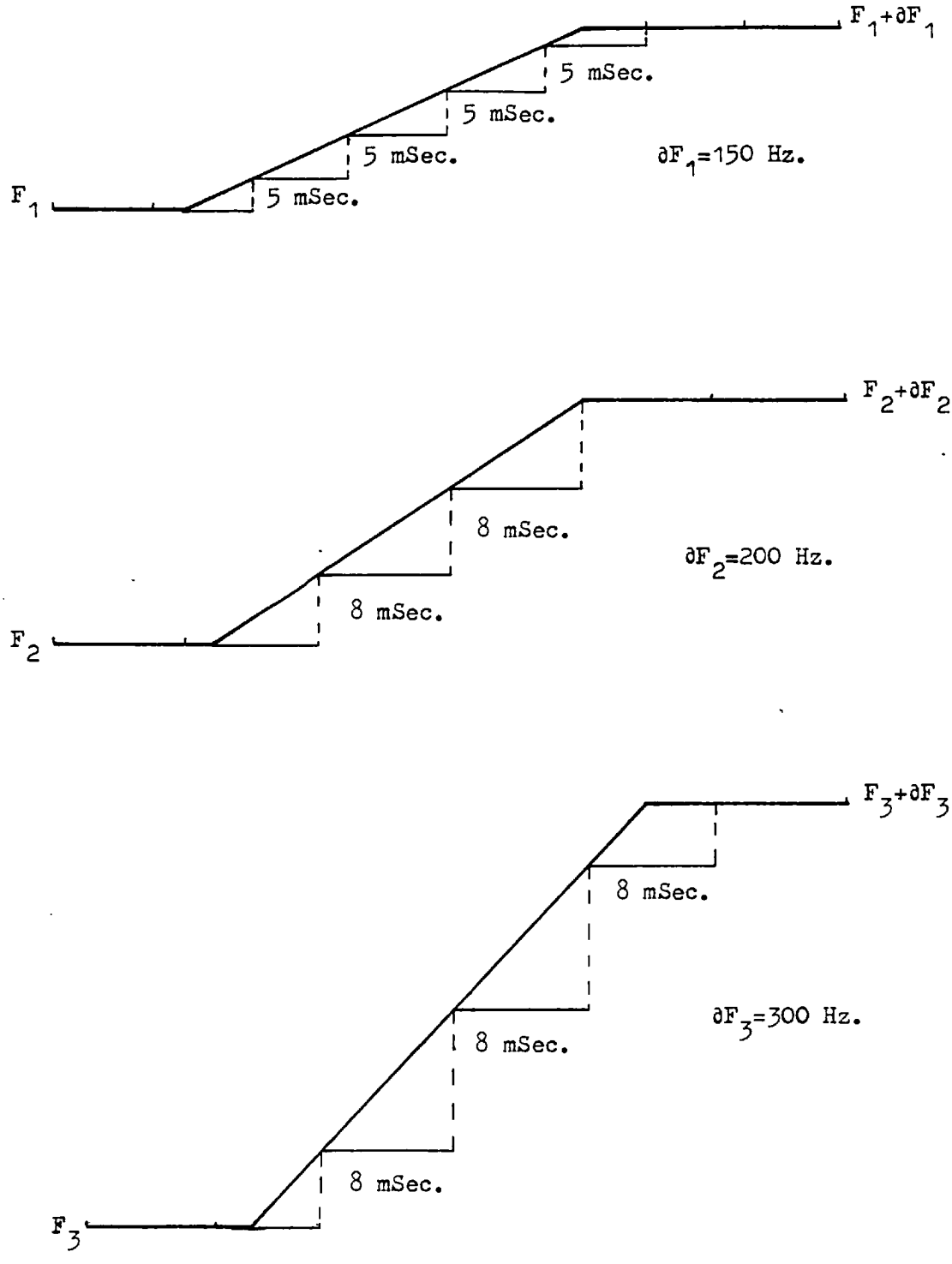


Fig. 6.3: Formant Transitions with different slopes and their staircase approximations for the Consonant sounds

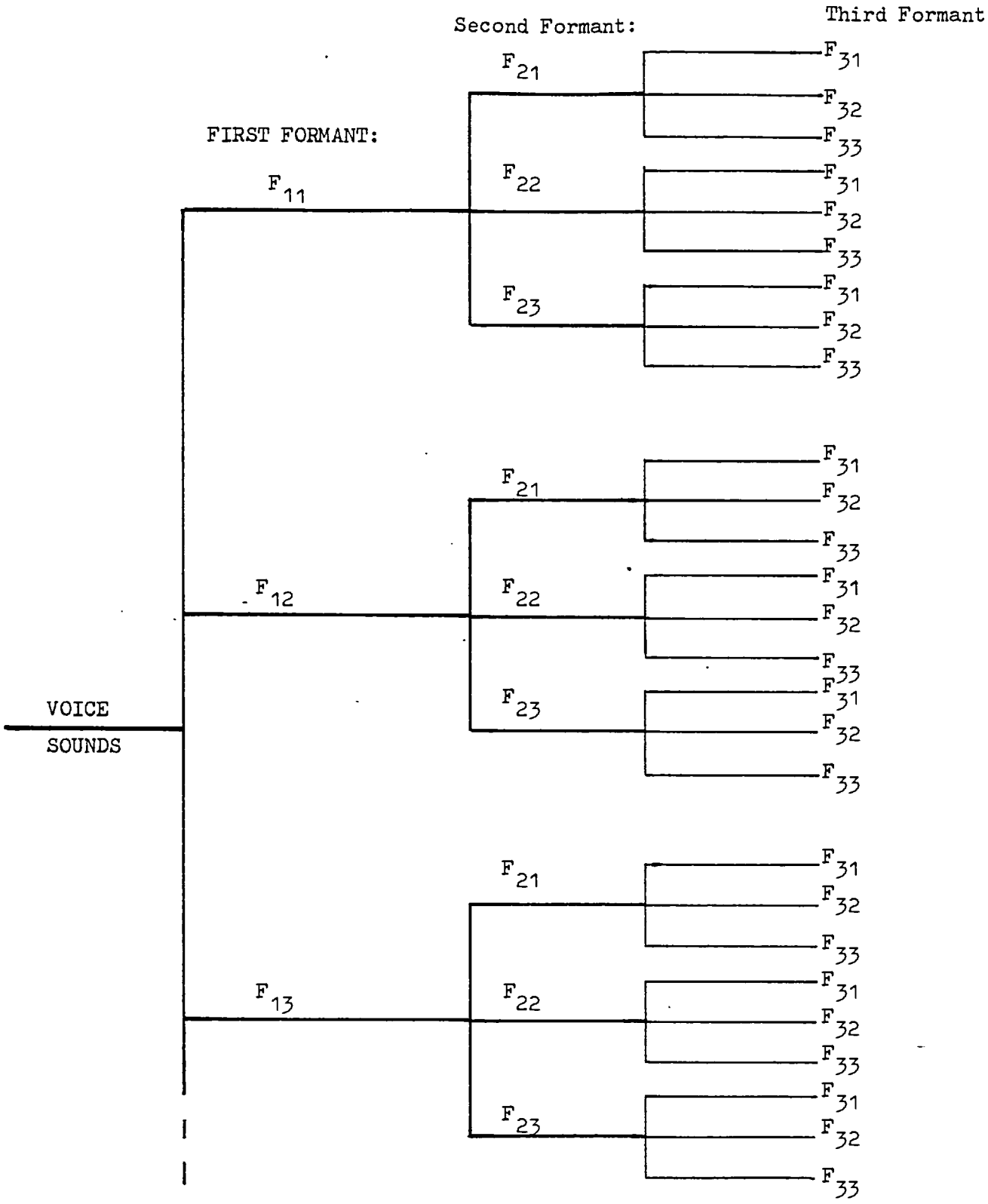


Fig. 6.4: A Tree representation of the Library Spectrum Shapes

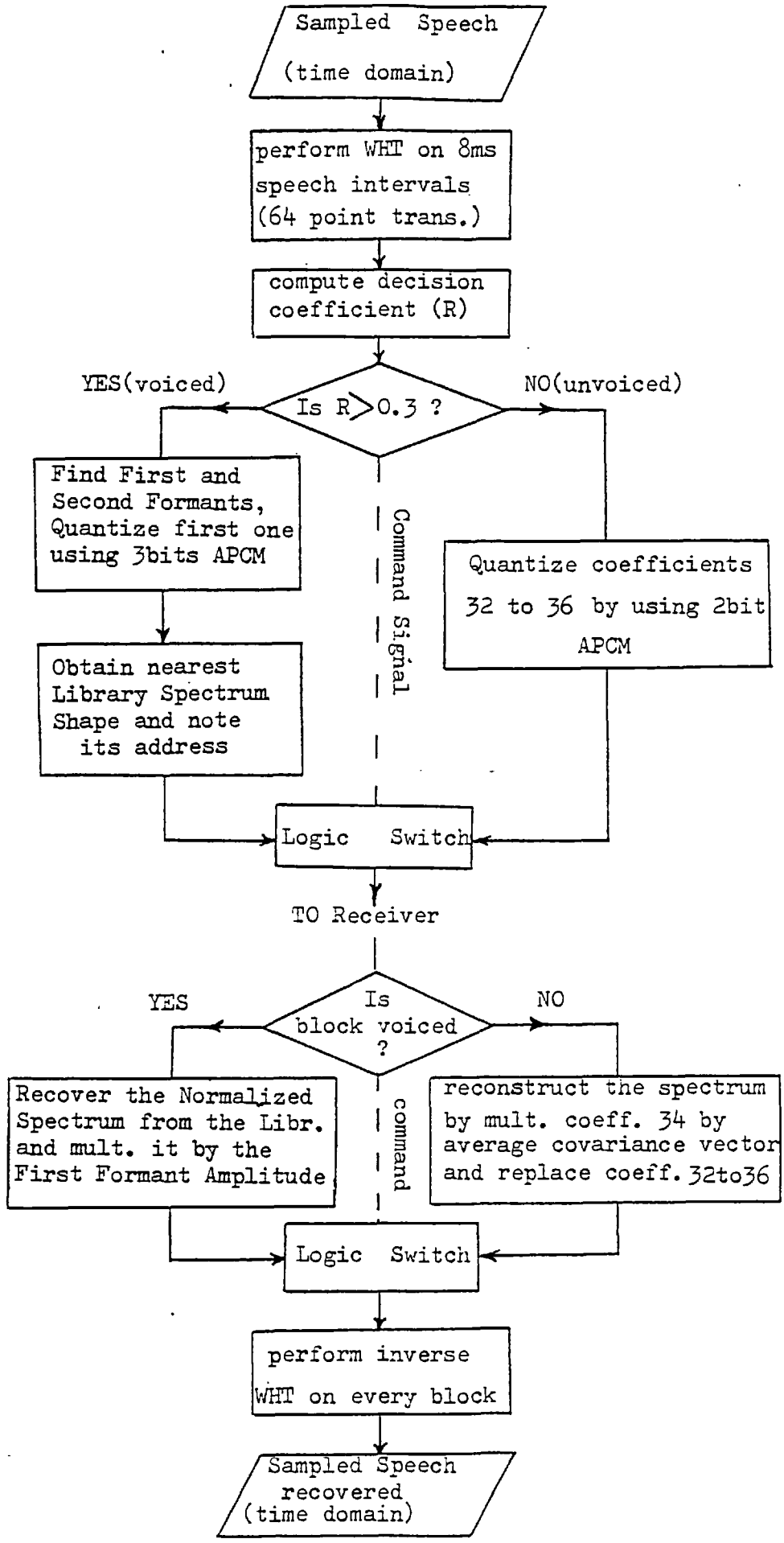


Fig. 6.5: Block Diagram of the Clustring System based on the Walsh domain

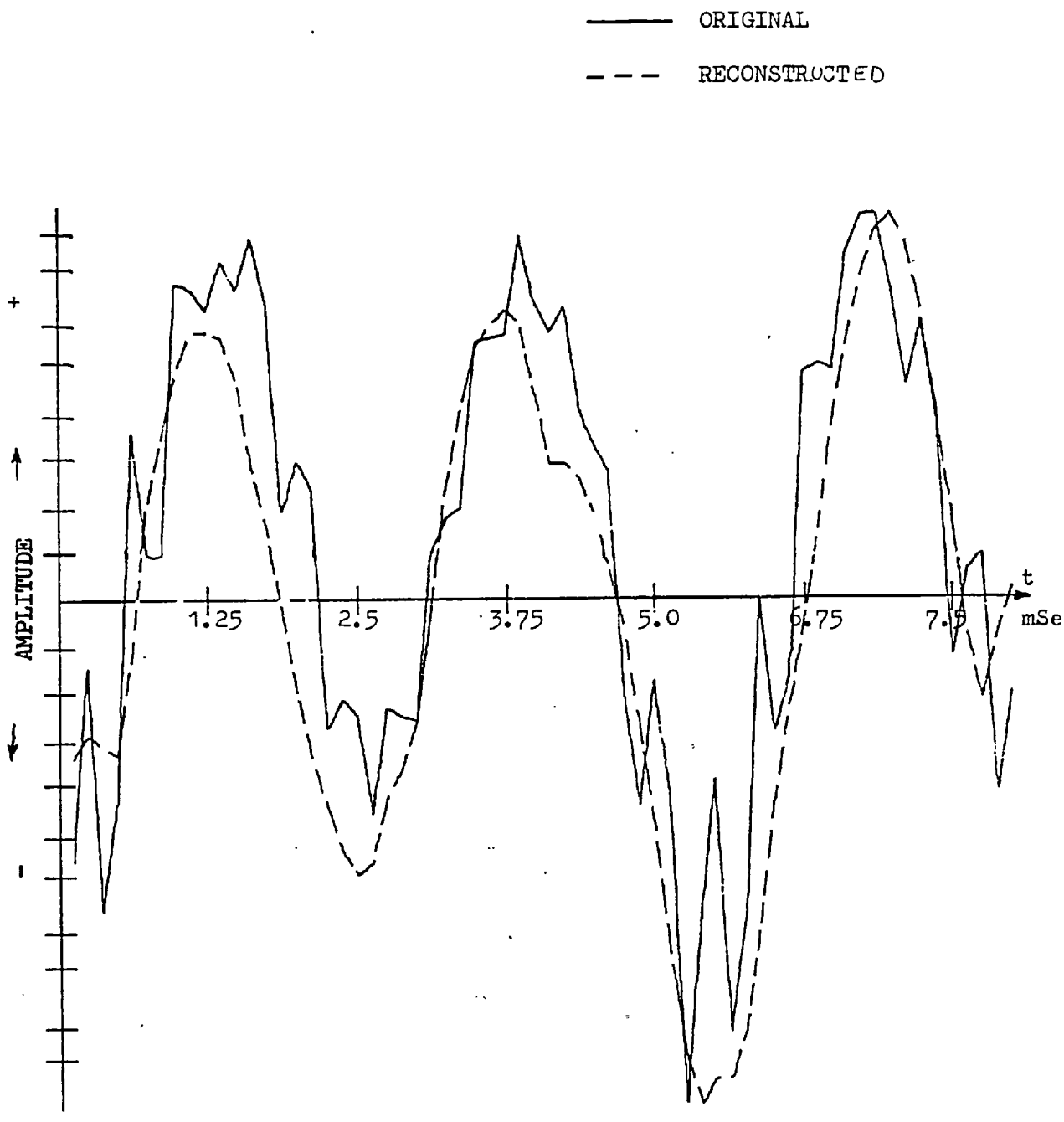


Fig. 6.6: Time domain representation of part of the Original and the Reconstructed speech signal using the Clustering approach(Walsh Domain case)

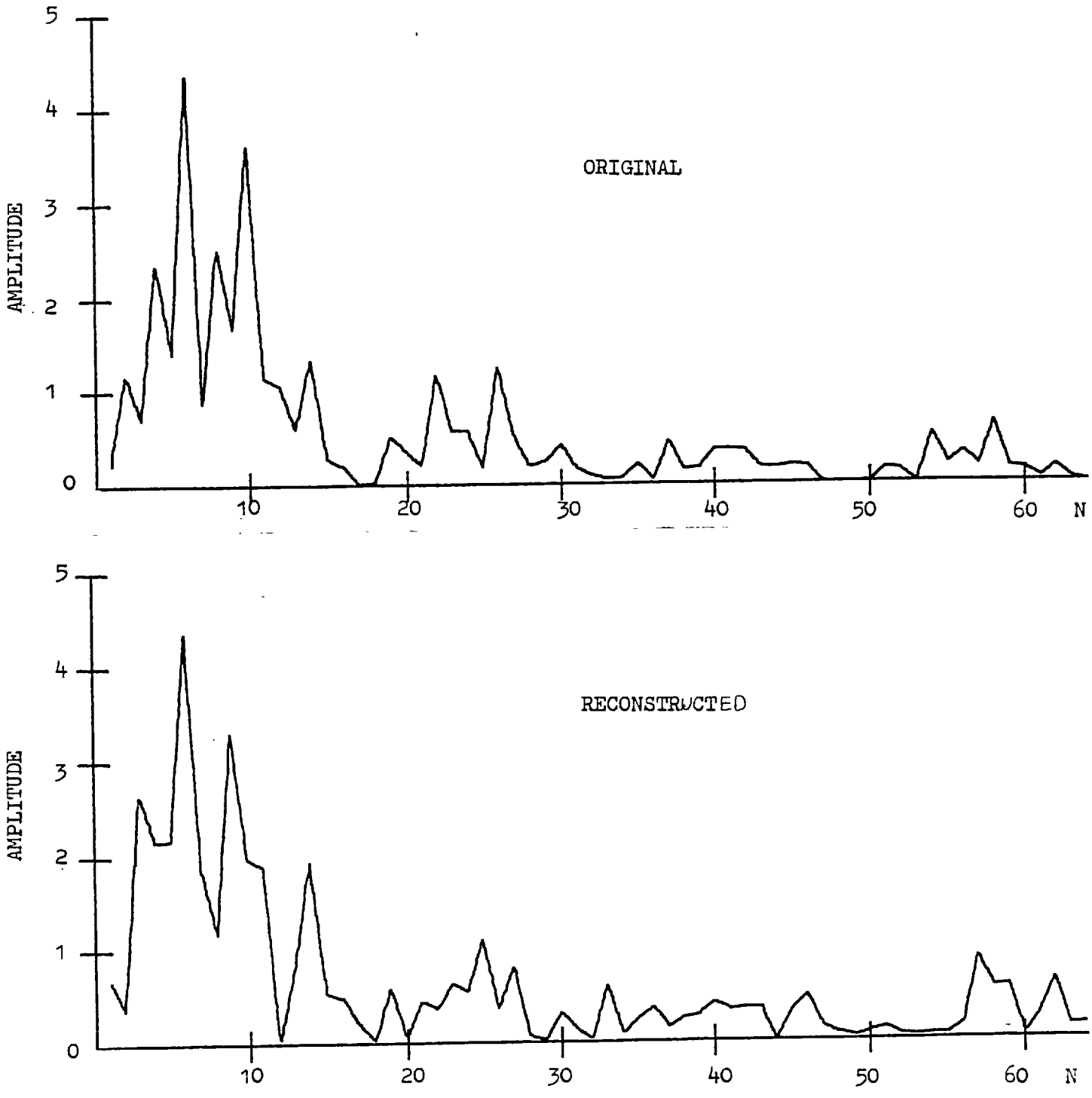


Fig. 6.7: The Walsh domain representation of part of the Original and the Reconstructed speech signal using the Clustering approach

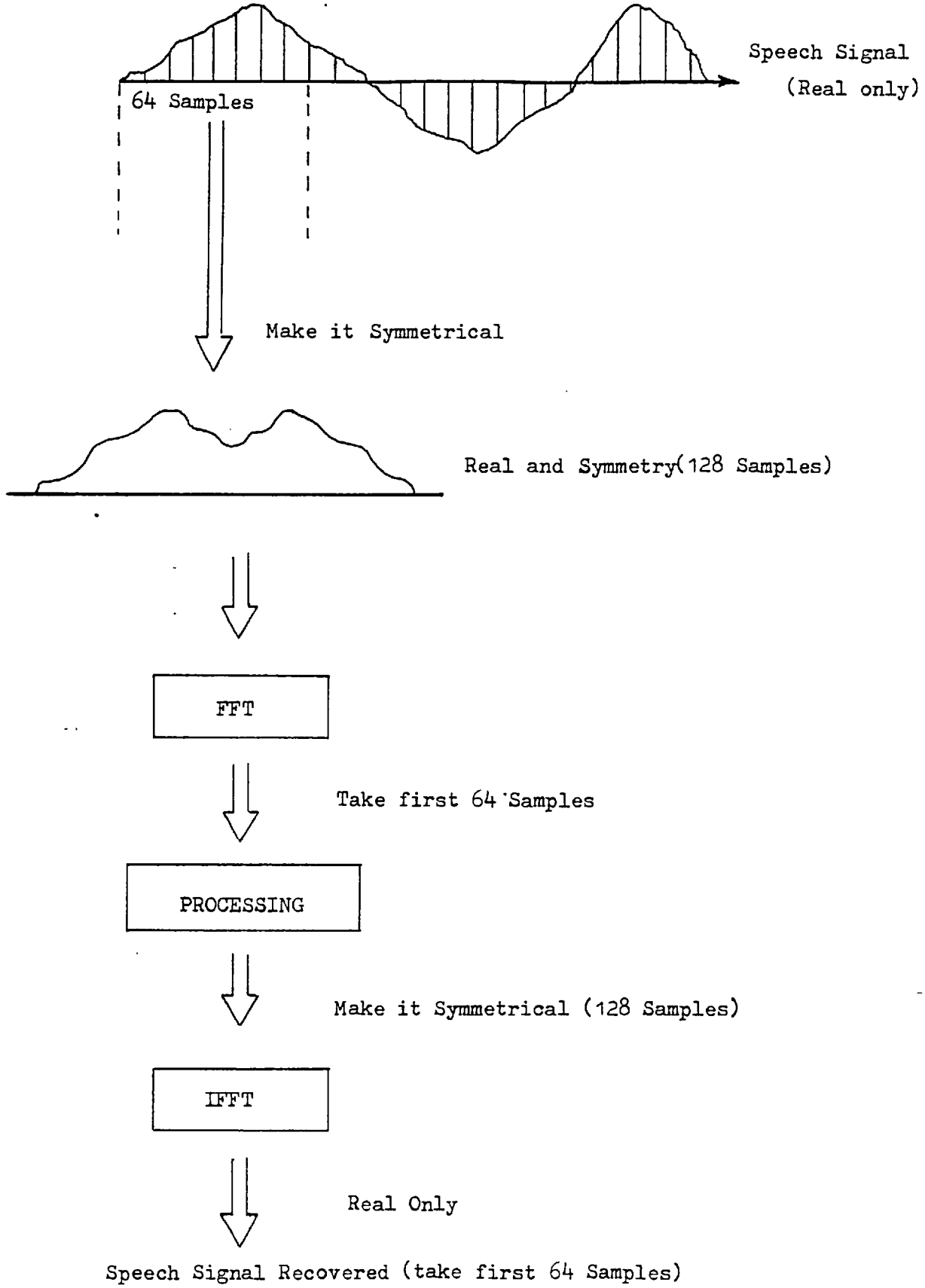


Fig. 6.8:The illustration of the symmetrical real input method for FFT usage

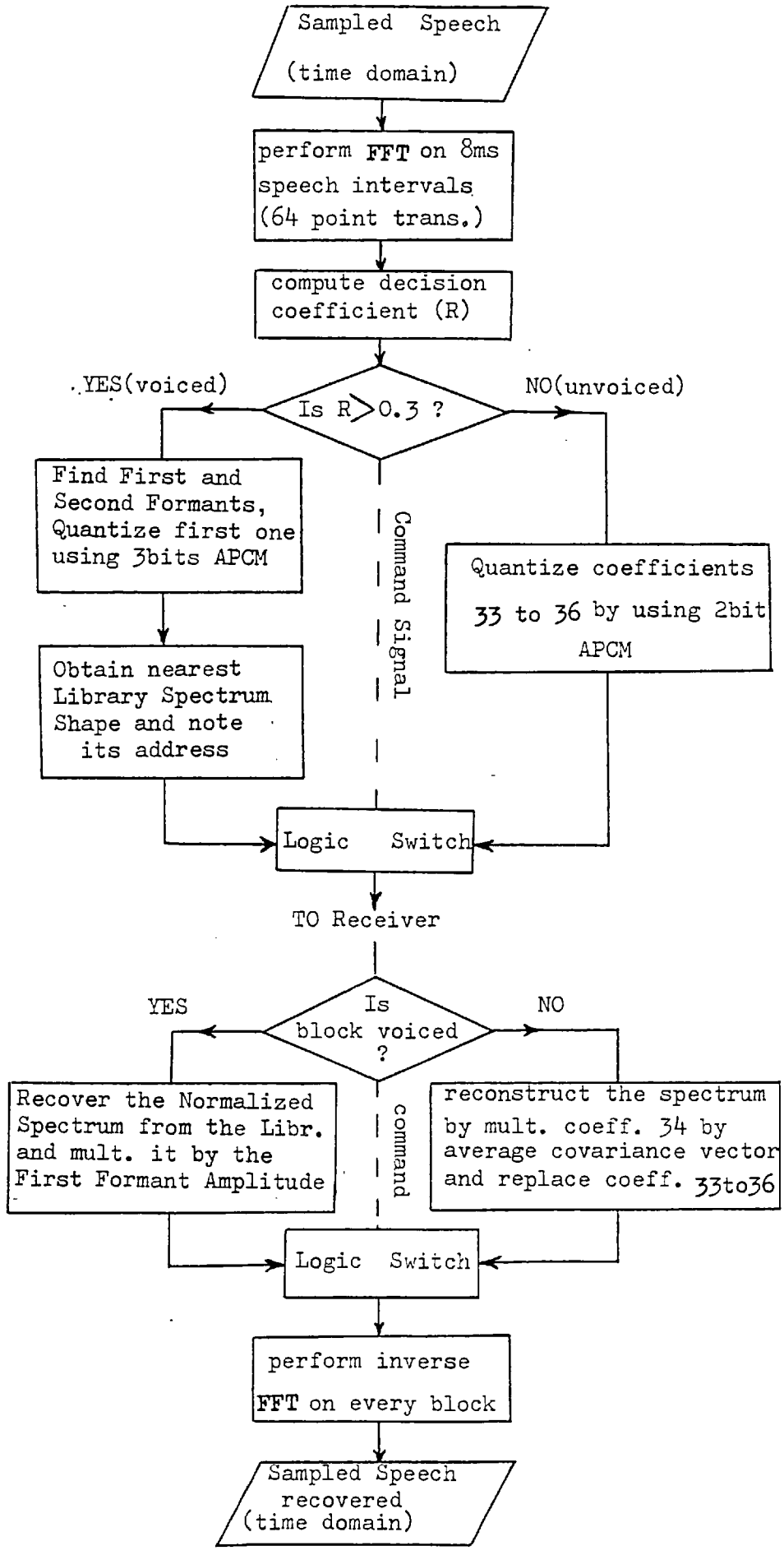


Fig. 6.9: Block diagram of the Clustering system based on the Frequency domain

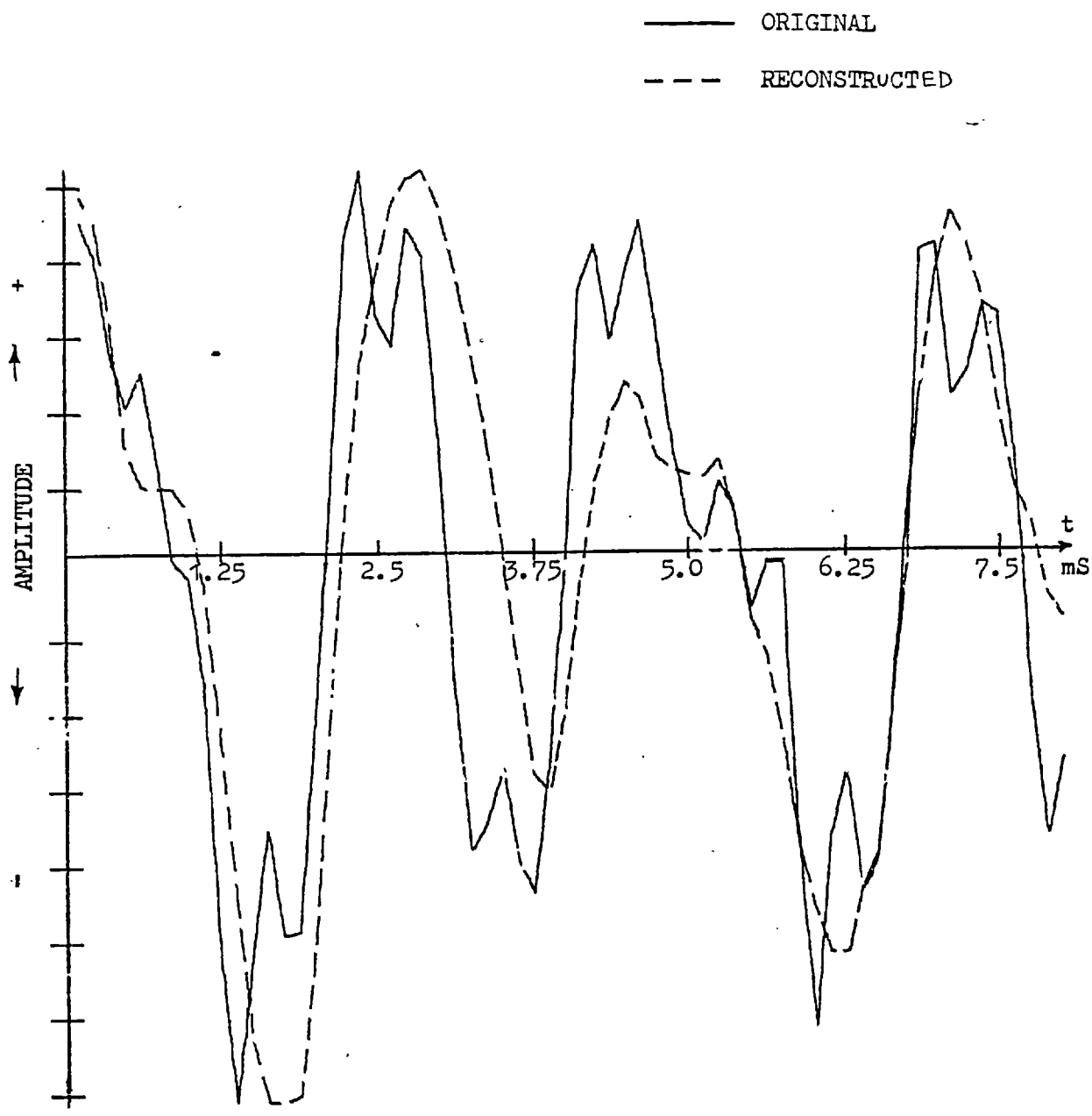


Fig. 6.10: Time domain representation of part of the Original and the Reconstructed speech signal using the Clustering approach(Frequency domain case)

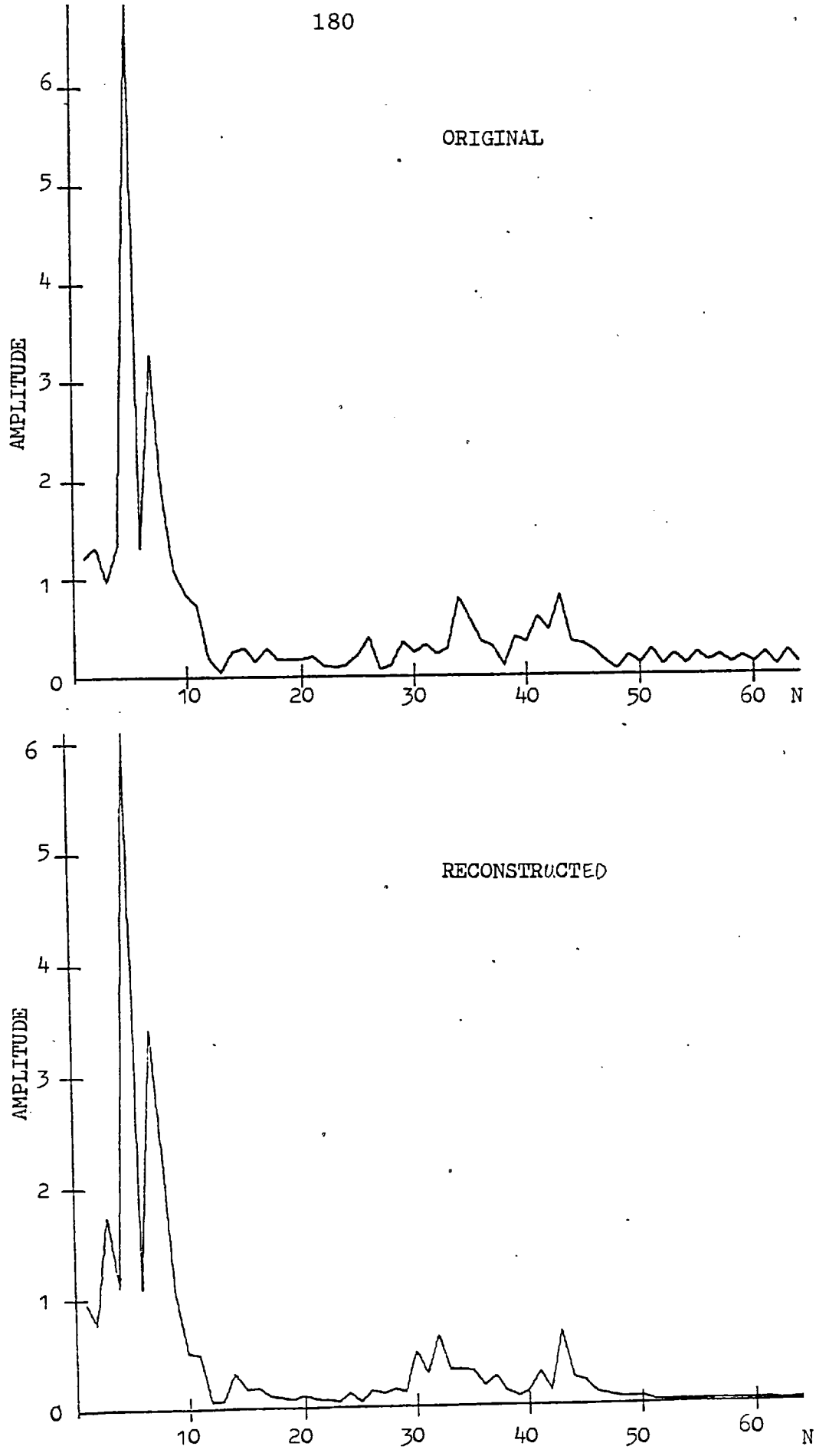


Fig. 6.11: The Frequency domain representation of part of the Original and the Reconstructed speech signals using the Clustering approach.

CHAPTER SEVEN

CONCLUSIONS

AND SUGGESTIONS FOR FURTHER RESEARCH

7.1. CONCLUSIONS

The aim of the research reported in this thesis has been to develop relatively simple and efficient systems for the digital transmission of speech. Three new low-bit-rate digital speech processing systems have been presented here.

In order to achieve the initial aim different possibilities were investigated to extract the information bearing parameters from the speech signals, and use them at the receiver to recover intelligible speech. The fact that the speech is produced by slow movements of the articulatory mechanism is taken into account, and information related to these movements has been extracted. The aim is then achieved by making use of the speech characteristics to their fullest extent for removing redundancy. For efficiency the speech signals are analyzed in transform domains other than the time domain. This is because the transform domains offer a non-uniform energy distribution in the "spectrum" and highlights the information bearing parameters of the speech waveforms. Moreover they remove, to some degree, the correlation between the speech samples. Almost any orthogonal transform can be used as the analyzing domain. Among them the Walsh-Hadamard domain is more suitable for digital systems and is used in the proposed

systems. The frequency domain has also been used as the system analysis domain. The voiced-unvoiced sounds of speech segment are processed separately because they exhibit different characteristics both in the time and frequency domains. The parameters considered to be vital for the perception of the voiced sounds are the resonances of the vocal tract referred to as the Formants. The formants are the distinct peaks available in the transform domain representation of the voiced sounds.

The systems in this thesis have been developed in an evolutionary way and the merits of each system are discussed below.

(i) The perceptual importance of the formants in the voiced segments of speech led to the design of a simple speech processing system with transmission rate of around 4 Kbits per second whereby fixed speech segments of 8 msec. duration are analyzed in the Walsh-Hadamard domain where the voiced-unvoiced sounds are separated according to their energy distribution. The voiced sounds are identified and recovered using their first three formants. The unvoiced sounds are recovered from their high energy regions. The system is simple when compared with those of similar transmission rate and speech quality. Indeed the speech quality can be improved if regions around the formants are also transmitted which, of course, will lead to higher transmission rates. The system can be used in circumstances where a simple system of low transmission rate is required and the intelligibility is the prime objective rather than speech quality.

The next step of the research was to design a system with lower transmission rate than the above, but of a similar speech quality.

(ii) To fulfill the original aim of achieving a simple low bit rate digital speech transmission system further investigation was carried out to exploit the speech characteristics. In this respect the linear prediction technique in the transform domain has been used to predict the higher order formants of the voiced sounds from the first formant. The unvoiced sounds in this system are recovered by employing a so-called average covariance vector obtained through extensive statistical investigations.

The Walsh-Hadamard domain has been employed in the proposed system but in principle any other orthogonal domain can be used. The transmission rate for intelligible speech recovery has been reduced from 4 K bits/sec obtained for the previous system to 1.2 K bits/sec for this system. Improved speech quality can be achieved by predicting regions around the formants which would increase the transmission rate to 2.4 K bits/sec. This proposed system has a simple structure compared with those of similar transmission rate. It can be used for applications where a simple structure is essential and intelligibility is of more importance than speech quality.

(iii) Thus far transmission rate achieved was around 1.2 K bits/sec with a relatively simple system for intelligible speech recovery. The concern at this stage was to keep the transmission rate and complexity approximately at the same level and improve the speech quality. In the proposed

systems, only few coefficients from the spectrum have been transmitted and the rest of the spectrum has been put to zero. This fact proved to be the main determining reason for the speech quality obtained. The investigation was then concentrated on developing a system with transmission rate of approximately 1K bits/sec and improved speech quality. To achieve this a novel method was introduced in which the short time spectrum shapes of the speech segments are stored with some degree of accuracy in a memory. Speech segments of 8 msec duration are considered for analysis for all types of voiced sounds. Then a library of 32 elements of the typical short time spectrum shapes was formed in the case of voiced sounds with a smaller library for the unvoiced sounds of 16 elements.

The speech waveform is then segmented into short sections and compared to the relevant library elements. The nearest spectrum shape in then chosen and its address transmitted. The short-time spectrum is recovered at the receiver by referring to the identical library as in the transmitter. The system is fairly simple when compared to digital speech processing systems with similar transmission rates. The speech quality is improved in comparison with the two systems discussed earlier, even though the transmission rate is in the region of 1 K bits/sec. Generally any orthogonal domain can be employed as the analysis domain but in the simulation carried out in this thesis the Walsh-Hadamard as well as the normal frequency domain have been used. The approach presented here gives a new insight into the problem of speech

processing in the transform domain. The expansion of the library would improve the speech quality in the proposed system and the system itself can be used in situations where very low transmission rate is essential and intelligible speech with reasonable quality is needed. The approach can also be used for speech synthesis systems where certain sounds, words or even sentences can be synthesied by appropriate addressing of the library spectrum shapes.

In its entire form the system is compatible with existing speech synthesiers or speech transmission systems of similar transmission rates.

7.2 SUGGESTIONS FOR FURTHER RESEARCH

The clustering approach as developed in this thesis, introduces new insight into the low bit-rate speech analysis-syn thesis systems. The general idea of classification can be applied to many types of speech processing systems. This is due to the fact that there normally exists some degree of correlation between various parameters of a speech processing method. These parameters usually have limited variations and therefore a sizeable set can represent them with a high degree of accuracy. To follow this approach a promising area of research is to apply the clustering method in the Linear Prediction Coding system with the intention of reducing the transmission rate while keeping the quality of recovered speech almost unchanged. In the LPC speech processing system normally a set of twelve parameters are used to represent the speech waveform of 30 msec duration where 72 bits are

are transmitted to represent these parameters (53). On the other hand there exists some degree of correlation between successive blocks due to the nature of speech. This correlation can be exploited through different techniques, one of which is the clustering approach. Employing this approach a library of typical parameter sets can be formed where the input set is checked against the library entries and the nearest one is chosen and its relevant address transmitted. This could lead to a substantial reduction of transmission rate while keeping the speech quality similar to normal LPC systems.

Further research can be carried out on the clustering system approach presented here. Investigations on system performance for different orthogonal domains could lead to wider applications of such speech processing systems. Improvements can be achieved by increasing the short time spectrum library size. The reduction of transmission rate would be possible if the correlation between successive speech blocks is exploited. Other techniques such as overlapping the analysis blocks and windowing of such blocks might lead to quality improvement in certain domains. Different techniques can be used for the processing of unvoiced sounds with the aim of reducing bits per block necessary for their transmission compared to the voiced sounds. The most obvious disadvantage would be the variable transmission rate resulting from such an approach.

More exact voiced-unvoiced decision algorithms can be used which result in a more complex system. Reducing the analysis length might also be useful with the concomitant disadvantage of increasing the transmission rate.

REFERENCES:

1. Dudley, H. and Tarnoczy, T.H.,
"The Speaking Machine of Wolfgang von Kempelen".
J. Acoust. Soc. Amer. Vol. 22. pp. 151-166, 1950.

2. Wheatstone, Charles; "Scientific papers of Sir Charles Wheatstone".
Physical Society of London, 1879.

3. Bell, A.G., "Prehistoric telephone days".
Nat. Geograph. Mag. Vol. 41, pp. 223-242, 1922.

4. Paget, Sir Richard, "Human Speech".
Kegan Paul, Trench, Trubner and Co. Ltd.
London, 1930.

5. Dudley, H. "Remaking Speech",
J. Acoust. Soc. Amer., Vol. II, pp. 169, 1939.

6. Potter, R.K., Kopp, G.A., Green, H.C.
"Visible Speech".
D. Van Nostrand, Co. Inc., New York 1947.

7. Reeves, A.H. "Electrical Signaling System".
International Standard Electric Corporation,
U.S. patent 2, 272, 070, Feb 3, 1942.

8. Dunn, H.K. "On vowel resonances and an electrical vocal tract".
J. Acoust. Soc. Amer,, Vol 22, pp. 740, 1950.

9. Shannon, C.E. "A mathematical theory of Communications".
Bell System Tech. Jour. Vol. 27, pp. 623-656, 1948.

10. Jacobson, R., Fant, C.G.M., Halle, M. "Preliminaries to Speech Analysis"
Technical report No. 13, Acoustic Laboratory, M.I.T., 1952.

11. Flanagan, J.L. "Automatic Extraction of formant frequencies from Continuous Speech".
J. Acoust. Soc. Amer. Vol. 28, pp.110-118, 1956.

12. Fant, G. "On the predictability of formant levels and spectrum Envelopes from formant frequencies."
In For Roman Jackobson, S. Gravenhage, pp. 109-120, 1956.

13. Smith, B. "Instantaneous Companding of Quantized Signals".
Bell System Tech. Jour. 1957. pp. 653-709.

14. Flanagan, J.L. "Speech Analysis, Synthesis and Perception.
Springer-Verlag, 1965.

15. Golden, R.M. "Digital Computer Simulation of a Sampled-data Voice-excited vocoder".
J.Acoust. Soc. Amer. Vol. 35, 1964.

16. Noll, A.M. "Short-time Spectrum and Cepstrum techniques for Vocal pitch detection". J. Acoust. Soc. Amer. vol. 36, pp. 296-302, Feb. 1964.

17. Kulya, V.I. "Experimental investigation of the correlation relations in the speech spectrum and a comparison of some variants of orthogonal vocoder". Telecommunications, Vol. 18, No. 4, pp. 39-50, 1966.

18. Mcdonald, R.A. "Signal to noise and idle channel performance of differential Pulse Code Modulation Systems - Particular application to Voice Signals". Bell System Tech. Jour. Vol 45, pp. 1123 - 1151. 1966.

19. Crowther, W.R. Rader, C.M. "Efficient Coding of Vocoder Channel Signals using Linear Transformation". Proceedings of the IEEE, Vol. 54, pp. 1594-1595, Nov. 1966.

20. Oppenheim, A.V., Schafer, R.W., "Homomorphic Analysis of Speech", IEEE Trans. Audio Electroacoust., Vol. AU-16, pp. 221-226, June 1968.

21. Atal, B.S., Schroeder, M.R., "Predictive Coding of Speech Signals". Proc. Int. Congr. Acoust. C-5-4, Tokyo, Japan, August 1968.

22. Cattermole, K.W. "Principles of Pulse Code Modulation".
Iliffe books, London, 1969.

23. Bower, D.E. "Walsh functions, Hadamard matrices and data compression".
Proc. of Symp. on Applied Walsh function.
Wash. D.C. pp. 33-37, 1970.

24. Klein, W. "The transformation error of the Walsh Vocoder".
Nachrichtent, Zeitchtent, Zeitschrift, NTZ, 23, pp. 126-128, 1970.

25. Harmuth, H.F. "Transmission of Information by orthogonal functions".
2nd edition. Springer-Verlag, New York 1970.

26. Ahmed, N., Rao, K.R. "Orthogonal transforms for digital Signal Processing".
Springer-Verlag, 1975.

27. Shum, Y.Y., Elliott, A.R. Brown, W.
"Speech processing with Walsh-Hadamard transforms".
IEEE Trans. on Audio Electroacoustic, Vol. Au-21, pp.174-17o, 1973.

28. Ahmed, N. Natarajan, T., Rao, K.A.
"Discrete Cosine transform".
IEEE Trans. on Computers. Vol. C-23, pp.90-93, 1974.

29. Pratt, W.K., Chen, W.H., Welch, L.R.
"Slant Transform Image Coding".
IEEE Trans. on Communications, Vol. COM-22, No. 8, Aug. 74.

30.. Rader, M. "Discrete Convolutions via Mersenne transforms".
IEEE Trans. on Computers, Vol. C-21, pp. 1269-1273, Dec. 1972.

31. Tribolet, J.M., Crochiere, R.E. "Frequency Domain Coding of Speech".
IEEE Trans. ASSP-27, No. 5, pp. 512-530. Oct. 1979.

32. Markel, J.D. Grey, A.H.Jr., "On Auto-Correlation Equations as applied to Speech Analysis."
IEEE Trans. Audio and Electroacoust. Vol. AU-21, No. 2, pp. 68-79, April 1973.

33. Makhoul, J. "Spectral Analysis of speech by linear prediction".
IEEE Trans. Audio Electroacoust. Vol. AU-21, pp. 140-148. June 1973.

34. O'Neal, J.B., Jr., Stroh, R.W., "Differential PCM for Speech and data Signals".
IEEE Trans. on Communications, Vol. Com-20, pp. 900-912, Oct. 1972.

35. Cummiskey, P., Jayant, N.S., Flanagan, J.L. "Adaptive quantization in differential PCM coding of speech".
Bell System Tech. Jour. pp. 1105.1118. Sept. 1973.

36. Jayant, N.S. "Adaptive quantization with a one-word memory".
Bell System Tech. Jour. Vol. 52, pp. 1119-1144, 1973.

37. Jayant, N.S. "Digital coding of speech waveforms, PCM, DPCM and DM quantizers".
Proc. IEEE, Vol. 62, pp. 611-632, May 1974.

38. Noll, P. "A comparative study of various quantization schemes for speech encoding".
Bell System Tech. Jour. Vol-54, pp. 1597-1614, Nov. 1975.

39. King, R.A., Gosling, W. "Time-Encoded Speech".
Electronics Letters, No. 15, pp.456-457, July 1978.

40. Zelinski, R. Noll, P. "Adaptive Transform Coding of Speech Signals".
IEEE Trans. on ASSP. Vol. ASSP-25, No. 4. pp. 299-309, August 1977.

41. Zelinski, R. Noll, P. "Approaches to Adaptive transform Speech Coding at Low bit rates".
IEEE Trans. on ASSP, Vol ASSP-27, No. 1, Feb. 1979.

42. Crochiere, R.E., Sambur, M.R.,"A Variable Band Coding Scheme for Speech encoding at 4.8 K bits/sec."
Bell System Tech. Jour. Vol. 56, No. 5, May-June 1977.

43. Wilkinson, M. "An adaptive pulse Code Modulator for Speech'
Proc. Int. Conference on Communications, pp. 1-11 Montreal, June 1971.

44. Castellino, P., Modena-Nebbia, G., Scagliola, C., "Bit rate reduction by automatic adaptation of quantizer step-size in DPCM systems".
International Zurich Seminar on digital communications, April 1974.

45. Steele, R. "Delta Modulation Systems".
Pentech Press, London, 1975.

46. Delaraine, E.M., Van Mierlo, S., Derjavitch, B. "Delta Modulation Systems"
French patent 932, 140, August 10, 1946.
U.S. patent No. 2,629, 857, Feb. 24, 1953.

47. Beranek, L.L. "The design of Communications Systems".
Proc. IRE, 35, pp. 880-890, Sept. 1947.

48. UN.C.K. "A Low Rate digital formant Vocoder".
IEEE Trans. on Comm., Vol. Com-26 No. 3, March 1978.

49. Crochiere, R.E., Webber, S.A., Flanagan, J.L. "Digital Coding of Speech in Sub-bands".
Bell System Tech. Jour. Vol. 55, No. 8, 1976.

50. Atal, B.S., Manauer, S.L. "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave".
J. Acoust. Soc. Amer., Vol. 50, No. 2, pp. 637-655, August 1971.

51. Makhoul, J. "Spectral Linear Prediction: Properties and Applications". IEEE Trans. Acoust. Speech and Signal Processing, Vol. ASSP-23, Nol 3, June 1975, pp. 283-296.

52. Markel, J.D., Grey, A.H., Jr. "Linear Prediction of Speech". Springer-Verlag, 1976.

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

54. Chandra, S. Lin, W.C. "Linear Prediction with a Variable Analysis Frame Size". IEEE Trans. on Acoust. Speech Signal Process. Vol. ASSP-25, 1977.

55. Itakura, F., Saito, S. "On the optimum quantization of feature parameters in the PARCOR speech synthesizer". Proc. IEEE Conf. Speech Communication Process. pp. 434-437, 1972.

56. Ray, W.D. Driver, R.M. "Further decomposition of the Karhunen-Loève series representation of a stationary random process". IEEE Trans. on Information Theory, Vol. IT-16, No. 6, Nov. 1970, pp. 663-668

57. Ashouri, M.R. "Transform Coding techniques for low-bit-rate speech communication". Ph.D. Thesis, Imperial College, January 1979.

58. Oppenheim, A.V., Schafer, R.W. "Digital Signal Processing".
Prentice-Hall, Inc., New Jersey, 1975.

59. Geckinli, N.C., Yavuz, D. "Some novel windows and a concise tutorial comparison of window families".
IEEE Trans. on Acoust. Speech Sign. Process. Vol. ASSP-26, No. 6, pp. 501, 507, 1978.

60. Walsh, J.L. "A closed set of normal orthogonal functions".
Amer. Jour. Maths., 45, pp. 5-24, 1923.

61. Shanks, J.L. "Computation of the Fast Walsh-Fourier Transform".
IEEE Trans. on Computers, pp. 457-459. May 1969.

62. Berauer, G. "Fast in place computation of the discrete Walsh transform in sequency order".
Proc. Symp. Application of Walsh Functions, Wash. D.C., pp. 272-275, 1972.

63. Shore, J.E., "On the applications of Haar functions".
IEEE Trans. on Communications, pp. 209-216, March 1973.

64. Frangoulis, E.D. "Orthogonal transform methods of speech coding".
Ph.D. Thesis, Imperial College, June 1978.

65. Nussbaumer, H.J. "Digital filtering using Psuedo Fermat number transforms".
IEEE Trans. Acoust. Speech Sign. Proc., Vol. ASSP-25, pp. 29-83, Feb. 1977.

66. Flanagan, J.L., Schroeder, M., Atal, B., Crochiere, R.E., Jayant, N.S., Tribolet, J.M. "Speech Coding". IEEE trans. on Communications, Vol Com-27, No. 4, April 1979.

67, Fant, G. "Acoustic Theory of Speech Production". Mouton & Co. S-Gravenhage 1960.

68. Bell, C.G., Fujisaki, H. Heinz, J.M. Stevens, K.N. House, A.S. "Reduction of Speech Spectra by Analysis by Synthesis Techniques". J. Acoust. Soc. Amer. Vol. 33, pp. 1725-1736. 1961.

69. Schafer, R.W., Rabiner, L.R., "System for Automatic formant Analysis of Voiced Speech". J. Acoust. Soc. Amer., Vol. 47(2), pp. 634-648, 1970.

70. Rabiner, L.R., Schafer, W. Rader, C.M. "The chirp Z-transform algorithm and its applications". Bell System Tech. Jour. Vol. 48, pp. 1249-1292, 1969.

71. Suzuki, J., Kadokawa, Y., Nakata, K. "Formant-frequency extraction by the method of moment calculations". J. Acoust. Soc. Amer., Vol. 35, Sept. 1963.

72. Christensen, R.L., Strong, W.J., Palmer, E.P. "A comparison of three methods of extracting resonance information from Predictor-Coefficient Coded Speech".

IEEE, Acoust. Speech Sign. Process.
Vol. ASSP-24, No. 1, pp. 8p14, 1976.

73. Cooley, J.W., Tukey, J.W.,"An algorithm for Machine Computation of Complex Fourier Series".
Mathematics of Computation, Vol. 19. pp. 297-301, 1965.

74. Flanagan, J.L., Coker, C.H.,Bird, C.M. "Computer Simulation of a formant vocoder . Synthesizer".
J. Acoust. Soc. Amer. Vol. 35, pp. 2003(a), 1962.

75. Carlson, R., Granstrom, B., Fant, G. "Speech Perception, some studies concerning perception of isolated vowels".
Haskins Laboratories, Report No. STL-QPSR 2-3/1970.

76. Edmondson, W.H. "Novel frequency analysis system for a vibrotactile speech training aid for the deaf".
Electronic Circuits and Systems, Vol. 1. No. 2, Jan. 1977.

77. Robinson, G.S., "Logical convolution and discrete Walsh and Fourier Power Spectra".
IEEE Trans. on Audio Elec troacoustis. Vol. AU-20, No. 4, 1972.

78. Rabiner, L.R., Cheng, M.J., Rosenberg, A.E., Mcgonegal, C.A. "A comparative performance study of several pitch detection algorithms". IEEE Trans. on Acoust. Speech Sig. Process. Vol. ASSP-24, No. 5, pp. 399-418, Oct. 1976.

79. Knorr, S.G. "Reliable voiced/unvoiced decision". IEEE Trans. on Acoust. Speech Sign. Process. Vol. ASSP-27, No. 3, pp. 263-267, 1979.

80. Gethöffer, H. "Transform Coding of Speech using Walsh functions". Proc. Symp. on Theory and Applications of Walsh functions, Naval Res. Lab., Wash. D.C., pp. 194-201, 1973.

81. IEEE, "IEEE Recommended Practice for Speech Quality Measurements". IEEE Trans. on Audio and Electroacoustics. September 1969.

82. Smith, C.P. "Perception of Vocoder Speech Processed by Pattern Matching". J. Acoust. Soc. Amer., Vol. 46, No. 6(2), 1969.

83. Mattingly, I.G., Liberman, A.M., Syrdal, A.K., Halwes, T. "Discrimination in Speech and non-Speech Modes". Cognitive Psychology 2, pp. 131-157, 1971.

84. Johnson, N.L. , Leone,F.C. "Statistics and experimental design in engineering and the physical sciences". Vol. I and II, Wiley & Sons, 1977.

85. Freund, J.E. "Mathematical Statistics" Prentice-Hall, N.J., 1965.

86. Cappellini, V., Constantinides, A.G. Emiliani, P. "Digital Filters and their Applications". Academic Press, 1978.

87. Bonnerot, G., Bellanger, M. "Odd-time odd-frequency Discrete Fourier Transform for Symmetric Real-Valued Series". IEEE Proc., Vol. 64, No. 3, pp. 392-393, March 1976.

88. Flanagan, J.L. Rabiner, L.R. "Speech Synthesis", Dowden, Hutchinson and Ross, Inc., 1973.

89. Ahmadi, S., Constantinides, A.G. "Formant Sequency Approach for Low Bit Rate Speech Transmission". Proc. Int. Conf. on Signal Processing, pp. 481-489, Florence, Italy, September 1978.

90. Ahmadi, S., Constantinides, A.G. "Low bit rate digital speech transmission based on formant sequency correlation". Institute of Acoustics meeting on speech transmission, Royal Military College of Science, Dec. 1978.

91. Ahmadi, S., Constantinides, A.G. "Linear Prediction of formants for low bit rate digital speech transmission".
Proc. Int. Conf. on Acoustics, Speech and Signal Processing, Wash. D.C., pp. 48-51, April 1979.

92. Ahmadi, S., Constantinides, A.G. "Low bit rate digital speech transmission based on vowel categorization".
Proc. 22nd Midwest Symposium on Circuits and Systems, Philadelphia, pp. 324-328, June 1979.