SPEAKER IDENTIFICATION USING MODELS FOR PHONEMES

A THESIS SUBMITTED BY
BABU ANTO P.
IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
UNDER THE
FACULTY OF TECHNOLOGY

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

1990

Dedicated to
my fellow citizens

CERTIFICATE

This is to certify that this thesis entitled Speaker Identification Using Models for Phonemes is a bona fide record of the research work carried out by Mr. Babu Anto P. under my supervision in the Department of Electronics, Cochin University of Science and Technology. The results embodied in this thesis or part of it have not been presented for any other degree.

Dr. C. S. Sridhar

(Supervising Teacher)
Professor
Department of Electronics
Cochin University of Science
and Technology.

Cochin - 22.
October 1990.

DECLARATION

I hereby declare that this thesis entitled Speaker Identification Using Models for Phonemes is a bona fide record of the research work done by me under the supervision of Dr. C. S. Sridhar in the Department of Electronics, Cochin University of Science and Technology and that no part thereof has been presented for any other degree.

Cochin - 22.
October 1990.

Babu Anto P.

ACKNOWLEDGEMENT

I am extremely grateful to my supervisor Professor C.S.Sridhar, Dean, Faculty of Technology, Cochin University of Science and Technology for his guidance and support throughout this work. His professional assistance, as well as his friendship and understanding, have made the completion of this work possible.

I wish to thank Professor K.G.Nair, Head of the Department of Electronics, for his timely help and inspiration. I would also like to acknowledge the encouragement and genuine interest shown by Professor K.G.Balakrishnan in this work.

Professor V.V.S.Sarma, School of Automation, IISc., Bangalore, Prof.B.Yegnanarayana, Computer Science & Engineering, IIT Madras, Dr.S.S.Agrawal, Scientist E, CEERI Delhi and Dr.K.K.Paliwal, Computer System and Communication Group, TIFR Bombay were very helpful to me by way of fruitful discussions and kind advices in the proper direction during the tenure of this work. I place on record my deep and sincere gratitude to each one of them.

Thanks are due to Mr.N.K.Narayanan, Lecturer, Govt. Brennen College, Tellicherry and Mrs.Tessamma Thomas, Lecturer, Department of Electronics for the effective discussions and fruitful association in this project. I wish to acknowledge the encouragement and help received from Dr.D.Devarajan, Scientist D, NPOL Cochin. I extend my sincere appreciation to Mr.I.G.Rameshkumar and Mr.U.Santhoshkumar with whom I could associate for some part of this work.

I am obliged to all faculty members, laboratory and nonteaching staff and research scholars of the Department of Electronics for their whole-hearted co-operation throughout the period of my research work. I wish to specially acknowledge the helps extended by Dr.C.K.Aanandan and Dr.K.A.Jose, Lecturers, Mr.K.K.Narayanan, Senior Research Fellow, Mr.James Kurian, Research Associate, Department of Electronics and Mr.P.Ramakrishnan, Lecturer, Govt. College, Madapally.

I also appreciate the co-operation and helpful attitude of Mr.Ajaikumar, Junior Research Fellow, Mr.Ajaikrishnan, Research Assistant and Mr.Jose Kuruvilla, Teacher Fellow, Department of Electronics.

In the course of research work I was enjoying Senior Research Fellowship from CSIR. During the earlier period of this work, I was enjoying research fellowship from University of Cochin. I acknowledge with thanks the financial support provided by these agencies.

I have been working as Project Associate in the Centre for Research in Microprocessor Applications, supported by Ministry of Human Resource Development, Govt. of India and as Research Associate in the project "Software Development for Echo Simulation and Processing", supported by Department of Electronics, Govt. of India. I am extremely thankful to these agencies for the financial support and the facilities provided for this research work.

My sincere thanks are due to Mr.C.B.Muraleedharan for talented technical assistance and co-operation for this work. I am also indebted to Mr.K.P.Sasidharan and Mr.V.M.Peter for bringing this thesis in the present form.

Finally, I wish to acknowledge the patience and love of my mother and wife during the period of this research work without which I could not have completed this work. I am also grateful to all who have great concern for me.

ABSTRACT

Motivation for Speaker recognition work is presented in the first part of the thesis. An exhaustive survey of past work in this field is also presented. A low cost system not including complex computation has been chosen for implementation. Towards achieving this a PC based system is designed and developed. A front end analog to digital convertor (12 bit) is built and interfaced to a PC. Software to control the ADC and to perform various analytical functions including feature vector evaluation is developed. It is shown that a fixed set of phrases incorporating evenly balanced phonemes is aptly suited for the speaker recognition work at hand. A set of phrases are chosen for recognition. Two new methods are adopted for the feature evaluation. Some new measurements involving a symmetry check method for pitch period detection and ACF are used as featured.

Arguments are provided to show the need for a new model for speech production. Starting from heuristic, a knowledge based (KB) speech production model is presented. In this model, a KB provides impulses to a voice producing mechanism and constant correction is applied via a feedback path. It is this correction that differs from speaker to speaker. Methods of defining measurable parameters for use as features are described. Algorithms for speaker recognition are developed and implemented. Two methods are presented. The first is based on the model postulated. Here the entropy on the utterance of a phoneme is evaluated. The transitions of voiced regions are used as speaker dependent features. The second method presented uses features found in other works, but evaluated differently. A knock-out scheme is used to provide the weightage values for the selection of features.

Results of implementation are presented which show on an average of 80% recognition. It is also shown that if there are long gaps between sessions, the performance deteriorates and is speaker dependent. Cross recognition percentages are also presented and this in the worst case rises to 30% while the best case is 0%.

Suggestions for further work are given in the concluding chapter.

CONTENTS

Page

CHAPTER 1

INTRODUCTION

Chapter 1

INTRODUCTION

1.1 MOTIVATION FOR SPEAKER RECOGNITION

One of the important areas in man-machine communication is Automatic Speaker Recognition (ASR). The main motivation behind the recent interest is for dispensing with external artifacts like keys, badges and magnetic cards for personal identification and replacing them with the human intrinsic characteristics such as voice. The speaker recognition has many other applications where unique and reliable personal characteristics are required. In the case of external objects like keys, magnetic cards, badges, passwords etc., the possibility of missing, or forgetting exists. Since voice and speech are intrinsic characteristics and not easy to duplicate, the reliability of identification is very high. The possibility of misusing the artifacts can be eliminated with this method. Recognition through telephone provides opportunities for a wide variety of applications. Transaction through telephone, access to computer data banks etc., are some of these applications. In forensic applications, where the only clue is a recorded speech or a telephone call, Automatic Speaker Recognition is of great help. In all these applications reliability is the main factor and the degree of reliability varies in each application.

1.2 POSSIBILITY OF RECOGNISING PEOPLE FROM THEIR VOICES

It is a well accepted truth that the voice of a speaker is unique and different from that of others, and one can recognise a familiar speaker even over a telephone without seeing him. This fact suggests that the speech of each person contains unique information which contributes to his personal identity. Speech is a product of many transformations occurring at different levels. It starts with the semantic level and proceeds to the linguistics, articulatory and acoustics level where the result is an output speech. Two main reasons can be attributed for this variability among speakers. One is the anatomical/physiological difference in the vocal tract, which is the speech producing mechanism. The second reason is the difference in styles and habits of speech which the speaker develops himself. The physiological difference is a fixed difference due to the shape and size of the vocal tract and this cannot be changed at any level. Individual methods of learning to use vocal tract organs determines the second reason viz., habit and style. The variation between the speakers is known as inter-speaker variation.

Besides the same utterance made by the same speaker on different occasions does not sound identical. This variation for the same speaker is known as intra-speaker variation. The intra speaker variation is due to different factors such as the speaking rate, health of the speaker, the emotional state of the spea etc.

In ordinary speech, in addition to the semantic contents, the information regarding the identity of the speaker and various other factors are included. The fact that human beings can isolate the required information from speech, suggests the seperability of this information contents. One main difficulty in this area is that the information regarding the speaker is only secondary in the speech signals. While choosing the speaker-dependent features, care is to be taken to select features with low intra-speaker and high intra-speaker variability. Since many of the speech-related problems are handled by computers, the speaker recognition also can be handled by computer.

1.3 SCHEME FOR RECOGNITION

The major difference between speech recognition and speaker recognition is that in speech recognition the semantic content is searched, while in speaker recognition speaker dependent information is searched. A general scheme for recognition is shown in Fig.1.1 According to this scheme, features are extracted from the incoming speech and a pattern known as 'Template' is generated. These features are the measures of variability among speakers. Reference templates are created during training sessions and stored in the system. A similar template is created during test phase. This is known as Token/Test Template and using this a pattern matching is performed. The appropriate decision is taken by a 'Decision Logic' depending on the distances between the patterns.

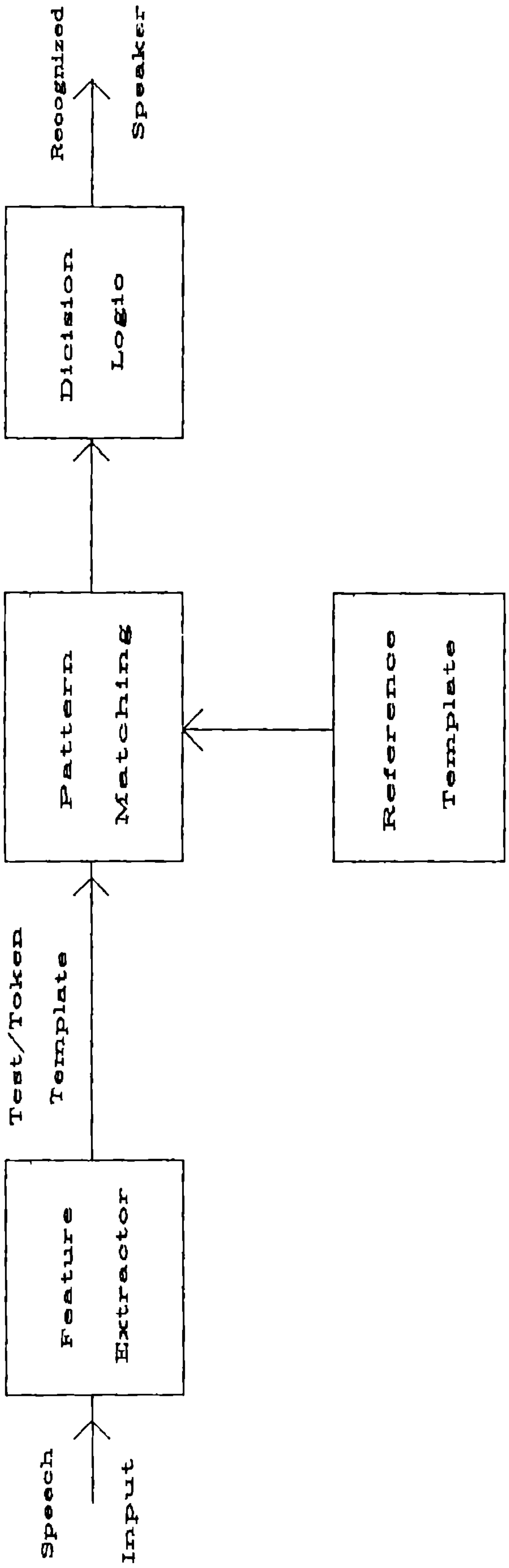


Fig.1.1 General Scheme for Speaker Recognition.

1.4 SPEAKER RECOGNITION TASKS

There are two distinct sub-areas of speaker recognition, namely speaker verification and speaker identification. Comparatively simpler is the speaker verification, in which the test pattern is compared with the claimed reference pattern and a binary decision, whether to accept or reject the claim, is taken. In speaker identification, the test pattern has to be matched with all the reference patterns available in the system and the decision about which speaker it belongs to, has to be taken. Though the method is similar in these two tasks, the number of comparisons and the decision logic are different. Since the verification task is a binary problem, two types of errors are possible in it; Imposter acceptance and the Customer rejection. But the error rate in speaker identification task depends on the population size of users and the performance falls when the population size is very large. The speaker verification is independent of the population size, because it compares only with the claimed reference [124].

Two types of tasks are possible for speaker identification; 'Open-set' identification and 'Close-set' identification. The most general one is the Close-set indentification, in which the test pattern belongs to any one of the reference patterns.

But in the Open-set identification, the test pattern may not belong to any of the reference patterns. In this case, the system has to take N+1 decisions, where N is the number of reference patterns and the additional decision-option is to declare that the token belongs to none of the reference templates. Here both identification and verification are involved.

If a good degree of control can be maintained, in which case the unknown speaker wishes to be identified and is co-operative, a fixed text or text-dependent recognition can be implemented. In this case, the same text is used for the reference and test templates. But in a situation where such control cannot be maintained and if the speaker is not co-operative, a free text or text-independent approach is adopted. With this approach, the reference and test templates are created from independent spoken texts. Depending on the nature of the application, the degree of control over the situation is varied.

1.5 SPEAKER RECOGNITION METHODS

There are three different speaker recognition methods. The earlier works on speaker recognition were mainly based on human listening. In this method, trained people are made to listen speech samples and asked to recognize people. One main interest at that time was to know how human beings recognized speakers [27]. In these studies, performance of different variables were investigated and tried to understand perceptual basis of speaker recognition. This method was used for certain forensic applications. The major difference between human listeners and machines is the ability of human being to recognize speakers even without comparing two similar speech samples. They can recognize from acquaintance with the speaker or from different reference voices of the speaker [1]. The reason for this is that the human brain is matching high level information such as speaker dialect, style of speech, verbal mannerism special type of laught etc., which are not possible to use in machines due to the difficulty in measuring and quantifying these features [124].

Speaker recognition by visual examination is another method. Spectrogram or 'Voice Prints' are used for comparing and recognizing speakers. This gives a three dimensional information about time, frequency and intensity. Various features such as segment duration, formant frequencies, pitch, formant amplitude etc., can be extracted from the voice print. But all these need human expertise and the performance mainly depends on the person who examines the spectrogram [9]. It was also proved that emotional state of speakers and the voice in disguise etc., have great impact on the spectrogram [8]. The performance of spectrographic methods was found to be inferior to human listening methods [124].

Introduction of computers had great impact on speaker recognition methods. Attention was mainly given to automatic recognition of speaker (ASR) using computers. Availability of many faster digital techniques have helped in this venture. Attempts are made to duplicate the human performance using computers. Since one does not know what exact feature human being uses for recognition, more and more techniques are tried in different applications.

1.6 OVERVIEW OF THE WORK

In chapter 2, a brief review of the previous works in the area of speaker recognition is presented. Different methods including listening, spectrographic and computer methods are reviewed in this chapter. A wide variety of features and techniques are available. This chapter highlights some of the important results and recognition accuracies achieved by these techniques.

One of the important hardware requirements of any speaker recognition system is a speech digitizer. The design and development of a 12 bit speech digitizer interfaced to a personal computer is discussed in chapter 3. This chapter explains the hardware consisting of filters which band limits the speech signal at 4 KHz and Sample and Hold, ADC circuits and DAC circuits. It also explains how sampling frequency is selected by software and other software features. Many features available in PC are made use in the design of the system at hardware and software levels. The software for the system is developed in a mixture of high level and machine level programs.

Chapter 4 discusses the advantage of using fixed text approach for speaker recognition and its limitations. Selection criteria of the phrase set is explained here. A pitch consistency study conducted in this regard has been discussed in this chapter. In the light of these criteria and results, the phrase set selected and its modification due to some practical difficulties are explained in detail.

Feature selection and extraction for the speaker identification is explained in chapter 5. In the first part, some standard criteria for choosing features are examined. Considering the feasibility of features used in many other successful methods, a set of features is selected in the present work. They are energy, auto correlation, zerocrossings, pitch by symmetry check method and entropies from time measurements. The first three features are proved to be successful in many earlier methods and found to be very much speaker dependent. The fourth feature, viz., pitch by symmetry is a new method developed for pitch determination. Towards the latter part of the chapter the extraction of these features from the speech signal is explained. The determination of pitch by symmetry check is explained in detail and some important results are presented.

A model for speech production has been suggested in chapter 6. The necessity for a knowledge base is established here. It is established that the speaker variations are due to the difference in the learning process, knowledge base and the feedback system. The variation among speakers are considered to be the error due to approximations and dynamic correction by feedback network. The chapter also explains how these factors can be related to the measurable parameters. The possibility of choosing the speaker dependent parameters is described. The a priori knowledge while using a fixed text approach improves the recognition accuracy.

Chapter 7 explains the algorithm for speaker identification. Two approaches have been employed for identification. In the first approach, a Fixed Text Phoneme Model (FTPM) is used. Probability of occurrence of the selected phonemes are determined from the text of the speech. Using these values, entropy is determined for each phrase. Thus for a phrase set consisting of four phrases, a Phoneme Entropy Vector (PEV) is obtained which is considered to be standard for all speakers. From the actual speech, phoneme time and transition times of each selected phoneme are detected using an automatic segmentation algorithm. Entropies are calculated from these timing data in a similar manner. This entropy vector corresponding to the feature measurements are known as Feature Entropy Vector (FEV). The difference between the PEV and FEV of each speaker is taken as the distortion vector or error for that speaker and is stored as reference template. These templates are used for identification after comparison with a test template.

In the second approach, a similarity measurement technique is employed. In this method, measurements corresponding to various selected features are used. A matrix is formed using these feature measurements corresponding to the selected phonemes of the spoken text. This feature matrix is stored as reference template for each speaker. During identification, the test template, which is also a feature matrix, is compared with the reference templates. Each element in the test matrix is compared with the corresponding element in the reference matrix and an appropriate weightage is assigned, if it falls within a range. The sum total of these weightages is taken as the similarity measure between the reference template and test template. The speaker, whose reference template shows a maximum similarity measure, is identified as the true speaker. The selection of feature measurements based on a knock-out strategy is also presented in this chapter. The measures which contribute more to the performance of the system are finally chosen.

In chapter 8, the computer implementation of the algorithms is explained. The experiment and the working of the system is discussed in detail. The automatic segmentation algorithm based on the energy envelope is also presented here. The results of the above two algorithms are presented with the help of confusion matrices. The performance of each algorithm is evaluated based on these confusion matrices and error rates. It is observed that the FTPM method is inferior to the similarity measure method in performance. The FTPM method shows an identification accuracy of 53.33% in total sessions while the similarity measure method shows 65% accuracy. When the test template and reference template are considered from the same session, the performance in FTPM improved hardly where as the similarity measure shows a very good performance. An accuracy of 93% is obtained in separate session tests for similarity measure method. Different combinations of the feature sets with different weightages are also tried in the similarity measure approach. A feature set showing maximum performance is selected. This chapter also analyses the individual error rates from various tests and misclassification among speakers.

In the last chapter, various conclusions drawn from the experiment are presented. Scope for future work in this field is also discussed.

The algorithm of symmetry check method for pitch detection is explained in Appendix A. The results compared with cepstrum method is presented in this section. A digit recognition system using energy envelope and zerocrossing rate is discussed in Appendix B. In this appendix the algorithm is described in detail with a tree classifying approach. The confusion matrix obtained using this system is also presented here.

CHAPTER 2

REVIEW OF THE PAST WORK IN THE FIELD

Chapter 2

REVIEW OF THE PAST WORK IN THE FIELD

In this chapter, previous works carried out in the field of Speaker Recognition are briefly reviewed. It reviews various experiments and tries to highlight the important features used with their results in various approaches and recognition methods.

J.Pollack et al. [1] have examined the effect of several factors on voice identification. The factors considered for their study were the size of the class of possible voices, the duration of the speech signal, the frequency range of the speech signal, voicing and non-voicing speech characteristics and simultaneous presentation of several voices etc. Duration of the speech signal was found to be the most efficient factor of all these factors.

S.Pruzansky [2] used a pattern matching procedure for automatic recognition of talkers to study the effect of variation in pattern on recognition performance. A time-frequency-energy pattern was generated from common word utterance of ten speakers. Cross correlation of the reference patterns with the test pattern was performed and pattern with highest correlation was chosen. Recognition score of this study was 89%. This study has also considered two dimensional patterns of time-energy and score was found to be very low. But spectral information alone gave the same result as the three dimensional pattern.

A real time speaker verification technique using computers have been presented by R.C.Lummis [3]. In this method utterances were represented by their pitch, amplitude, and formant frequency profiles and these functions were smoothed by low-pass filtering. Dissimilarity measures were then computed. Eight true speakers and 32 casual imposters were included in this experimental study. The acceptance-rejection criterion was adjusted a posterior i for equal rates of false acceptance and false rejection. An average error rate of 1% was obtained for the speakers.

A.E.Rosenberg [4] has presented a listener performance study for speaker verification task. In his subjective testing 10 listeners participated and each was presented withd a paired comparison task between challenge and reference utterances. The reference utterance was from a true speaker while the challenge utterance was either from a true speaker or an imposter with equal likelihood. The overall average error was 4.2% for false acceptance and false rejection, while the best false acceptance rate was 1.6% and best false rejection rate was 0.5%.

S.K.Das and W.H.Mohn [5] have reported an experiment in automatic speaker verification using adaptive pattern recognition. One hundred and eighteen speakers and seven thousand phrases were used for the study. An average misclassification rate of one per cent with a "no decision" rate of ten per cent was obtained in this study.

Another reported work on implementation of an on-line speaker verification system by R.C.Lummis [6] converts the utterances to pitch and gain contours and comparison was made using automatic temporal registration. Updating capability was reported to be a novel aspect in this work along with graphic display.

A text-independent speaker recognition system was described by B.S.Atal [7]. The experiment included 10 female speakers and 12 predictor coefficients were extracted from the 50 msec. duration speech. The extracted features were represented as a vector and new sets of co-ordinates which minimized the intra speaker variances were determined by linear transformations of original vector space. Correlation measures were taken over several segments and average largest correlation was considered for recognition. An overall identification accuracy of 93% was reported in this study.

C.E.Williams and K.N.Stevens [8] tried to extract the parameters that reflect the emotional state of a speaker from a speech signal. They have also investigated the changes that happen in the speech signal of a speaker in different situations. A comparison was also made with real-life situations simulated by an actor. It was found that anger, fear and sad situations produce characteristic difference in contour of fundamental frequency, average speech spectrum, temporal characteristics, precision of articulation and wave form regularity of successive glottal pulses. It was also established that attributes for a given emotional situation are not consistent from one speaker to another.

Voice identification based on visual inspection of spectrograms was performed by O.Tosi et al. [9]. Experimental trials of this study correlated with forensic models have yielded an error of approximately 6% false identification and approximately 13% false elimination.

J.J.Wolf [10] has conducted a study to find out the efficient acoustic parameters for speaker recognition. He found that fundamental frequency, features of vowel and nasal consonant spectra, glottal source spectrum slope, word duration and word onset time are very useful parameters for speaker recognition. A simple linear classification method was employed in this study. In the experiment using 17 such parameters, no speaker identification error was found for a set of 21 adult male speakers and an error rate of 2% was found for speaker verification.

Intensively trained professional mimics were tested with an automatic speaker verification method by R.C.Lummis and A.E.Rosenberg [11]. A computer was made use of to compare the mimics and the customer utterances. The features used in this system were pitch, level and first and third formant frequencies. They have reported that 27% of the best utterances of the best mimics were accepted corresponding to a false acceptance rate of 1.2% for non-nimicking impostors.

G.D.Hair and T.W.Rekieta [12] had experimentally tried speaker verification using phoneme spectra. In this study, phonemes were automatically edited and power density spectra was computed over 7.5 KHz bandwidth. Phoneme spectra for five repetitions were averaged to produce speaker standard pattern vector. In these types of pattern vectors, the error rates were found to be very low using a simple hypersphere decision rule.

D.Meltzer [13] considered formant values for men, women and children, combined with the fundamental frequency and then presented to trained listeners. A respective score of 90.03% was obtained in this listening experiment.

Listener performance in children had been studied by S.Cort and T.Murry [14]. The study indicated that the repertoire rather than the actual duration of the sample provids cues to the children.

T.W.Rekieta and G.D.Hair [15] have tested mimic resistance of the speaker identification system using phoneme spectra. It was found in the study that mimic even after familiarization with the user voice, was unsuccessful in his attempt.

An evaluation study was conducted by A.E.Rosenberg [16] for the listener performance. Thirtytwo casual imposters and four professional mimics were employed with eight true speakers. The test pattern was from an impostor or a mimic against the reference pattern of a true speaker. (The study showed accepted results).

A statistical model based, decision-theoretic approach to speaker verification was reported by N.S.Jayant [17]. In this model, probability density function of a measurement was compared with the PDF measurement of the true pattern and accept, reject and refuse action decisions were taken.

An automatic speaker recognition method, using temporal variation of pitch in speech as speaker-identifying characteristic, is described by B.S.Atal [18]. The 20-dimensional vector representing the pitch contours were linearly transformed to maximize the inter speaker to intra speaker variance. A Euclidean distance measure was used for the decision. A 97% of identification accuracy was claimed with this method.

R.C.Lummis has reported a method for speaker verification by computers using speech intensity for temporal registration [19]. The features used in this method were voice pitch, low-frequency intensity and three lowest formant frequencies; all represented as functions of time. Before comparison of test and reference patterns, the time dimension of the test utterance is warped to optimally register it's intensity pattern onto the reference intensity pattern. This verification was conducted for moderate size of speakers and an error rate of less than 1% was obtained for verification solely based on voice pitch and intensity.

Linear Prediction characteristics were used for automatic speaker identification and verification by B.S.Atal [20]. The linear prediction coefficients, impulse response function, the auto-correlation function, the log area function and the cepstrum functions were used as inputs of an automatic speaker recognition system. Six repetitions of 10 speakers were used. In all these parameters, cepstrum was found to be the most effective. An identification accuracy of 70% for 50 msec. speech and 98% for 0.5 sec. was reported. Verification accuracy of 83% for 50 msec. of speech and 98% for 1 sec. of speech was also claimed in this paper. For a text-independent speaker identification experiment, 93% of accuracy with a 2 sec. of speech was achieved.

An approach based on statistical properties of nasal spectra and study on nasal coarticulation indicated [21] that coarticulation between [m] and [v] have strong speaker dependence and this could be used for speaker identification.

Automatic speaker verification system which includes Linear Predictive parameters with the pitch and intensity analysis of sentence long utterances was presented by A.E.Rosenberg and M.R.Sambur [22]. A method for selecting optimum speaker dependent features have also been discussed in this paper. A verification error of 1% with casual imposters and 4% with well trained mimics were reported in this work.

A study to examine the most effective features for speaker identification by M.R.Sambur [23] revealed that, second resonance (around 1000 Hz) in /n/, third or fourth resonance

(1700-2000 Hz) in /m/, the values of the second, third and fourth formant frequencies in vowels and the average fundamental frequency of the speaker are the most important speaker dependent features. A probability of error criterion was used to determine the relative merits of the features. An identification experiment with 11 speakers using the best five features was also presented by him. One error was observed in 320 separate identification experiments.

J.E.Paul et al. [24] reported an analytical study for semi automatic speaker identification system. Study of discriminatory power of individual phonetic events along with the study of co articulation effects on some events was conducted in this work. Weighted Euclidean distance measure was used for measuring speaker similarity within a phonetic category. Desensitized Fisher discriminant was used to convert the individual distance measures into an overall measure of similarity between the speakers of two spoken utterances.

Speech amplitude, low pass (below 1 KHz) and high pass (above 1 KHz) zero crossing rates were used as weighted sum of orthonormal functions and these weighting coefficients were used to identify unknown speakers by D.A.Wasson and R.W.Donaldson [25]. A recognition rate of 96.6% was obtained with a speaker population size of 10 and using the same data for training and testing.

A long-term avaraged speech spectrum of short sentences was used for feature parameter representation by S.Furui et al. [26]. Though variation was observed in pattern for individual talkers, by defining a suitable measure of the multidimensional distance, it was possible to recognize talker with unknown samples of 3 months or longer.

An overall review of the automatic speaker recognition was presented by B.S.Atal [27]. It discusses the speaker dependent properties of speech signal, methods of selecting efficient set of speech measurement and results of experimental studies of various methods. It also compares the performance of human listener and automatic method using computers. This paper deals with text-independent and text-dependent speaker recognition techniques.

Another review on speaker verification was presented by A.E.Rosenberg [28]. It discusses techniques, evaluation and implementation of various proposed speaker recognition system.

A.Furui and F.Itakura [29] have presented a method of speech wave analysis by Partial Auto Correlation (PARCOR) coefficients and fundamental frequency. Statistical measures like Averaged value, standard deviation and correlation coefficients between parameters in voiced portion etc., were derived and used for the experiment. The experiment indicated that this statistical features are useful in discriminating speakers and also long term variation and combination of words were studied with these features.

G.R.Doddington [30] has developed an entry control system using voice verification. The system was a text dependent one and chooses word from a set of 16 monosyllabic words. The impostor rejection rate of the system was 99.3% and user acceptance rate is 99.9% with a verification time of 5.6 sec.

An evaluation study of an automatic speaker verification system over telephone lines has been performed by R.A.Rosenberg [31]. The experiment included 100 male and female speakers and acoustic analysis of a fixed sentence long utterance was conducted. Computation in the system was done off line and updating of analyzed data with accepted utterance was also tried. Error rates of approximately 10% for new customers and 5% for adapted customers were reported.

An approach to speaker recognition using orthogonal linear prediction parameters was proposed by M.R.Sambur [32]. The orthogonal parameters obtained by linear transformation of Linear Prediction parameters were found to be important speaker characterizing features. The experiment conducted with this approach was using the same sentence spoken by 21 male speakers. For identification and verification, the recognition accuracy exceeded 99% with high quality input speech. Telephone speech produced an accuracy above 96% and in text independent speaker identification a 94% of accuracy was obtained with high quality speech.

S.Furui [33] extracted two types of speaker dependent feature parameters. One was a long time average spectrum of a short sentence and the other the statistical parameters derived from PARCOR coefficients and fundamental frequency. He observed that by eliminating vocal chord characteristics stable personal information could be obtained.

Speaker-Speech interaction was investigated by R.L.Kashyap [34]. It considered problem of clustering speaker to aid speaker verification and optimal choice of phonemes for speaker recognition. An experiment for both speech and speaker recognition using the same speech data base was also conducted.

A survey of automatic speaker recognition system was presented by V.V.S.Sarma and B.Yegnanarayana [35]. Along with the review of the state of art, this paper also discusses computer algorithms developed for extraction and evaluation of several features like glottal waves, spectrographic information, linear prediction coefficients, LPC contours etc. The problem of data reduction and computational needs are also discussed in this paper.

Temporal variation of talker dependent features was analyzed by S.Furui [36]. Temporal variations within and between talkers of feature parameters related to exciting source and vocal tract characteristics of speech sound were analyzed. The variation of exciting source was found to be large with longer time interval and small in shorter time intervals. Talker recognition experiments indicated that the removal of exciting source characteristics from original speech improves talker recognition rate.

An automatic speaker recognition system over communication channel was designed by M.J.Hunt et al. [37]. Statistics on fundamental frequency and spectral shape information by real-time cepstrum were used in this study. Fundamental frequency was found to be not degrading over the communication channel. A reasonably good performance was reported in this study.

The structure of modular system for automatic speaker identification and verification for security system and forensic voice identification was described by E.Bunge [38]. The system response time of less than 1 sec. and the error rate of less than 1% were reported.

An investigation on long time feature averaging for speaker recognition was conducted by J.D.Markel et al. [39].

The parameters used for the study were pitch, gain and reflection coefficients. A thorough observation of the between-to-within speaker variance ratio indicated that performance was improved by long time averaging of the parameter set. The second and sixth reflection coefficients averages showed higher variance ratio.

A report on the AUROS (Automatic Recognition of Speakers by computers) project was presented by E.Bunge et al. [40]. It elaborately described the feature extraction, classification methods and finally the experiment and it's results. This paper claimed a false acceptance rate of 0.94% and false rejection rate of 0.87%.

A comparative study of two speaker recognition methods, using Linear Prediction model had been conducted by L.Fasolo and G.A.Mian [41]. The experiment was conducted with 500 phrases of 10 speakers recorded over a period of 3 months.

Techniques for extraction of speaker specific features had been explained by P.Jesorsky et al. [42]. In this paper a new, minimum-maximum-locating method for time normalization is presented. It also discussed the feature extraction techniques and some experimental results.

Another method of text-independent speaker recognition which almost resembles the human perception technique had been developed by L.L.Pfiefer [43]. The vowel sounds were chosen as the basis and vowel pooling was done without vowel recognition. The sequential analysis in a dynamic fashion tested the samples until a required confidence level has been achieved.

J.D.Markel and S.B.Davis [44] have conducted experiments on Text-independent speaker identification with long term averaged features of large data base. It was found that if the averaging interval is increased, the probability of correct identification increases monotonically. For an average interval of 39 seconds a text independent speaker recognition rate of 98.05% was obtained.

A canonical discriminant analysis was applied to some sub-spaces of observation space and personal factor spaces were constructed which were used for text-independent speaker identification by H.Matsumoto et al. [45]. Decision was made by likelihood measure derived from a posterior i probabilities in the factor spaces. An identification accuracy compatible to human listeners has been claimed using 21 dimensional observation vectors obtained from voiced segments of every 40 msec.

H.M.Dante and V.V.S.Sarma [46] developed a sequential identification technique for large number of speakers. Some predetermined stages were included to reduce by classifying the number of speakers and a tree classif er is used for final decision.

An investigation to find out the possibility of application of zero crossing analysis data in speaker identification was carried out by C.Basztura and J.Jurkeiwicz [47]. An experiment with 20 male speakers proved the possibility of tracking the individual characteristics from their voices using the zero crossing analysis of speech.

An entry control system with voice identification with less than 6 seconds using four randomly chosen monosyllabic words amongst sixteen words was designed by G.R.Doddington [48]. A sequential decision strategy was used in the method. The data storage requirement is 9408 bits for the whole 16 words.

The long term zero crossing analysis of speech signal was used for speaker identification by C.Basztura and W.Majewski [49]. This work presented a method for defining minimum length of window, in which the statistical distribution is stationary in successive zero crossing intervals.

Speaker verificatiion by human listeners over several communication systems has been studied by C.A.Mc Gonegal et al. [50]. The study reveals that speaker verification by human listeners cannot be performed as accurately over mixed speech transmission system as over the same transmission system.

Dynamic programming was successfully applied to selection of features in text independent speaker identification by R.S.Cheung and B.A.Eisenstein [51].

T.Tran Cao and F.Spronck [52] have presented a work on the use of measurements of fundamental period and rate of zero crossings in speaker identification. They measured the rate of zero passage of audio signal and the first two derivatives of French non-nasal vowels taking two time windows: the fundamental period and a standard interval of 10 ms.

Dynamic programming procedures to extract the personal characteristics using PARCOR parameters and the fundamental frequency were described by S.Saito and S.Furui [53]. It was found that the rate of similarity from the matched path is close to the diagonal within the speaker whereas it is spread around the diagonal in the case of match between the speakers. The error rate in this method was found to be reduced to about 2%, when tested with two words.

Y.Grenier [54] presented another experiment for speaker identification, where Linear Prediction parameters were found to use co variance matrix. For decision, two methods were used viz., Mahalanobis distance measure and the likelihood ratio.

An elaborate study on speaker verification through telephone suggests that normalization is required for the telephone voice. A.Ichikawa et al. [55], first conducted an experiment on identification through microphone using auto correlation coefficient, linear prediction coefficients, PARCOR parameters and Log area ratio and found that PARCOR parameters give an identification score of 99.8%. When PARCOR parameters were adopted to telephone speech, it was found that the score was as low as 65%. The authors suggested that a technique involving average-self-inverse filtering which normalizes the linear distortion of actual telephone voice and selected band analysis with the use of pitch frequency for normalization of non-linear distortion improves the verification score to 94%.

S.Nakakagawa and T.Sakai [56] reported the results of a statistical analysis on static features of voiced sound spectra and three dimensional representation of consonants on the basis of dynamic features of the spectra. The paper also presented different conclusions on the use of these spectra.

A mini-computer based speaker recognition scheme was presented by V.V.S.Sarma et al. [57]. A simple scheme for recognizing 8 speakers using long term averaged features extracted from short code words was found to give a recognition accuracy of 94%.

An evaluation of the speaker recognition method using LPC parameters in different environment viz., quiet room, dialed up telephone line via direct hook up and suction cup tap were performed by G.A.Mian [58]. Sentences were manually segmented and evaluation was conducted on phoneme, on breath groups and the whole sentence using minimum distance classifier.

M.Shridhar and M.Baraniecki [59] have discussed the problem of noise in the speaker verification system and investigated the possibility of a speaker verification algorithm derived from an orthogonal parameter representation of speech.

A system for on line speaker verification was described by U.Hofker and P.Jesorsky [60]. This paper discusses various considerations for pre-processing, feature extraction and classification to get a better performance.

W.D.Voiers [61] developed a listener method of speaker recognition in which he employed factor analysis to identify the elementary perceptual parameters of individual differences in speech.

H.M.Dante, V.V.S.Sarma and G.R.Dattatreya [62] presented a multistage decision scheme for speaker recognition. In this method one feature was used for decision at each stage and final decision was made when the number of classes becomes less than a predetermined value. The procedure was based on an optimal stochastic control problem and a population of 60 speakers was tested in this method.

A speaker recognition system suitable for forensic application was designed by E.Bunge [63]. He also investigated the influence of telephone transmission, which is common in forensic applications.

Another speaker verification system based on a feature set of selected short-time spectra, long-time spectra, intensity contour and stationary contour was developed by U.Hoefker et al. [64].

A speaker verification with two stage classifier employing only 391 bits of reference storage on a standard magnetic identity card was presented by R.Geppert et al. [65].

R.Dubes et al. [66] used a method based on choral speech for speaker recognition. The authors claimed that when the voice channel is translated into choral speech, many of the factors affecting the system will be averaging out and the system will beome more or less independent of text, language, time and media of recording.

A review on Principles of Automatic Speaker Recognition was given by P.Jesorsky [67]. In this he describes speaker specific features, pre-processing and parameter representation of speech signal, feature extraction methods by statistical analysis, segment analysis and contour analysis and also different classification schemes. Finally he investigates various recognition schemes available.

A scheme for automatic speaker identificaiton for a large population based on multistage, decision tree classifier was proposed by H.M.Dante and V.V.S.Sarma [68]. A large number of classes, after comparing with the given pattern, were rejected in the first stage based on a subset of the features. The final decision was taken using the remaining features after a predetermined number of classifications. In this scheme the authors used a population size of 30.

The importance of mood of the speaker in speaker recognition problem was investigated by A.I.Menkiti [69]. He examined the speaker identification in two methods viz., aural and by analysis for different speaker moods.

The problem of variability of speaker voice with organic change as well as change of time was found to be tackled by applying cluster analysis. M.H.Kuhn [70] has shown that with this technique of excluding extremely untypical samples from the training session, the performance was improved and the error rate of 20% was reduced to 4%.

A system for access control using speaker verification implemented on a mini-computer was developed by M.H.Kuhn [71]. Here he proposed a method of reduction of storage size, so that the pattern can be stored in a magnetic identity card and the verification system will have no restriction on the population size of speakers.

M.Baraniecki and M.Shridhar [72] introduced a new scheme to verify speaker from a speech corrupted by noise with unknown statistics. When preprocessed with an adaptive noise cancelling algorithm, the verification performance was improved from 45% to 95%.

R.E.Wohlford et al. [73] conducted a comparative study of four speaker verification methods. The four methods were: short and long term averages, cepstral measurements of long term spectral averages, orthogonal linear prediction of speech waveforms and long term averages of LPC reflection coefficients combined with pitch and overall power.

S.Furui and A.E.Rosenberg [74] have suggested a new technique where the fixed text utterance were represented by time functions of cepstral coefficients expanded by an orthogonal polynomial. A dynamic programming technique was also used to bring sample utterance representation into time registration with reference pattern. The decision, based on the overall distance, which was again a sum of local distances along the optimal path associated with the time registration.

A real time automatic speaker recognition system was implemented on a mini-computer by C.E.Chafei [75]. The four characteristic features are extracted from the pitch of the speaker. The system worked satisfactorily for 15 speakers with utterances taken over a period of 6 months.

A low cost microprocesssor based speaker verification system was presented by M.H.Khun [76]. The verification procedure involves Bayes classifier with histogram approximation of the probability density. The system requires, 256 bytes of RAM memory and this was tested for 127 speakers. The reference was updated after every successful verification and the verification time was about 8 sec. with 1% false rejection and 2% false acceptance.

H.M.Dante and V.V.S.Sarma [77] have studied speaker verification based on a theoretical model useful for forensic application. Three methods using single feature, multiple independent measurements of single feature and multiple independent features were analysed in this paper.

M.H.Kuhn et al. [78] in another paper have discussed the differences of customer acceptance in different application areas like banking and military environment. He also indicated the possibility of online user evaluation.

A study by H.Ney and M.H.Kuhn [79] revealed that the time frequency matrix obtained from a 13 channel filter bank, normalized with respect to long time averaged spectrum and normalized via dynamic programming could be a very strong feature set for telephone line speaker recognition.

Orthogonal parameters formed from the eigen values and eigen vectors of the covariance matrix of the speech of a speaker were found to be very useful in speaker verification. R.E.Bogner [80] studied this technique and also the factors that effect the distortion over telephone line. The method was based on covariance of logarithmic spectral estimates and the results gave an accuracy of about 95%.

A.E.Rosenberg and K.L.Shilpey [81] proposed a speaker identification and verification system combined with a word recognition system in which a template distance measure was made use of.

The auto correlation function was subjected for study in speaker recognition on telephone lines. H.Ney [82] came to the conclusion that the clipped auto correlation function was superior to simple auto correlation function both in the simplicity of computation and reduced dynamic variability. Dynamic programming was used for time registration and an error rate of around 2% was obtained.

A text-independent real time speaker recognition system was developed by E.H.Wrench Jr. [83]. A test conducted with 30 speakers with 10 seconds of speech gave a recognition accuracy of 93-100%.

A dynamic programming based feature selection was attempted by M.Shridhar et al. [84]. He has also investigated the possibility of the use of orthogonal linear prediction parameters in the text-independent speaker recognition. A verification accuracy of 96.5% was claimed with 8 optimally chosen parameters generated from 100 seconds of time spaced voiced speech for reference pattern and 5 seconds of speech for test pattern.

A composite model of speech for speaker and word recognition was presented by R.J.Fontana and M.S.Fox [85]. The method comprised of estimating the underlying sub-source using data compression technique and the switch sequences were derived from these estimates. These switch sequences were then compared in time domain and decision was made via variation distance.

A fixed text approach for telephone line speaker recognition, based on the cepstral coefficients obtained from LPC analysis of time functions and removing the frequency response distortions introduced from the transmission lines was adopted by S.Furui [86]. Time registration and distance calculation were performed by dynamic programming technique and the decision was made on the overall distance measure. A recognition accuracy less than 1% was obtained in spite of different transmission conditions.

One real time speaker verification system implemented on a mini-computer was introduced by M.De George [87]. The techniques for optimization like word selection, microphone selection, training technique and reference template updating were discussed in this paper. The particular work claimed a recognition accuracy of 98.8% with a speaker base of 27 males and 15 females.

A software developed for interactive text-independent speaker verification was presented by A.Kohen and I.Froind [88]. The different aspects of the problem and the results of a verification system were also discussed.

A comparative study of statistical features and dynamic features were conducted by S.Furui [89]. The speech wave transformed into a set of log area ratios and fundamental frequency were used for the extraction of statistical and dynamic features. In the case of statistical features, the mean value and standard deviation for each time function and correlation matrix between these functions were calculated for voiced regions of speech. Using dynamic features, time registration of time functions was performed. The long term effect of these features were also studied in this work.

H.Ney [90] reported a method for speaker recognition of telephone speech using the intensity contours and fundamental period. The fundamental frequency was measured by estimation procedure and a real time operation was assured by implementing the system on a mini-computer or a microprocessor.

A study on the effects of acoustic features on speaker identification by K.Itoh and S.Saito [91] concluded that frequency spectrum envelope (FSE) is contributing too mcuh to the speech and only by removing FSE the fundamental frequency and tempo become significant. The study was conducted by using synthesized speech signal processed by PARCOR speech analysis-synthesis technique.

S.Furui [92] developed a scheme which employed the statistical feature of cepstrum viz., mean value, standard deviation, fourier coefficients of each parameter and cross correlation between parameters. More than 99.9% of accuracy was obtained using a microphone. False accpetance rate of 2% and false rejection rate of 3% was achieved for an online verification experiment using telephone speech.

Another comparative study on statistical and dynamic features using for text-dependent speaker recognition method presented by S.Furui [93] suggested that the statistical features have advantages in calculation, memory size for feature and recognition. It also attempted to combine these two features and the effectiveness of the spectral equalization technique.

H.Ney and R.Gierioff [94] introduced a feature weighting technique in speaker recognition. In which, weights were predetermined for individual feature components according to the ability to distinguish between classes of speakers. These weights are depending on both time and frequency.

R.Schwartz et al. [95] have investigated the application of probability density estimation to text-independent speaker identification. The study was performed for two parametric and one non-parametric PDF estimation and the performance was found to be better in non-parametric estimation.

A method for text-independent speaker recognition was devised by N.Mohanakrishnan et al. [96]. In this method, they included LPC, reflection cepstrum, log area ratio coefficients, speech power spectrum parameters and inverse filter spectral coefficients, as features and selected any two features for the first stage of recognition. If there was any disagreement in the result, another feature was made use of and the result was found to improve.

A unified scheme for speaker verification was proposed by M.Shridhar et al. [97]. They used an orthogonal linear prediction model for verification and claimed a good result for the test speech recorded in a noisy environment. The test speech was preprocessed using a modified adaptive noise cancellation filter.

H.Ney [98] suggested a speaker recognition method which employs the spectrogram of a sentence. The time synchronization of the test and reference template was done using dynamic programming and the decision was made by the dissimilarity measure after comparison of the two spectrograms.

A text-independent speaker recognition procedure described by M.Shridhar and N.Mohanakrishnan [99] claimed an accuracy greater than 99%. The procedure was implemented at two stages using two different parameter sets and a confirmation was done with a third set of parameter if controversy arose in the first stages.

A comparative study by J.Wolf et al. [100] showed that the probabilistic classifier is much more superior to that of a minimum distance classifier. A data base created from a radio channel speech was tested and compared with the performance of lab quality speech in this sutdy.

Another method based on a statistical model of the speaker's vector quantized speech was introduced by K.P.Li and E.H.Wrench Jr. [101]. The frequency-occurring vectors or characters formed a model of multiple points with n-dimensional speech space instead of the usual single point models. The distance was also measured depending on the statistical distribution of the models. An accuracy as high as 96% was claimed by this particular sutdy.

A comparative study of the distance measures used in the speaker recognition methods was conducted by M.Shridhar et al. [102]. Among the four commonly used methods viz., Mahalanobis distance, maximum a posteriori probability, nearest neighbour criterion and the correlation distance measures, the nearest neighbour criterion with a modification was found to be superior.

From a previous expeirmental analysis on the telephone speech, based on the FO statistics and low order cepstrum coefficient, the results are found to be poor. In a study by M.J.Hunt [103] feature set based on the frequencies of peaks in the short term smoothed spectrum was found to perform better for the telephone speech, mainly because of the greater resistance to the noise and non-linear distortion.

D.Helling and H.W.Strube [104] have described a speaker identification and verification system in real time using signal processors. In this method, a Euclidean distance approach was used for decision and two feature reduction techniques were discussed.

A study to find the best features for consonant and speaker recognition by S.Nakagawa and M.Sakamoto [105] showed that FFT cepstrum and LPC cepstrum are the most suitable parameters. In this work, it was specifically found that 18th FFT cepstrum at an analysis frame length of 25.6 ms. with 5 ms. of shift between frames are most suitable for the speaker recognition.

M.K.Krasner et al. [106] conducted a study on the speaker recognition using a radio channel speech data, which is of poor quality and full of noise. This work suggested ways to extract robust feature sets and a method for modelling and classification.

A performance verification study of the communication system for speaker identification was carried out by P.E.Papamichialis and G.R.Doddington [107]. The listeners were made to identify the utterance comparing with a reference utterance and the experiment was repeated for processed and unprocessed speech, from same and different data base.

H.Hollien [108] suggested that speaker identification can be effectively done using multiple vectors. In his paper he considered four speech vectors viz., Speaker fundamental frequency, long term speech spectra, vowel formant tracking and temporal analysis vector.

In another approach for speaker identification by C.B.A.Shaw Cross et al. [109] two dimensional half-plane lattice parameters of spectrogram were employed. An FFT and Wigner transform were used to process the raw speech data into a two dimensional form suitable for lattice modelling.

A speaker verification system constructed from off-the shelf components consisting of both hardware and software was explained by D.E.Crabbs and D.P.Conard [110]. They described the system, starting from the analog speech acceptance, feature extraction and template creation to the decision classification and also performance verification.

H.Tananka et al. [111] proposed a novel approach for speaker identification based on a new parameter named Time Sequence Matrix. In this method the zero crossing interval was extracted and plotted into a two dimensional 10x10 matrix. The experimental results gave an accuracy of 94% with this method.

An operational evaluation of the speaker verification system was presented by D.E.Crabbs and J.R.Clymer [112]. The optimization of the system performance based on the results of the experiment was also explained by them.

A speaker identification based on known voices familiarity and narrow-band coding was presented by A.S.Nielson and K.R.Stern [113]. A listening test was conducted over an unprocessed and LPC processed channels from the voices of speakers who were familiar with the listeners and it was found that the processed speech was better for identification.

The channel variability problem was analysed for speaker identification by H.Gish et al. [114].They have shown that the channel invariant features can discard more speaker dependent information and found that it was more effective when channel variability was incorporated during training.

Vocoded speech was put into test for speaker recognition by S.S.Everett [115]. Six different voice processors were used in the input side of the speaker recognition system. An accuracy as high as 95% was obtained for different input filter bandwidths.

The use of vector quantization in speaker verification was tried by F.K.Soong et al. [116]. A vector quantization code book was used as an efficient means of characterizing the short term spectral features of a speaker. A data base consisting of isolated digits was used for experiment and 98% accuracy for speaker identification was obtained. The system was also tested for different code book size, number of digits, different recording sessions etc.

Another work on vector quantization was reported by J.T.Buck et al. [117]. For an experiment on 16 speaker population, they obtained a false rejection rate of 0.8% and false acceptance rate of 0.0%.

A speaker verification system for access control discusses the issues of interfacing different microphones and also the environmental variation. This was studied by W.Flix and M.De George [118].

H.Kashiwagi et al. [119] used spectral envelope of the linear prediction residual of speech for speaker identification. The spectral envelope was obtained from the low time portion of the cepstrum of the residual and a Euclidean distance between the envelopes were used for identification. An identification rate above 80% was obtained by this method.

H.Noda [120] developed a speaker verification system for forensic application where concatenation of inventory of recorded syllabic units was tried for fixing the threshold. Here a text independent speaker verification via telephone lines was conducted. It was found that this method is superior to the normal fixed threshold.

A.E.Rosenberg and K.L.Shipley [121] have introduced a talker recognition system in tandem with a talker independent isolated word recognition. A template based approach was adopted for both these recognitions. An experiment with relatively large population gave an identification error of 3.6 to 14% and an error rate of 8% when tested in a speaker verification mode.

Performance of the human speaker recognition of LPC voice processor was studied by Z.Uzdy [122] This study pointed out that for a very good speaker recognition a high-frequency data bandwidth is necessary.

M.Ganesan et al. [123] have demonstrated this speaker identification and verification based on an acoustic data of Hindi vowels using a sound spectrograph. The features extracted from the spectrograph was fed to the computer for creation of templates. A classification was done based on the Euclidean distance between the test and reference templates. The experiment was done by changing the parameters and utterances and results were compared.

A general review of the speaker recognition methods was given by G.R.Doddington [124]. He has discussed the different methods of speaker recognition like listening, visual and computer techniques and the limitations of each method. The paper also deals with text-dependent and text-independent methods and also covers the state of the art.

G.Audiso and A.Caramella [125] have developed a speaker verification algorithm which evaluate the probability density function using a Parzen estimator with a hyperconic kernel. A method for evaluating the radius of the kernel was also given in this paper.

A system using template matching for both text-dependent and text-independent speaker recognition was presented by A.L.Higgins [126]. A 10 sec. speech conversation could be identified without any error using the above template matching method.

A speaker recognition system for the field use (outside) the lab environment), has been explained by M.C.K.Yang et al. [127]. The system makes use of a regression technique for the combination of several features.

Assuming that the recognition of speech is a prerequisite for a good quality speaker recognition, M.De George and

W.Fiex [128] have presented a speaker verification system integrated to an access control environment which utilizes short term spectral and temporal features. A field experiment was also conducted by them using this system.

Cepstral features and energy are extracted in real time and used in the speaker verification system by M.Burnbaum et al. [129]. A dynamic time warping method was used for distance measurement and the reference template was updated after each successful verification. Since the system was used in a dialed telephone line, channel normalization and noise floor were used to reduce the telephone line variation. An equal error rate of 19% was claimed by the authors.

S.Furui [130] has given an overview of the speaker recognition technology and discusses various issues in this particular field. The paper focusses on the effect of long term spectral variability on the recognition accuracy and the ways to reduce this effect.

An optimal decision threshold was determined by N.Fakotakis et al. [131]. This was calculated from the distribution of the intra and inter speaker distances by minimizing the false acceptance and false rejection errors.

An overview of speaker recognition was presented by D.O'Shaughnessy [132]. The paper reviews various recognition techniques, and different distance measures, timing considerations and dynamic time warping. It also discusses template matching technique and use of dynamic and statistical features. The author looks into various techniques like vector quantization, cepstral analysis and use of orthogonal LPC parameters. In general the paper touches upon almost all the aspects of speaker recognition.

A text-dependent speaker verification using vector quantization source coding was presented by D.K.Burton [133]. A source code book was designed to represent a speaker based on his utterance and deviation is measured in this code book. The experiment resulted into a false acceptance rate as low as 0.7% and 0.6% false rejection rate with a speaker population of 16 true speakers and 11 imposters.

A PC based speaker verification, using statistically averaged parameters of speech was proposed by Michailov and D.Milev [134]. This method requires smaller memory capacity and computation power. In this method, the parametric vector representation was transformed into a new one, which reduces the influence of phonetic content and the condition of recording and also the information redundancy of the LPC.

The technique of frequency warping was found to be very effective for speaker verification in noise by H.Noda [135]. In this method the higher energy portion like formant regions were expanded and low energy portion which is affected by the noise was shortened in the spectrum. Then an LPC analysis on this frequency warped spectrum was performed and the auto correlation function derived from this to form an all pole model. The spectral distance measures from this method was found to be very effective in speaker verification especially in noisy environments.

A new method of Circular Hidden Markov Model (CHMM) was applied in speaker identification by Y.C.Zheng and B.Z.Yuan [136]. A distinct reference CHMM was produced for each person using Baum's forward and backward algorithm. Classification was effected depending up on the highest probability and results show 94% speaker recognition accuracy.

A template based approach for speaker verification, using different distance measures was presented by G.Velius [137]. He has also studied the variation by changing the length of window and the parameters chosen and the order of LPC-cepstrum analysis. Euclidean, inverse variance weighting, differential mean weighting, Khan's simplified weighting, Mahalanobis distance, and the Fisher linear discriminant were the distance measures tested. It was found that the performance varies depending upon the vocabulary and an average performance of 5% Equal Error Rate was obtained.

J.Wilbur and F.J.Taylor [138] used a derivative of Wigner Distribution function called smoothed discrete Wigner distribution. This was proved to be significantly more consistent estimate over differing samples of a given word. With this the formant frequencies are well defined and consistent over the samples, which helped in speaker identification.

A Texas Instruments TMS32020 digital signal processor based speaker verification system was designed by J.B.Attili et al. [139]. They have used a 37 dimensional feature vector consisting of 12 PARCOR coefficients, 12 log-area coefficients, 12 LPC cepstrum coefficients and one normalized gain coefficient. An overall error rate of 1.9% in text independent verification and 0.94% in text dependent verification was observed using this hardware based system which carried out the verification in almost real time.

2.1 PRESENT WORK

Text-dependent speaker recognition has shown a better performance compared to the text-independent recognition. The a priori knowledge of the text gives an opportunity of modelling the phonemes. In this present work a model with knowledge base and feed back network is proposed for speech production. Attempts have been made to relate the measurable parameters with this model, since, the actual parameters in the model are not measurable.

Two approaches have been attempted for speaker identification. Both the methods are employing the popular template matching approach. In the first approach, a Fixed Text Phoneme Model has been used. Using entropy from the probability of occurrence of the text and the entropy obtained from the timings of the spoken text, a template is generated for comparison.

In the second approach, a Similarity Measure Method is made use of. The feature measurements from the spoken fixed-text are taken for the creation of the template. A weightage for each measure is assigned when a similarity is observed during comparison. The sum total of the weightages are taken as the similarity measure for each speaker template. A knock-out method is also used for the selection of the feature measurements depending on their performance. A good performance is obtained using a set of final feature measurements.

CHAPTER 3

SYSTEM USED FOR THE EXPERIMENT

Chapter 3

SYSTEM USED FOR THE EXPERIMENT

3.1 INTRODUCTION

Most of the modern signal analysis and processing is performed by digital computers. Using a small general purpose digital computer one can apply a wide variety of digital signal processing techniques to the speech communication problem. The complexity of analog circuit is very high compared to the digital processing. Moreover, the digital data is less prone to noise. In any speech analysis problem, one always faces the non-availability of speech data in a suitable digital form. Usually a large data base is required for speech experiments. Most preferably, the data should be available in the same system.

A good solution for this is, a digitizing system designed to work with the computer, used for the processing. So, the same computer can control the digitizer and get on-line data in the required form and can also store the data in its own memory. Ultimately one can think of a computer system, in which one function key converts analog speech to digital data and another key activates on-line processing and a third key converts the processed data into analog speech and so on.

In an application like speaker identification and verification, user expects a quick response from the system. Considering the huge amount of computational analysis and large number of comparisons involved in the problem, the system used for the purpose must be fast enough. Though there are not many real time identification or verification systems available, one expects a reasonably fast response and result. Availability of faster algorithms for analysis is another major issue. Above all, one has to think of an economic system with all these facilities. Taking all these factors into consideration, a system built around a personal computer (PC) is an optimized solution. It is the most inexpensive computer, with fast processing capabilities. There are many fast software algorithms, specially suited for digital signal processing, available in PC. Besides, the development of hardware on a Personal Computer is easy due to its open architecture.

3.2 DIGITAL CODING OF SPEECH

There are different schemes for representing the digital speech viz., PCM, DM, DPCM and ADPCM. Among all these coding schemes, PCM being direct and simple, is chosen for this present system. A digital representation of analog speech is always discrete in both time and amplitude. The sampling operation makes the analog signal discrete and the quantization in amplitude makes the signal completely digital.

For a faithful representation of speech signal, at least a 12 bit resolution is necessary which gives a reasonably good SNR also. Using a 12 bit digitizer along with a 16 bit computer makes the system faster and gives a good quality representation of speech. In order to check the quality of speech before and after processing, it is desirable to reconstruct the digital data to analog signal. For this, an 8 bit DAC is used with the system. Thus the whole system with a 12 bit ADC, 8 bit DAC and an IBM PC/AT gives the power of a speech workstation.

3.3 SYSTEM OVERVIEW

The block diagram of the system is shown in Fig.3.1. In this speech digitizing system one of the expansion slots in a Personal Computer is used with an interfacing card. The address lines from the computer are decoded in such a way that it avoids clash with standard I/O addresses used in computer. In the digitizing circuit, the microphone input is amplified properly and band limited to 4 KHz by an active low pass filter and fed to the ADC through a Sample and Hold circuit. Since the system was designed for PC and PC/XT, initially the 12 bit data had to be read into the computer through 2 latches and in two cycles to match the 8 data lines of 8088 CPU. The sampling rate of the ADC is controlled by feeding pulses to the Start of Conversion (SOC) pin of ADC. The End of Conversion (EOC) pin

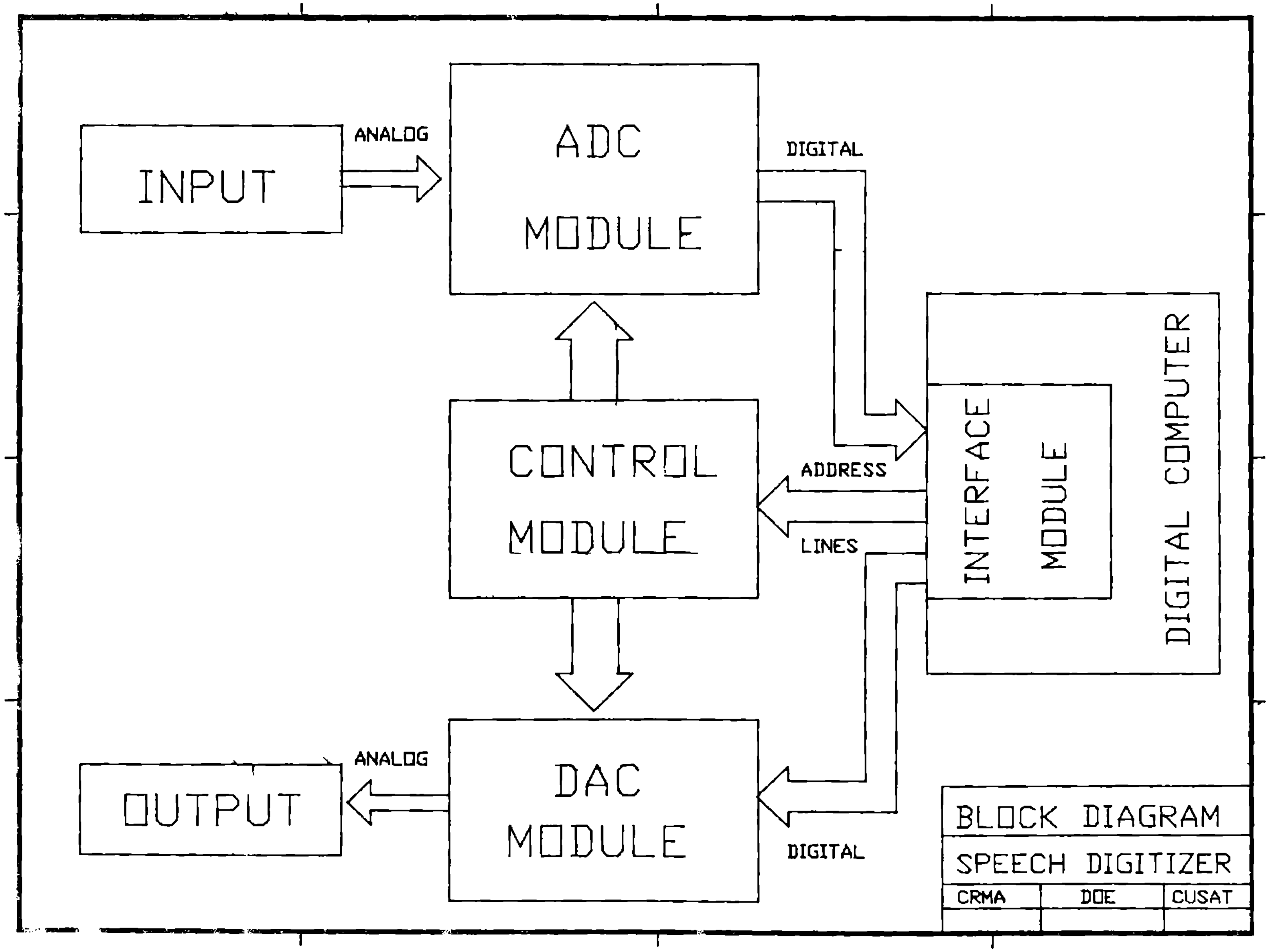


Fig.3.1 Block Diagram of Speech Digitizer.

is connected to the Sample/Hold pin of Sample and Hold circuit, so that the data will not change during conversion. The digitized data is stored in the main memory of the computer in a separate extra segment. This data can be viewed by plotting on the screen or listened back after reconstruction. For the reconstruction of digital data, a DAC is connected through a latch and the output of DAC is lowpass filtered and after proper amplification fed to an output speaker. The selected portion of data can be stored on to the secondary storage unit viz., Floppy Disk or Winchester, for later use of data.

The complete control program for the system was developed in a mixture of high level language and assembler. This gives the user an easier way of programming. One can make use of the memory management, graphic features and other advanced facilities available in the higher level language. At the same time, using the assembler enables, fine control of devices, which is not possible through high level language. Through the software, it is possible to select the sampling rate of data as well as the reproduction rate of output speech after reconstruction. Since 8088 and 80286 processors have segment addressing, the data can be stored in different segment addresses available in PC. Thus a large amount of data can be stored. At a normal sampling rate of 8 KHz, one segment of memory can hold about 4 sec. of speech. Data stored in the memory can be examined by graphically plotting it on the graphic screen of PC, which is easily done through high level language. The system software developed is very powerful, that the system can handle many functions with the help of available hardware and deliver the power of a fairly good speech workstation.

3.4 HARDWARE DESCRIPTION

Only essential hardware is developed for this system. Special care has been taken to make use of all available facility of the PC. So, the hardware developed for this system is sharing most of the resources in the PC. The extra hardware developed is kept outside the PC and interfaced with an interface card plugged in the expansion slot of the PC mother board. On this interface card, we have bidirectional gated tristate buffers for data lines. This precaution has been taken to prevent any clash with the normal operation of the computer. These gated tristate buffers are enabled only in the selected address range by a decoder circuit. The I/O addresses are also chosen with special care to avoid clash with the standard I/O peripherals used with the PC. The range selected for this particular I/O operations is 8300H to 830FH, which is not used by any other devices. All the address lines are buffered and address range selection is done through a small logic circuit. An $\overline{\text{XUSER}}$ signal is generated using logic gates, which will be active only in the selected range of address. The interfacing circuit is shown in Fig.3.2.

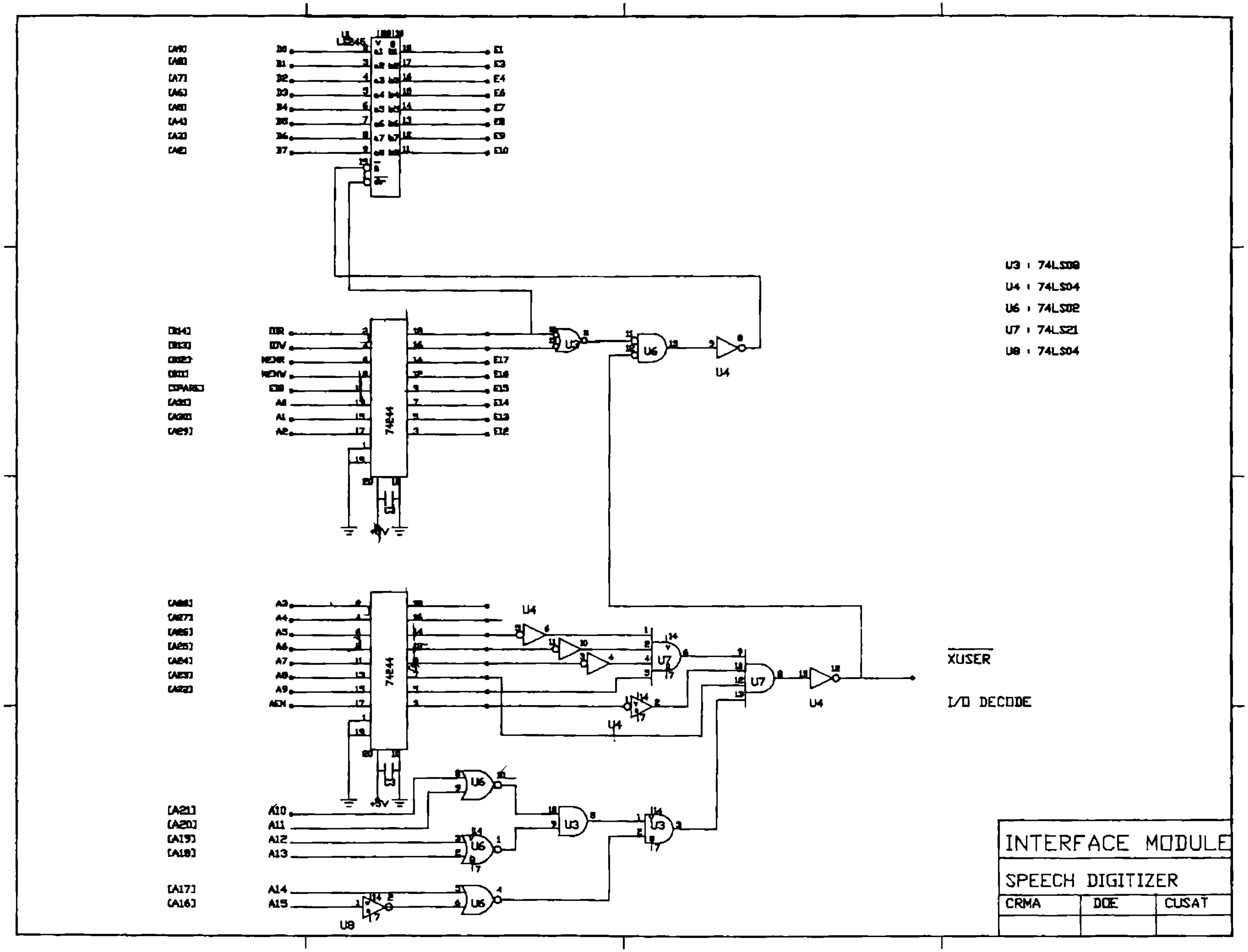


Fig.3.2 Interface Circuit with PC.

The least significant 4 bits of address lines along with the $\overline{\text{XUSER}}$ is decoded using a four-to-sixteen decoder, which generates 16 separate addresses in the range 8300H to 830FH. Then these decoded address lines along with $\overline{\text{IN}}$ and $\overline{\text{OUT}}$ signals from the computer are used to generate separate IN and OUT addresses using an OR gate. Thus the addresses 8300H through 8303H are decoded as input ports and 8304H through 8307H are decoded as output ports. These address lines control and monitor the devices and operations of the system. The decoder and address generation is shown in Fig.3.3.

In the digitizing circuit, the microphone is the front end. The signal from this microphone, which is only of the order of a few milli volts is amplified to appropriate level and fed to the Analog-to-Digital Converter. This ADC is working at bipolar ±10 volts, so that the level of the signal should be amplified using an IC 741 operational amplifier. Two 741 IC's are used for this amplification and a gain control potentiometer is provided in the second stage of the amplifier for proper control of input signal level. A second order Butterworth active low pass filter is designed for bandlimitting this amplified input signal at 4 KHz. This filter circuit is also designed based on an IC 741 op-amp. The band limitted signal, which is the output of the filter circuit is now fed to a Sample and Hold circuit. An LF 398 along with a holding capacitance constitute a S/H circuit.

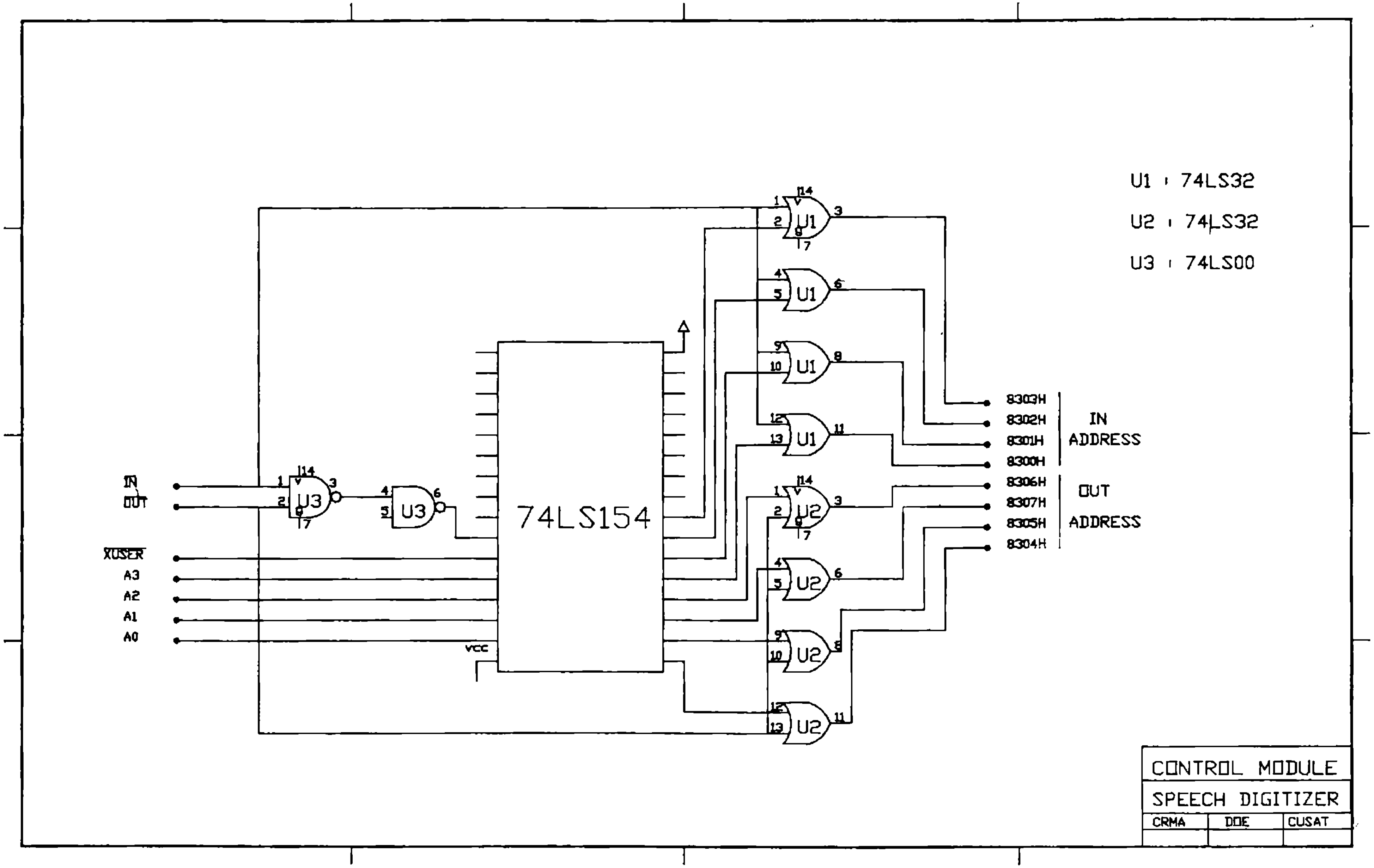


Fig.3.3 Control Circuit - Decoder and Address Generator.

Binary (COB) for bipolar input signal ranges. In this present system the input range is selected as ±5 V and the output mode as Complementary Two's Complement (CTC).

Once the analog signal is digitized by the ADC, it has to be transferred to the computer. The EOC line status is checked by the computer by periodically scanning this line which is connected to a data line through a tristate buffer. Once the EOC signal is issued by the ADC, the processor senses it and then steps are taken to read the converted data into the memory. Since the card was originally designed for PC and PC/XT, the data width is considered to be 8 bit. Hence the 12 bit data from the ADC had to be read into the computer in two cycles. This is done by connecting this output lines from ADC to two latches and the output of the latches are connected to the 8 bit data bus. Two 8282, 8 bit bipolar latches with tristate output buffer are employed for this purpose in the circuit. The two latches are write enabled simultaneously by the EOC signal from ADC and the output data from the latches are enabled separately. The data read from the digitizer is appropriately stored in the main memory of the computer at a pre-specified extra segment. The conversion continues till the entire segment gets filled. Hence at a sampling frequency of 8 KHz, it can hold 4 seconds of 12 bit data. The complete digitizing circuit is shown in Fig.3.4.

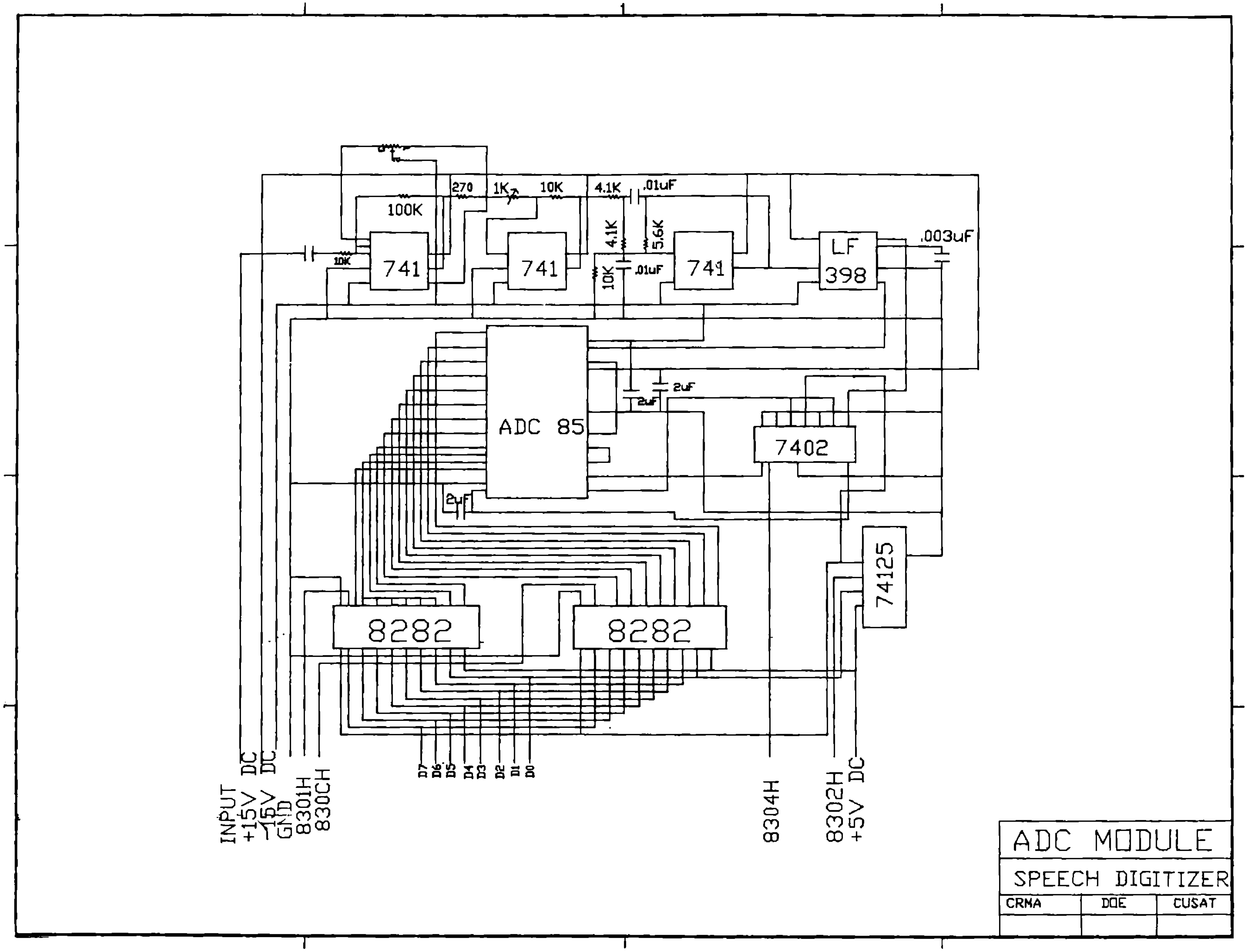


Fig.3.4 Analog-to-Digital Conversion Circuit.

The logic control input of the Sample and Hold is connected to END OF CONVERSION (EOC) pin of the ADC chip, so that the data will be held unchanged during the conversion of the analog signal by the ADC. The output of the Sample and Hold circuit is fed to the analog input pin of the ADC. The ADC used for this is Burr Brown ADC 85, which has 12 bit resolution and very fast conversion capability with internal clock, comparators, reference voltage and input buffer amplifier. The conversion time of this chip is 10 microseconds. The Start of Conversion (SOC) signal fed from the computer through the output address 8304H controls the sampling rate of the signal. This signal is generated by software. When an SOC is fed to ADC, the EOC pin will go low, which holds the signal in the holding capacitance of S/H circuit and at the same time start the conversion. After the conversion of data, the EOC signal goes high and the S/H circuit will be in sampling mode, which will start charging the capacitance. In the ADC 85, various analog input signal ranges are available with ±2.5 V, ±5 V, ±10 V, 0 to 5 V and 0 to 10 V which can be chosen according to the requirement. Facility is provided to trim externally the gain and offset error. The converted data from the chip can be taken out either in serial or parallel form according to the clock and status signal. The output is available in three different selectable binary codes. They are Complementary Straight Binary (CSB) for unipolar input signal range, Complementary Two's Complement (CTC) and Complementary Offset

There are many options for the user with the digitized data. This data can be checked by plotting or listened back by reconstructing or stored in the secondary storage units for further applications. One can also straightaway go for the analysis of speech. For most of these above mentioned processes, PC facilities are exploited. But for the reconstruction of the digitized data a special Digital-to-Analog Conversion circuit is required. This circuit consists of an 8 bit DAC (National DAC 0800), latches, filter and amplifier. The data bus of the computer is connected to the input digital lines of DAC through an 8282 latch, which is enabled when required using the address 8305H. This DAC 0800 is a high speed, monolithic and current output device. Since the output filter quality was not as demanding as the input filter, a simple RC low pass filter is used for the output analog signal from the DAC, which will band limit the signal at 4 KHz. The function of the filter is to smoothen the quantization effect and to move the high frequency noise from the system. A power amplifier, designed with 810 IC is used in the output side of the filter, which properly amplifies the signal and feeds to the speaker. This reconstruction facility will enable the speaker to check a processed data file for quality enhancement/deterioration by listening back in the subjective testing. Fig.3.5 shows this DAC circuit.

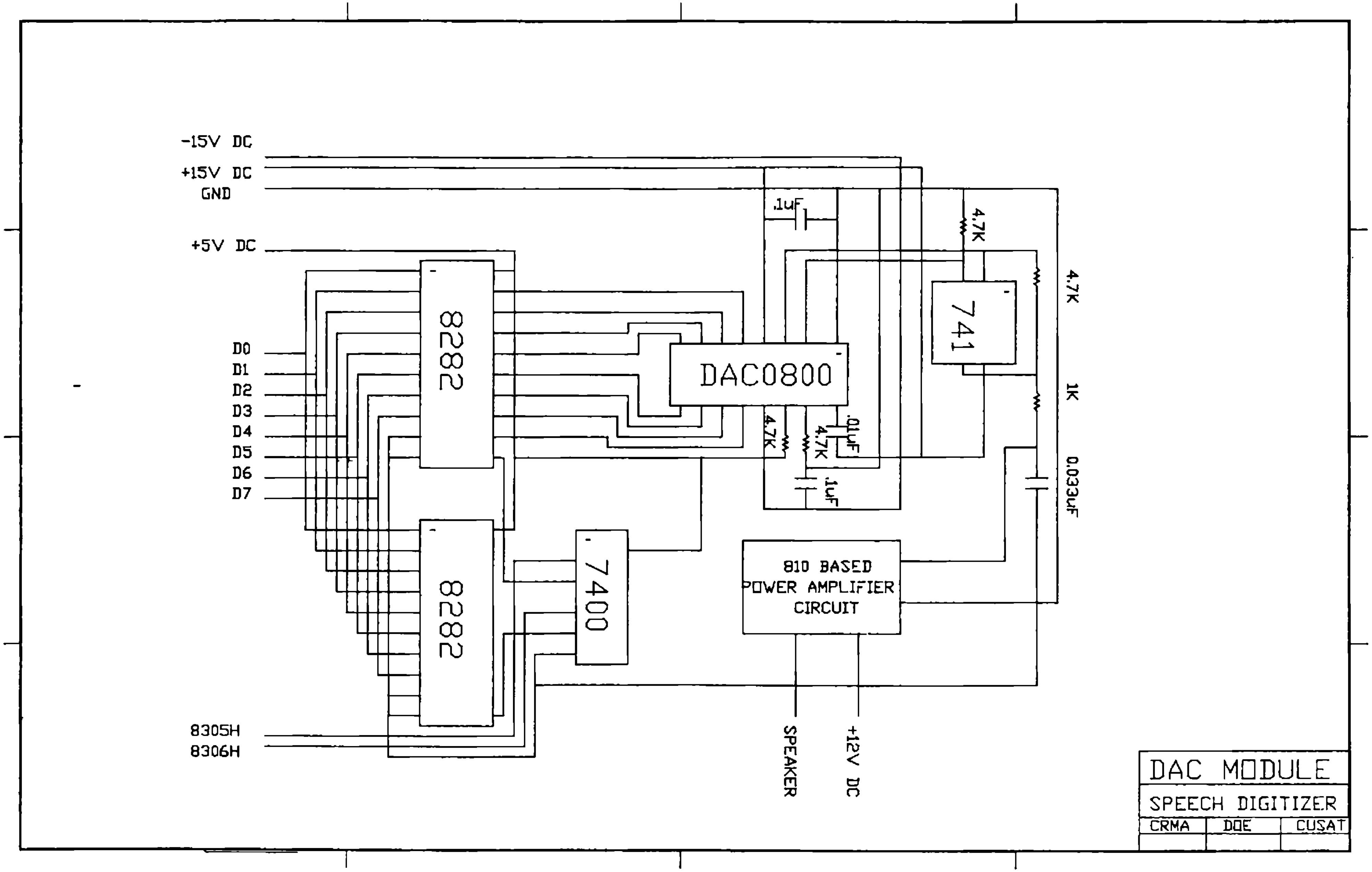


Fig.3.5 Digital-to-Analog Conversion Circuit.

3.5 SOFTWARE FOR THE SYSTEM

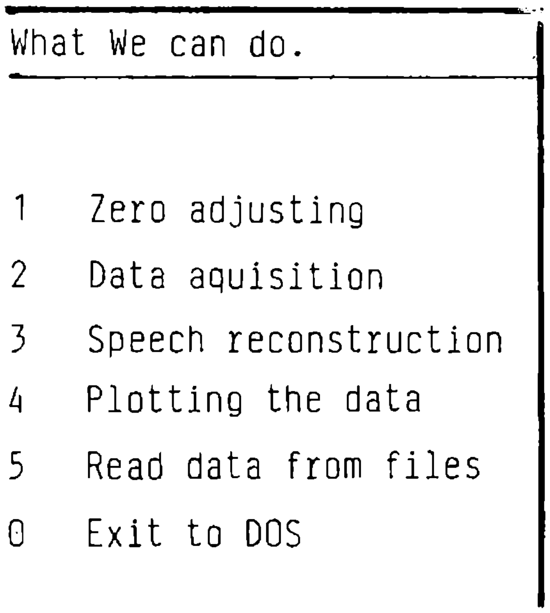
The hardware explained in the above section is supported by appropriate software. The software developed for this system is exploiting many inherent facilities of PC using high level and machine level language. Some BIOS interrupts and subroutines are made use of at certain points of software development to make the system a self contained and powerful speech station. The main program is developed in BASIC and when there is some faster operation and some machine level control is required, separate subroutines in ASSEMBLER is developed and linked with the main BASIC program. This is achieved by the facility provided for calling external subroutines in the advanced BASIC available in PC. By developing programs in this mixed level, the programmer need not bother about the file management, screen manipulation, graphics and other complex features of the computer which will be taken care by the Operating System (OS) and compiler. But at the same time fine control is possible whenever required with machine level programs, especially in real time control of I/O devices. The approach adopted is that when a faster and flexible operation is required, separate assembler subroutines are developed and assembled into relocatable object files. The main program is also compiled to an object file. The subroutines are called from the main routine. The LINKER program, available in the PC, links all these relocatable object code files into a single executable file with absolute address.

When subroutine calls are made from the main program, there are some parameters to be passed between the modules. This parameter passing from high level to assembly level and vice-versa is a tricky affair. Instead of passing the parameter, its address is usually passed. The address of the parameter to be passed are kept in the stack just below the 32 bit return address (Segment and Offset: which is done automatically when a call is made from the program). Since no POP operation could be carried out first, the BP is pushed to the stack and then SP is copied to BP as the initial operation of the assembly subroutine. Now pointing the BP appropriately to the parameter addresses by adding number of locations, one can exactly access the parameter from the memory. This can be done by putting the address of the parameter in BX register and moving the contents of the memory points by BX to the required register. While returning from the subroutine, the BASIC program assumes that the parameters are lost from the stack. So to take care of this, one has to ignore the contents of the stack when the parameter address is fixed. Thus a RET 2*N instruction is used in the subroutine, which will discard 2*N bytes from the stack, where N is the number of parameters passed. The BP is popped from the stack before this RET instruction.

The software developed is user interactive. When the main program is invoked, it will display a main menu on the screen as shown in Fig.3.6, from which the user can choose any function he desires. This menu has six options with six different functions. The functions available in the program are:

3.5.1 Zero Adjustment

This zero adjustment is meant for offset level of the input signal. When this function is invoked, the program jumps into the corresponding subroutine, which scans the data from the digitizer and displays it on the screen at the same time. The program initially issues an SOC signal to the ADC and then waits for EOC signal from the ADC. When an EOC signal is detected, this program will read the digitized data into a previously specified location. This data indicates the offset level of the input signal, without actual signal from the microphone. This is also displayed graphically and numerically on the screen. A scaled area corresponding to this value is displayed on the screen in the shape of a box. This box will have the same size in positive and negative offset, but numeric value displayed will indicate the sign. This complete process is repeated till another key is pressed. The user can adjust the offset to a minimum, bringing the size of the box to a point and thus ensuring the absence of noise in the input. A potentiometer is provided in the circuit for this purpose. Before all data storage sessions, this zero adjustment was made in order to avoid unnecessary noise in the digitized data.



```
What We can do.

1   Zero adjusting
2   Data aquisition
3   Speech reconstruction
4   Plotting the data
5   Read data from files
0   Exit to DOS
```

Fig.3.6 Display of Main menu.

3.5.2 Data Acquisition

This function is responsible for digitizing the analog data and storing it in the appropriate memory location. When the user selects this function, the program will ask for sampling frequency. In this system the sampling frequency can be varied according to the need. Since the speech signal is bandlimitted to 4 KHz in the input side, the sampling frequeny should be higher than 8 KHz. In some application, 10 KHz is also chosen as the sampling frequency. The data has to be sampled at an interval of 125 μ sec. if the sampling frequency is 8 KHz. So a delay has to be introduced between an EOC and next SOC. This delay is generated by software instruction loops, after calculating the execution time of each instruction. Hence the sampling frequency is dependent on this clock frequency of each system. Initially the delay is calculated with a system clock of 6 MHz. If PC is changed, then the sampling frequency has to be scaled according to the system clock. When the users specify the sampling frequency, system produces a beep along with a prompt to wait, while it initializes the devices. The system starts sampling when the beep stops. The program sends out SOC signal with the preset sampling time. When the EOC signal is detected from ADC, the program issues two successive output enable (OE) signals through address lines to latches which take the digitized data into the computer and stores this data in the successive locations starting from 0000H offset address of the pre-specified Extra Segment (ES). This offset addr ss is then incremented after storing the current data till the 64 KB locations gets exhausted. If one selects 8 KHz sampling frequency, then 4 sec. of speech can be stored in this 64 KB address available. If needed, similar available ES can be used, with each segment providing a 64 KB memory area. The MSDOS limitation in the main memory restricts this to a maximum size of 512 KB, which can hold about 32 sec. of speech data sampled at 8 KHz. But at present, only one segment is used for data storage. When this memory area gets filled, the program generates another beep along with a prompt message indicating that the time is up. Then the program automatically comes out of the function and displays the main menu.

3.5.3 Speech Reconstruction

The speech reconstruction function will enable the user to check whether the data stored is properly digitized by listening back the reconstructed speech. One can check the quality of a processed speech, either from the main memory or, by calling it from the secondary storage unit. In this subroutine, provision for reconstructing the whole segment of data or a selected portion of data is given. The program asks for number of samples to be reproduced and the starting point of the data so that exact desired portion can be listened to.

In this program also, the user has to specify the sampling rate at which rate the data is fed to the DAC input through a latch. The required delay between samples is generated by this subroutine. At present, an 8 bit DAC is used in this circuit. Hence the least significant four bits are truncated in the computer, and only the resultant 8 bit data is fed to the DAC. Facility is provided in this subroutine to plot or save the data which is reconstructed and listened back.

3.5.4 Plotting of Data

This function facilitates the user to visually inspect the data stored in the memory. For this function, the graphic function available in the PC is utilized and a waveform of the speech stored in the memory is displayed on the screen. In this routine also, one can view the complete data or a portion of the data stored in the memory. This selected view provides a sort of zooming facility, which enables the user for a closer examination of speech segments. A grid is also displayed for proper graduation of the waveform, which can be toggled to ON and OFF. The plotted data can be reconstructed for further checking or can be saved into a file.

3.5.5 File Saving

This facility helps the user to store the required portion or the full data files. The digitized data are stored in the main memory of the computer. After examining this data, either by reconstruction or by plotting, the required portion is stored permanently in floppy or winchester with file names. The data can be stored in two file formats viz., Binary format and ASCII format. In the binary format, the data are directly stored in the floppy using the BSAVE statement available in BASIC. When the data is to be used by other programs, it should be in ASCII format. In order to store in this format, the corresponding ASCII is calculated by the subroutine and then saved. The user has to specify the file name and the format in which it is to be stored. The ASCII file takes almost 5 to 6 times space compared to the BIN file for the same number of data. But the BIN file cannot be typed on the screen directly or accessed by external programs.

3.5.6 Reading Data from Stored File

This function in the program extends a facility to read stored data file from the secondary storage media into the main memory. The same two formats viz., Binary and ASCII, are available in this function. The data stored in binary format is loaded with a BLOAD statement in BASIC directly into the segment at a specified location. ASCII data has to be reconverted into binary and stored in the main memory. Once the data is read into the main memory all the functions discussed earlier can be utilized. The user can even concatenate speech by loading different data files in successive locations.

3.6 CONCLUSION

The indigenous and economic speech digitizing system described above with all the software and hardware deliver a powerful facility of digitizing and storing the speech data without much information degradation. From the stored data, the features are extracted and used for the speaker recognition experiment. With proper development of software one can modify the system for online recognition purpose.

CHAPTER 4

USE OF FIXED TEXT AS A TOOL

Chapter 4

USE OF FIXED TEXT AS A TOOL

4.1 INTRODUCTION

Speaker Recognition problem is generally tackled by two approaches namely, text-independent/fixed text approach and text independent/free text approach. As their names indicate, the former one makes use of a specified text for both reference and test and the latter makes decision from any text spoken by the speakers. In the earlier works of speaker recognition, the decision was taken by human beings either by listening tests or by spectrographic visual examinations. In either methods a text is involved on which the method mainly depends [2,1]. The usage of this text varies as in certain methods the words in isolation were used for recognition by listening techniques. In some other methods, these words were used in fixed contexts or in random contexts in listening test [9]. However in spectrographic methods, the visual comparison was made of a fixed text [8]. In the fixed text approach, the system searches for specific, previously fixed features at specific locations in the given text.

Later, when computers were put in place of human beings for decision making, new techniques were also evolved. At this stage, text-independent speaker recognition was also proposed [7]. In this method the speaker is free to speak any text and the system searches for speaker dependent features or patterns for taking appropriate decision.

Both these approaches have their own advantages and disadvantages. Depending on these, they are put in some specific applications. In the text-dependent approach, the system is more or less calibrated with both the reference and token and thus it has more control over the situation. In this approach different training sessions are possible from which one can choose the most appropriate and efficient text which can be reduced in size. Since an a priori knowledge of the text is available, the best speaker dependent features can be employed in this method. Hence one is free to choose the text as well as the features best suited. One can achieve better calibration of the input speech token with reference speech material which in turn will have control over the speaker as well as the speaking environment. Generally, recognition performance is better in fixed text approach due to the above mentioned reasons. But due to the same reasons, the approach cannot be employed in an application, where one has to recognize a hostile/non-cooperative speaker. Thus the applications of the fixed text approach is restricted to security applications like restricted entry and banking transactions, where the speakers wish to be identified and tend to be most co-operative.

In contrast, when any control cannot be maintained one has to depend on the free text/text-independent approach. This approach is most suited to forensic applications where the speakers are either hostile or the reference material could be a telephone speech or a recorded speech. Here, from this reference the system will formulate certain features which will be searched into the speech token of the suspected speaker. In this approach, the system does not have an a priori knowledge of the speech token and it can only look for the features/patterns until they occur in the text. So naturally a longer text is required to get a very good recognition performance. In this approach one cannot have any calibration or alignment with the reference sample. Since the size of the text required is very large, the memory size and computation time are very demanding in this approach. The most difficult problem in this type of application is to establish a valid statistical model upon which verification decision may be based. This is because of the lack of control over the speech signal and speaker and also because of the difficulty in predicting acoustical and transmission conditions. Even if one can have a statistical model for the verification decision, the lack of control in text-independent application leads to poorer performance than the application of fixed text approach [124].

4.2 FIXED TEXT APPROACH IN THE PRESENT METHOD

In the present method, which is a model based approach, a fixed text approach is preferred mainly because of its simplicity in implementation. It was felt that for a template matching technique adopted in this method, a fixed text approach is more suitable. Since the same text is used for creation of reference template during training session and the token template during the test phase, more control and better calibration can be maintained. Moreover, the a priori knowledge of the text and its pattern helps in choosing the most suitable features in advance which will improve the system performance and the computation can be effectively utilized. The analysis involved in text-independent approach is larger than in the text-dependent approach which will take considerable amount of computation time in this present set up and the memory requirement is also large. So a fixed text approach was finally decided upon. This reduces the problem dimension considerably. Several trial sessions and training sessions are conducted which finally reduced the size of the text and hence the algorithm works faster.

4.3 SELECTION OF PHRASES

Selection of the phrases for this experiment is very vital. Since a fixed text approach is adopted in a template matching method, care has been taken for including phonemes and words which are more speaker dependent. Though there is a view that all phonemes are speaker dependent, it is proved from the earlier works that some phonemes are more speaker dependent and have a better role in the recognition. Earlier studies indicate that voiced sounds or vowels are more speaker descriminating than unvoiced sounds [34]. Nasals are also found to be speaker dependent phonemes. The nasal coarticulation with other sounds was studied in detail by Su et al. [21]. This study has shown that nasal coarticulation is a very good feature for speaker recognition and it is specifically pointed out that coarticulation of /m/ with a vowel has strong speaker dependence. The characteristics of these coarticulation is found to be very difficult to change, which suggested that it is non-susceptible to mimicry.

In certain other previous works it is found that front vowels, high vowels and nasals possess greater speaker descrimination capability than other vowels [23,24].

Though the features used in these earlier works were different from the features used in the present method, it is assumed that speaker dependency remain the same for these phonemes. Since pitch is considered to be a speaker dependent feature, it is more appropriate to conduct a pitch consistency study for some selected phonemes. Thus long vowels /a/, /i/ and /u/ are taken for this study. Different words consisting of these sounds between different other phonemes at different environments are studied. The data samples are taken using 8 bit and 12 bit digitizers. The pitch is determined by a cepstral peak method. The words chosen are in Malayalam language and have

CVC or CVCV nature. Thirty words are selected for different phonemes and these three long vowels are examined. Plosives, nasals and affricates are also included in these phonemes. From the results it is found that absolute consistency in pitch is impossible. The pitch periods are slightly different for the same vowel in different contexts, environments and sampling set up. When a tolerance of ±500 microseconds is given, consistency is maintained in the pitch for all the three vowels in different phoneme contexts. But a variation of ±750 microseconds is observed for samples taken in different surrounding environments. It has very large variation in the pitch periods for the samples taken at 8 bit and 12 bit sampling set up. Among the three vowels, /u/ has shown maximum consistency especially when it follows the nasal /m/.

Considering the aforementioned factors and adopting to the specific requirement of this method, a phrase set consisting of four phrases is chosen. Each phrase consists of four words. In this method an entropy vector formed from the probability of occurrence of phoneme is taken for template creation. For this, the phonemesare to be distributed almost uniformly throughout the phrase set. Thus, after some trial and error method the following phrase set was used.

DROP	COIN	AFTER	TONE
PUSH	YELLOW	AFTER	BELL
CLOSE	DOOR	AFTER	PARTY
RIGHT	ROTATION	OPEN	DOOR

In this method, along with the entropy vector, the timings of the phonemes, words and phrases are also being used. So segmentation of the phrases is essential. Since the timings are critical and a manual segmentation will lead to errors, an automatic segmentation method is employed. This segmentation algorithm is based on the energy envelope of the words. At the first level word segmentation is tried. But it is observed that two phrases show error in segmentation for some speakers specifically. This error is due to the nature of the words in that phrase and also the speaking habit of the speakers. The problem could have been removed by asking the speakers to introduce silence in between the words. But this could affect the naturalness of the speaker. The trailing end of one word is merged with the starting of the next word making it difficult for word segmentation. This is because the region of ending of one word and starting of the next are closer in the vocal tract. To avoid this difficulty, an alternate set of phrases after making modifications and incorporating new words is selected. Here the error creating adjacent words are replaced with words which have natural silence in between. Since one word ends at one region and the next word starts at a different region, distinctly apart in the vocal tract, merging of words is avoided.

Thus the new set of phrases is

DROP	COIN	AFTER	TONE
PUSH	BLUE	AFTER	SPEECH
CLOSE	DOOR	AFTER	PARTY
RIGHT	MOVE	CLOSE	LOCK

4.4 CONCLUSION

The new model proposed for speech production and hence speaker recognition calls for a fixed set of phrases. These are chosen after due consideration given to speaker dependent phrases, brevity and ease of use. After trial run on the system, further modifications were introduced and the final set of phrases chosen is well balanced and free from segmentation problems.

CHAPTER 5

FEATURE SELECTION AND EXTRACTION

Chapter 5

FEATURE SELECTION AND EXTRACTION

5.1 INTRODUCTION

In pattern recognition problems, one has to always consider the dimensionality of the pattern. Features are pattern descriptors, having lower dimension. These features are important in representing a pattern and characterizing the discriminating properties of pattern classes. The complexity of the algorithm increases with the number of patterns and their dimensions Therefore it is essential to restrict the number of patterns to the most important and having discriminatory information. This in turn will reduce the hardware requirement and also the cost of measurement extraction. The dimensionality reduction in most of the cases improves the performance [43].

The reduction in dimensionality is carried out in two ways. In the first one, the least contributing measurements to the class separability are identified and eliminated. Thus a subset of the superset is chosen by ignoring the redundant and the dispensable measurements. This process is called feature selection. In the second method, the measurements are mapped into a lower dimensional feature space to reduce the dimension. This is called feature extraction. The important points to be considered for feature selection and extraction are the feature evaluation criteria, the dimensionality of the feature space, the optimization procedure and the form of mapping.

5.2 FEATURE SELECTION METHODS IN PATTERN RECOGNITION

5.2.1 Probablistic Method

This method is considered in terms of error 'e' in the two class problem. When the two probability density functions overlap, the error will be maximum and when the PDF's are non-overlapping the error will be zero. There are several probability distance measures available of which Mahalanobis distance measure is very popular and commonly used in many speaker recognition methods.

$$d = (\mu_2 - \mu_1)^T \Sigma^{-1} (\mu_2 - \mu_1) \qquad (5.1)$$

Where μ_i's are mean vector and Σ is the covariance matrix.

5.2.2 Probablistic Dependence Measure

In this approach two random variables are involved viz., a pattern vector and the class. An observation of the outcome of the former enables one to make a decision about the latter. Conditional density function is made use of in this method. In situations where the pattern vector and class are independent, the method fails in classification.

5.2.3 Entropy Method

Entropy measure is also employed for feature selection. As a posteriori probability is calculated to determine how much information has been gained from the experiment. If all the classes are equally probable, then the information gain is minimal and the entropy is maximum.

5.2.4 Interclass Distance Measures

In this method, it is assumed that the pattern vector of each class occupy distinct region in the observation space. The average pair-wise distance between the patterns in the set is a measure of class separability in this space. The Euclidean metric is a common interclass distance measure in speaker recognition problems.

5.3 FEATURE SELECTION CRITERIA IN SPEAKER RECOGNITION

Speaker recognition problems exclusively have to meet certain specific criteria for the feature selection, which depends on the measurements of speech characteristics. These criteria can be listed as shownbelow [10].

1. The features should occur naturally or frequently in normal speech.

2. These features should be easily measurable.

3. They must vary as much as possible among speakers, but must be as consistent as possible for each speaker.

4. They should neither change over time, nor be affected by the speaker's health.

5. They should neither be affected by reasonable background noise nor depend on specific transmission characteristics.

6. They should not be modifiable by conscious effort of the speaker or at least be unlikely to be affected by attempts to disguise the voice.

Practically it is impossible to incorporate all the above criteria to make the recognition system foolproof. Especially the last mentioned three are most difficult to adopt. So one has to make an engineering compromise.

5.4 FEATURE EVALUATION IN SPEAKER RECOGNITION

There are different popular methods used for feature evaluation in speaker recognition problems. The F-ratio analysis is one of these evaluation techniques for speaker discrimination ability of the features. In speaker recognition, a parameter is said to be good when the individual speaker probability distribution is as narrow and as widely separated as possible. The F-ratio can be defined in other words, as the ratio proportional to the variances of speaker means to the mean of the speaker variances. When the individual speaker distributions are farther apart and narrower, the F-ratio value is higher and the parameters are selected as suitable parameters [5].

Discriminant analysis is another feature evaluation technique in speaker recognition. In this method new features are created with linear combination of the original features. The optimum linear transformation of the original feature space is determined by a combination of eigen vector analysis and F-ratio technique [23].

The draw back in these two feature evaluation methods is that the features with high F-ratio may not contribute much to the performance of the recognition system, than a feature with lower F-ratio [23].

Another evaluation technique uses a knock out method. The stress in this technique is that the features selected should contribute to the performance of recognition. If a set of N features are available for evaluation, the effectiveness of a subset of N-1 features is considered and the error performance is determined. By this method, the most effective subset (N-1) is chosen and the single feature is eliminated or knocked out. Then a subset of (N-2) is chosen and the same elimination technique is used. This process is repeated till all the features are knocked out and the effectiveness is arranged in the reverse order [23].

Dynamic programming methods are also used for determining feature effectiveness [51].

5.5 FEATURES USED IN THE PRESENT METHOD

Though there are many feature selection and evaluation methods, there are not many standardized features in the speaker recognition systems. The features used in different methods are dissimilar. Many features are claimed to be successfully used in different recognition systems. Using all these features in a single system is not practical due to various reasons. A possibility of using a few features in this method is studied. Attention is given to features which are easily measurable and implementable in a small computer system at a reasonably good performance. Avoiding very complex features, one can save a lot of computation time and make the system faster. An appropriate knock out method is adopted by critically examining the performance. A few measurements which are felt as simple are chosen for this purpose. Different combinations of these features are used in the method and performance is determined. The features whose contribution is small are left out.

The features selected in this method are

1. Short-time energy
2. Zerocrossings
3. Autocorrelation function (ACF)
4. Pitch by time domain symmetry and
5. Phoneme timings, transition times and their entropy from the utterances.

The features energy, zerocrossings and ACF are successfully used in one or other forms in the earlier methods [19,20,22,25,50,92] while time domain symmetry measures is a new technique used in this method. The duration of speech was also considered in some earlier methods [1,10]. In some features, the measurements taken are different from those adopted in the earlier methods.

5.5.1 Feature Extraction

While some of the simpler techniques are adopted directly from previous methods, others are extracted using simple techniques as explained below.

For finding the short-time energy, a segment of fixed length (600 samples) at maximum amplitude region of each word in the phrase is chosen and the energy is determined.

The number of zerocrossing within a length of four pitch periods is determined. The number of positive peaks within this same period is also taken as a feature measurement.

The Autocorrelation Function of 512 points is determined using an FFT method. From this ACF, the distance between the first two prominent peaks, slope between these peaks and number of smaller peaks within these peaks are taken as measurements.

5.5.1.1 Time domain symmetry

Since time domain symmetry is a new approach for pitch detecting, it is discussed in detail. This feature is mainly selected because, the fundamental frequency or pitch is a very important speaker discriminating feature in many speaker recognition systems [5,10,18,23,33,37].

It is an accepted fact that the voiced sound has a periodicity and determination of this periodicity leads one to pitch detection. There are different standard algorithms used for this purpose viz., Autocorrelation, Center clipping method and cepstrum method [144,145]. In most of these methods, some type of transformation is involved and also the computation involved is lengthy. So a new approach was felt necessary and has been tried for detection of pitch which in turn is also a speaker dependent measurement. The symmetry check is purely a time domain method and mathematically less complex.

The basic principle is that of extraction of the periodicity of the voiced signal. It is well accepted that speech signal is a complex signal and it is the convolution product of the fundamental frequency generated by the vocal chords and the harmonics generated from the vocal tract. Thus there are two significant components viz., the slowly varying signal corresponding to the vocal chords and rapidly varying signal component corresponding to the vocal tract harmonics. In other words one component corresponds to the pitch and the other component corresponds to formants. Fig.5.1 shows a typical voiced sound segment. In this method, the main task is to filter out the effects of vocal tract and extract only the slowly varying fundamental frequency component.

An algorithm has been developed for this pitch detection. This algorithm works on a window based operation. The symmetry in the window is checked every time and if it shows a symmetry, this window is considered for pitch. Initially a small window of 32 samples is considered. Correlation coefficient of the samples in that window is determined. If the correlation coefficient exceeds a threshold value, the periodicity is checked

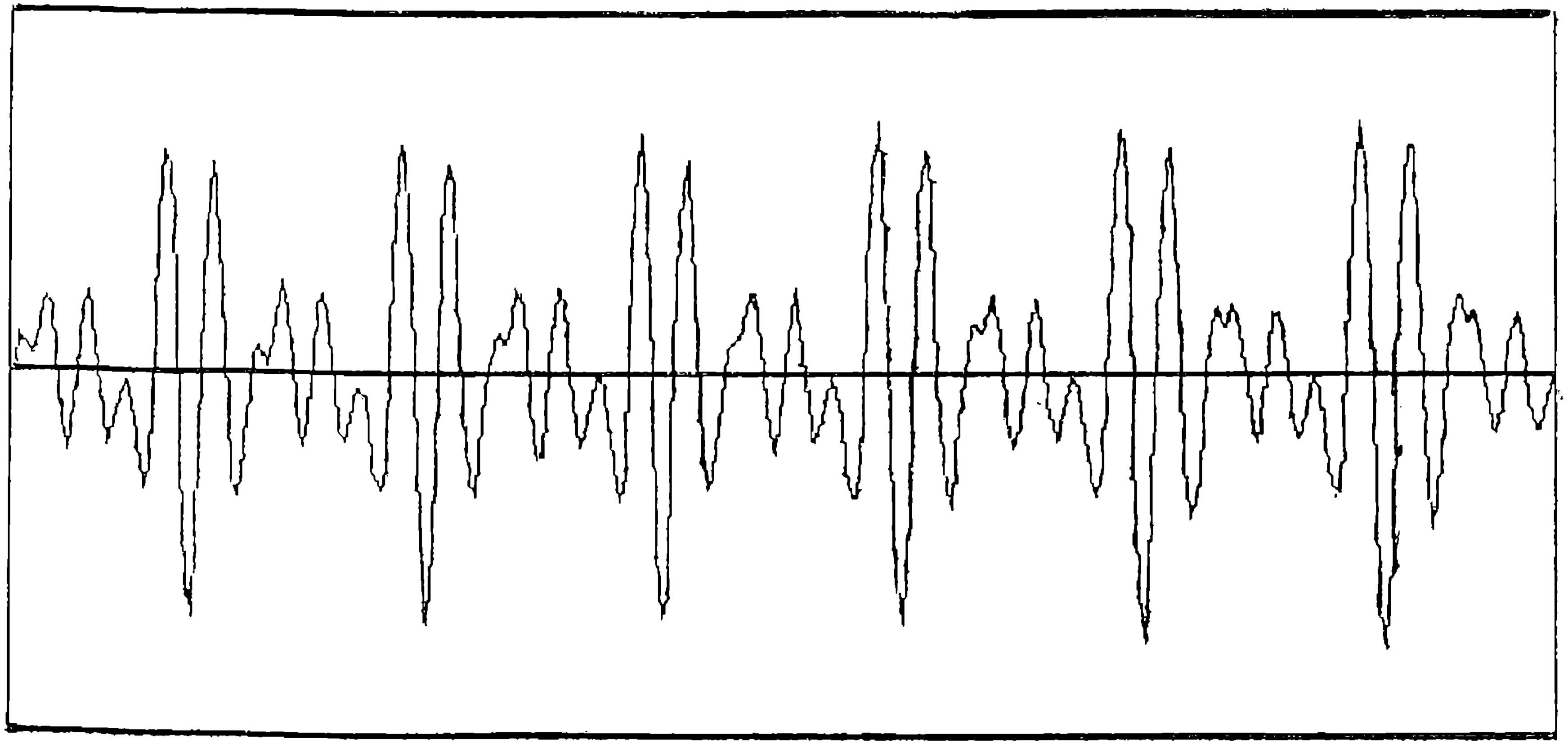


Fig.5.1 A Typical Voiced Speech Segment /a/

with other symmetry measures, these symmetry measures are average, rms and rate of change of the samples in the window. For this purpose, the samples in the window is divided into equal halves and these measurements are determined. The difference in each measure gives a symmetry measure. The three symmetries are determined by

$$A = \left(\frac{1}{N/2}\sum_{i=1}^{N/2} X_i - \frac{1}{N/2}\sum_{j=\frac{N}{2}+1}^{N} X_j \right) \Big/ \frac{1}{N}\sum_{i=1}^{N} X_i \tag{5.2}$$

$$R = \left(\frac{1}{N/2} \sum_{i=1}^{N/2} (X_i^2)^{\frac{1}{2}} - \frac{1}{N/2} \sum_{j=\frac{N}{2}+1}^{N} (X_j^2)^{\frac{1}{2}} \right) \Big/ \frac{1}{N} \sum_{i=1}^{N} (X_i^2)^{\frac{1}{2}} \tag{5.3}$$

$$C = \left(\sum_{i=1}^{N/2} (X_i - X_{i+1}) - \sum_{j=\frac{N}{2}+1}^{N} (X_j - X_{j+1}) \right) \Big/ \sum_{i=1}^{N} (X_i - X_{i+1}) \tag{5.4}$$

where A, R and C are normalized symmetry coefficients corresponding to Average, Root Mean Square, and Rate of change. N is the total number of samples in the window considered and X_i is the ith sample.

The total symmetry of the window can be computed to be

$$T = 1/3\ (A+R+C) \tag{5.5}$$

From eqn.(5.5) it can be observed that when T = 0, the signal corresponding to the samples in the window shows an absolute symmetry and when T approaches 1 it shows maximum asymmetry or lack of periodicity. A threshold value of T is fixed by trial and error and the window of samples falling within this threshold is taken as periodic. The period corresponding to half the number of samples in the window (N/2*ts, where ts is the sampling time of the signal) is taken as the pitch period. If there is no symmetry in the window considered, the window size is changed and symmetry is again checked in the same manner. The complete algorithm, experiment and results are given in Appendix A. Upon confirmation of consistency the technique can be accepted as a good and easy technique for determining pitch period.

Once the pitch period is detected, the intermediate measures viz., average, RMS and rate of change of symmetry coefficients and pitch number are taken as valid measurements for speaker recognition.

An attempt is also made to define a set of relationships between the features selected in the current work and elsewhere and a new model for speech production. This is presented in a subsequent chapter after presenting the model.

5.6 CONCLUSION

A method is explained in this chapter which selects features similar to those established method but with different techniques of measurements and evaluation. A suitable strategy for knock out of features is also developed. A reasonably good performance is obtained using a set of features chosen in this method.

CHAPTER 6

A KNOWLEDGE BASE SPEAKER RECOGNITION MODEL

Chapter 6

A KNOWLEDGE BASE SPEAKER RECOGNITION MODEL

6.1 INTRODUCTION

There have been many approaches to the Speaker recognition problem. Many of these approaches are based on models for speech production. A universally acceptable model for speech production in human beings is yet to be established. Most of the models are based on the physical human vocal organs as shown in Fig.6.1.

Generally, speech production is divided into three stages viz., Generation of voice source, Articulation and Radiation. These three stages can be shown in simple blocks as in the Fig.6.2.

A complete representation of the actual speech producing mechanism is not possible since, many of the related processes in human beings are yet to be explained. Electrical equivalence of the physiological organs is employed to explain the process [141] According to this equivalent circuit, the voice source is simulated by a train of periodic pulses and triangular waves whose period and amplitude correspond to pitch and intensity of a voice source. An unvoiced source is simulated by white random noise whose average power corresponds to average turbulence energy.

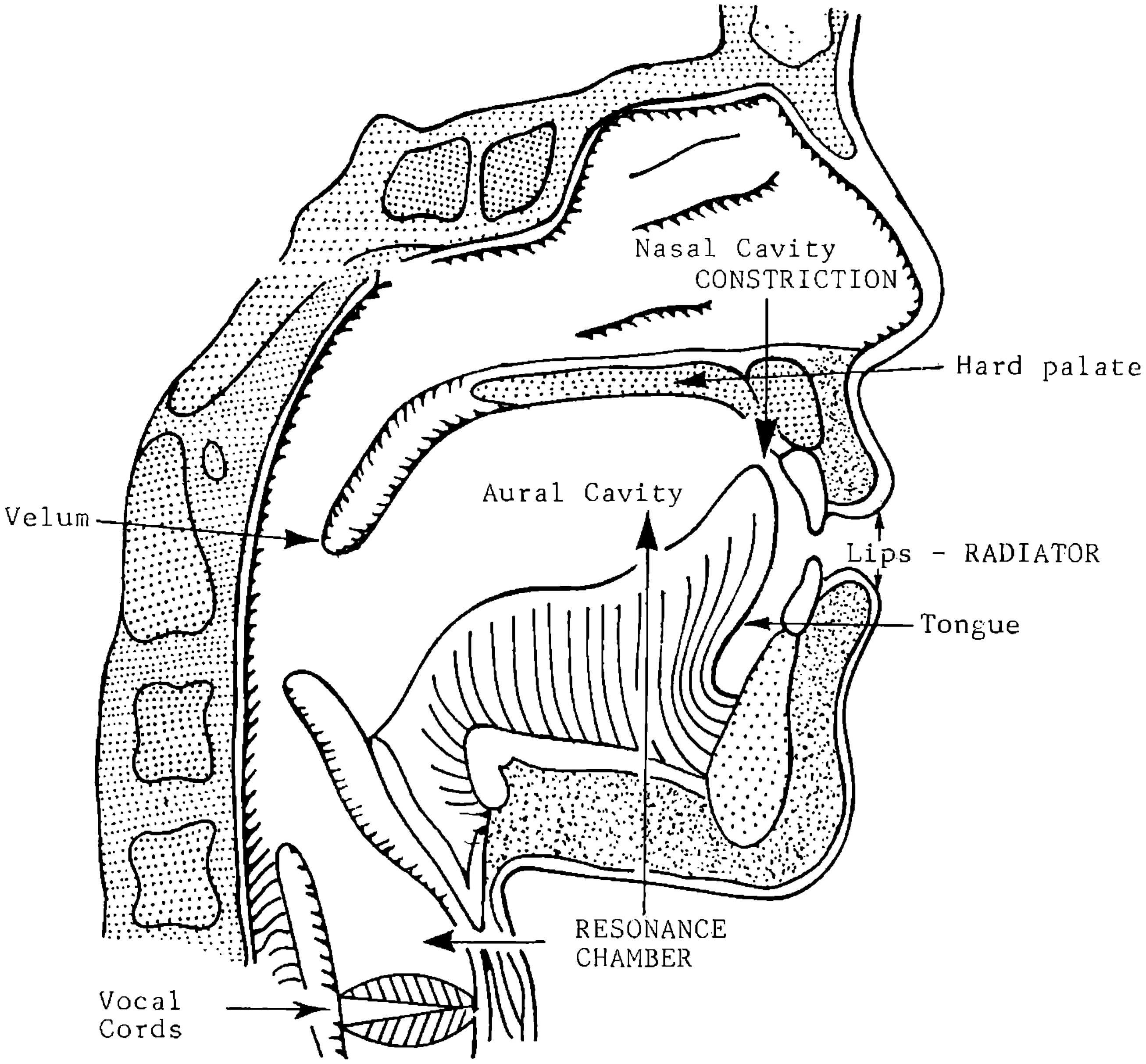


Fig.6.1 Physical Human Vocal Organs for Speech Production

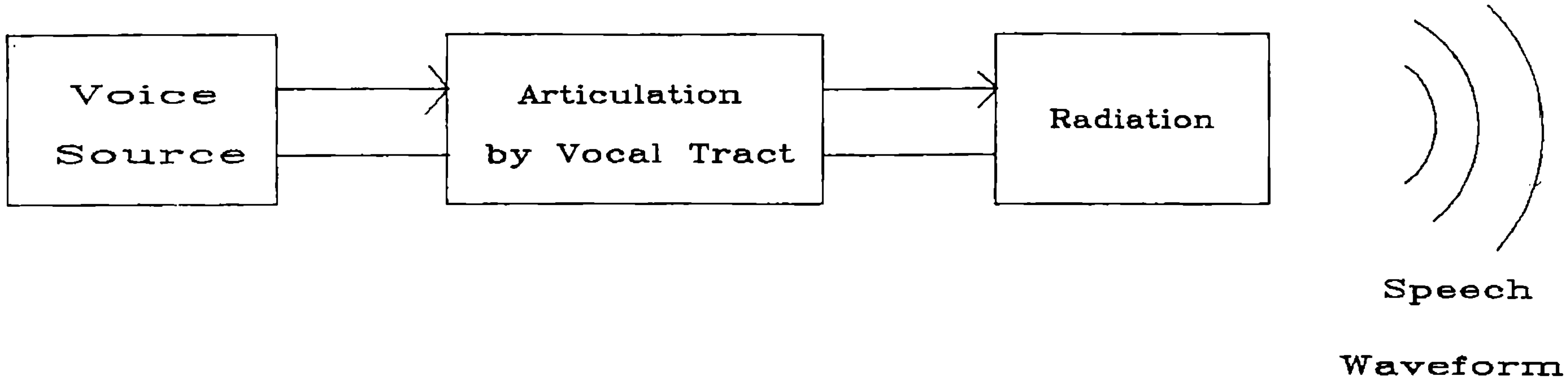


Fig.6.2 Fundamental Process of Speech Production (Block Diagram)

Articulation is simulated by the connection of resonance circuits in series and in parallel. Radiation is simulated by serial connection of L-R elements. The output is picked up as a voltage difference across the L element and R represents the loss of energy by radiation.

Large amounts of analyses have been conducted on speech, and as a result different models have been proposed. Fig.6.3, 6.4, 6.5 show some of these models. But all these models are only good approximations to speech generation. There is no model which accurately represents human speech system, particularly the 'impulse trains' and the motivation for producing a sound.

In the present discussion a model has been suggested, in which a Knowledge Base (KB) plays a pivotal role with a feed back system in the production of speech. As mentioned earlier the models accepted widely employ 'a train a periodic pulses' or 'gaussian random noise' as the source signal. The model presented here tries to find out how these trains of pulses are generated.

6.2 A HEURISTIC APPROACH

Instead of a mathematical model, a heuristic approach leading to a model is presented. Man communicates with the help of developed languages and speech. Hence speech is a very powerful communication tool. If we examine the history of spoken

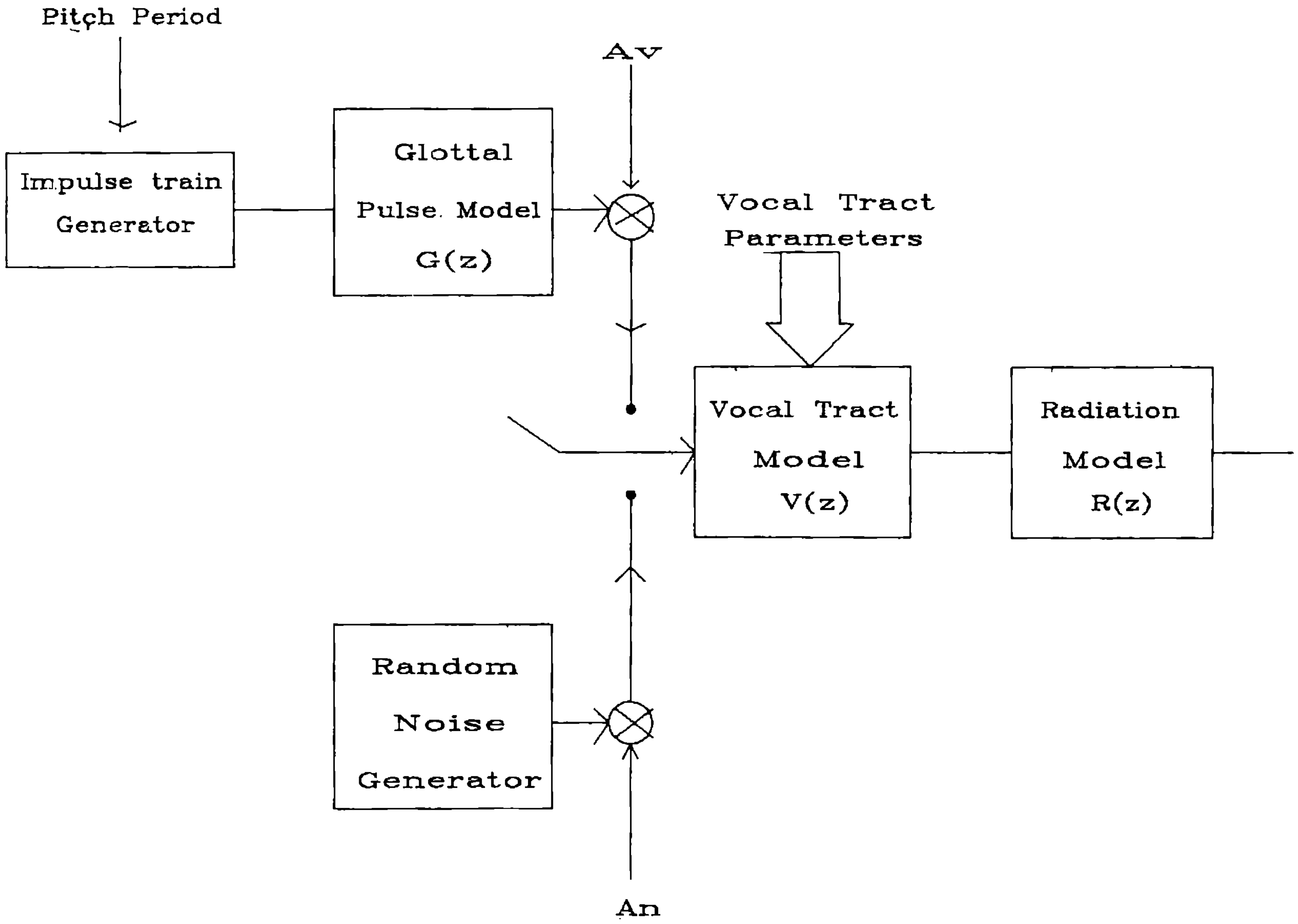


Fig.6.3 General Discrete-Time Model for Speech Production

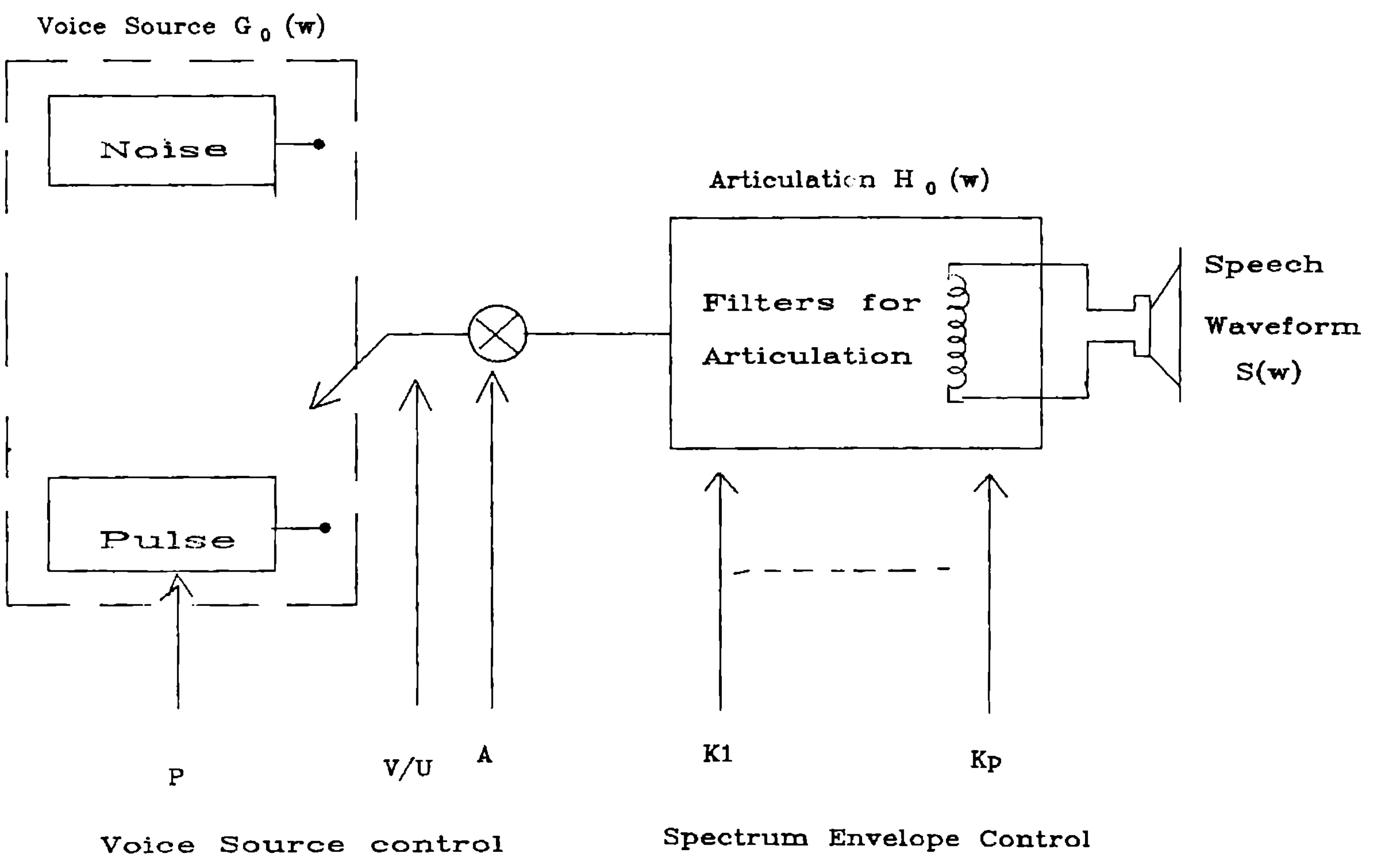


Fig.6.4 Model for Linear Independent Speech Production Mechanisms. P is the Pitch Period. V/U the Voiced Unvoiced decision. A the Source Intensity and {ki} the PARCOR coefficients.

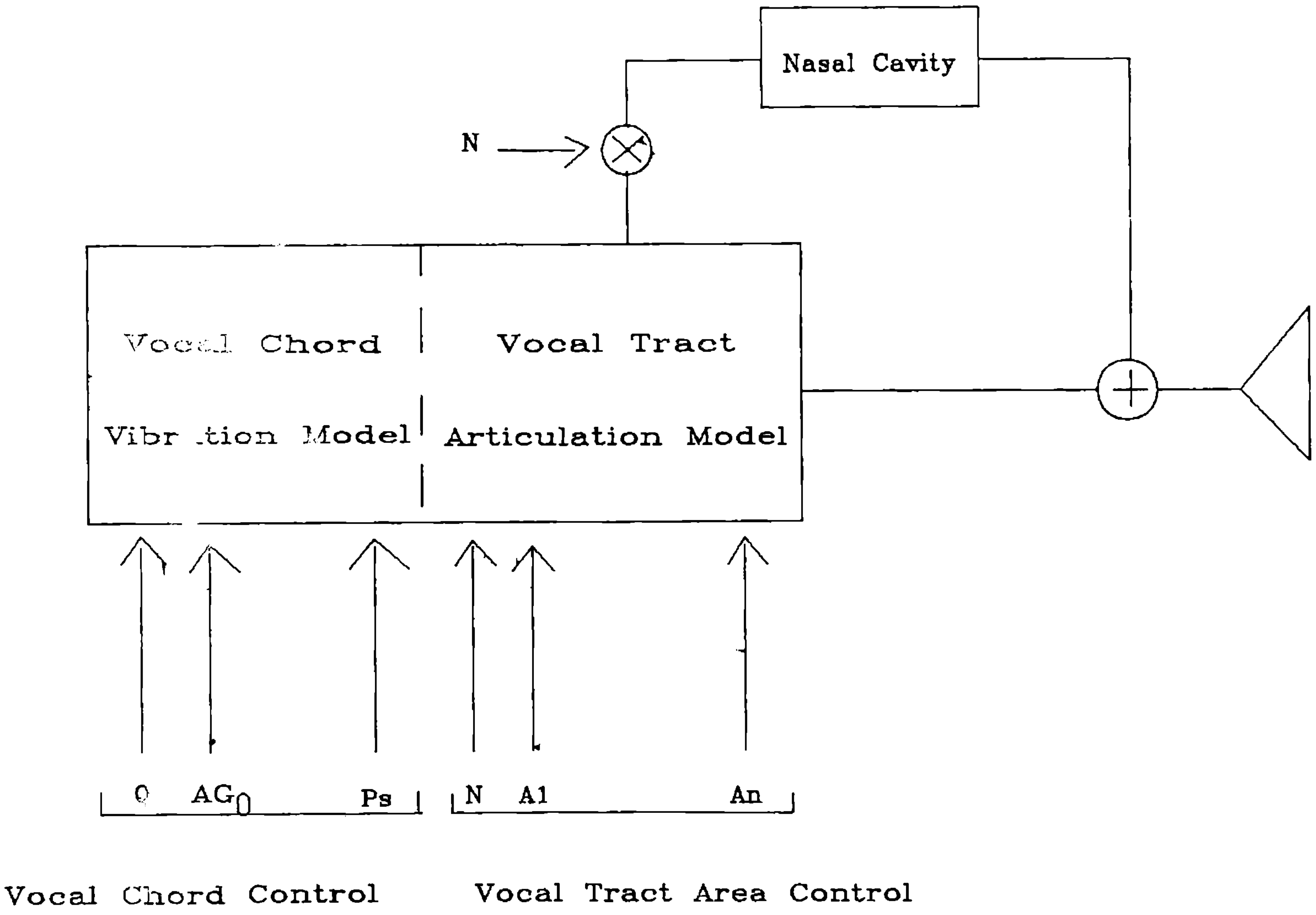


Fig.6.5 Substantial Model of Speech Production. Q is the Vocal-cord tension. AG0 the Vocal-cord opening area at rest. Ps the subglottal air pressure, N the Nasal Cavity coupling and {Ai} the Vocal-Tract area function.

languages in human beings, it can be seen that ancient people communicated using sounds and gestures. For every action and material objects around, they produced special types of sounds by which they could convey the idea. By a gradual learning process, man developed language which passed through different stages and reached the present day form of organized speech.

People in different parts of the world speak different languages and each language is spoken in various styles. Linguistic experts explain these aspects assuming that there is a basic set of 'root' sounds from which sets of languages developed. There are certain 'sounds' and terms common to more than one language. This might be because, speech did not originate from a single place. Human beings have developed their instincts to speak independently in different parts of the world and each of these independent endeavor gave birth to individual languages. It can therefore be said that eventhough the basic physiological features are universally the same in humans, geographical and other aspects are likely to influence the sound producing mechanisms to the extend of affecting the input pulse trains.

One can introduce the concept of Knowledge Base in human beings. This knowledge base contains different knowledge regarding various human related activities. Initially, the KB, which is the part of human brain, contains very little knowledge about speech. Gradually through learning process, the KB is expanded and man is able to generate speech in the required form. This process is observed in the case of children, who learn to speak. At the time of the birth, their knowledge regarding speech is practically nil. Currently some experiments have established that a certain amount of prenatal KB exists. The only instinct of a new born baby is to cry. Studies have shown that even the cry of a child is of different styles at different situations and are classified as Call cry, Hunger cry and Anger cry [142]. It is also noted that one who is closely associated with the infant (eg., Mother) can easily identify this. When the baby grows, it produces some other sounds and tries to imitate the elders. He is expanding his knowledge base by listening and watching the people around him.

The knowledge base is associated with 'learning process' and a feed back system. In the case of human beings, the feed back system for speech is the auditory system. This child tries to master speech by trial and error method. He closely watches the elders and tries to generate similar sounds. If he finds it matches with those sounds, the parameters corresponding to that speech is stored in his knowledge base. The knowledge base developed by each child depends on the environment in which he grows. That is why the child speaks the same language the people close to him speak and most often he adopts a similar style of speech.

Though the style of speech and language spoken by the children are similar to that of elders, there is distinct variation between their sounds. Each individual's speech is different from others' One main reason is that, physical size and shape of vocal organs in each person are different. Thus even with the same knowledge base, the sound produced might differ. Another reason is that, the knowledge base developed by each person is distinctive. During the learning process, the child is associated with different people, whose styles may be different. Thus the child develops a knowledge base, which is a mixture of all the styles and languages of the people close to him. Moreover, the parameters in the KB are only approximations due to the trial and error. Hence, depending on the people associated with each individual and the environment, each person develops a unique knowledge base.

6.3 A KNOWLEDGE BASE SPEECH MODEL

According to this model, sound producing physiological organs are influenced by an individual's knowledge base. When a person feels like speaking, the control parameters corresponding to those sounds are sent to the mechanism which controls each stage of speech production. A feed back mechanism is always active, which corrects the speech and knowledge base when required. Fig.6.6 shows the block diagram of this model.

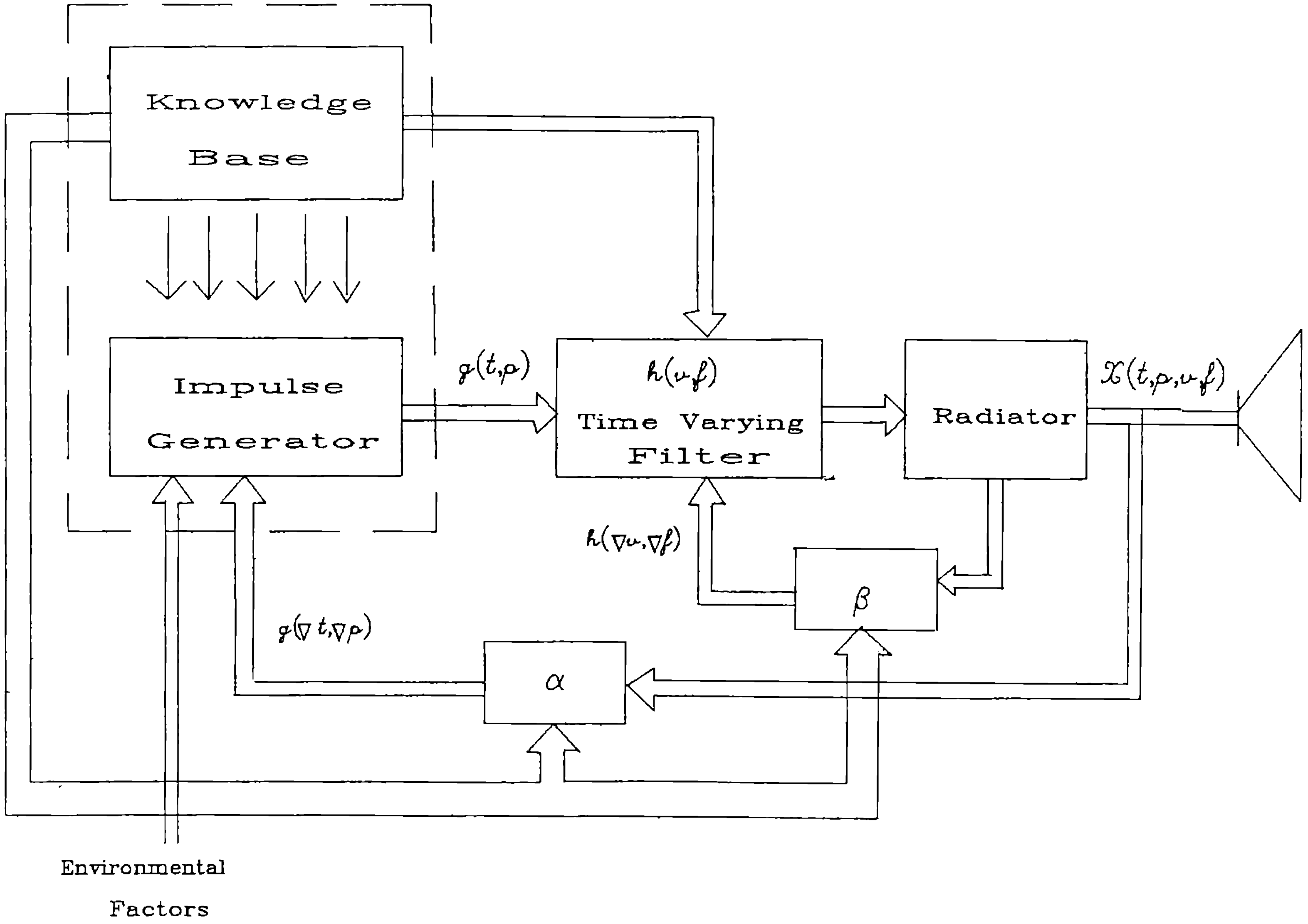


Fig.6.6 Knowledge Base Speech Model with Feedback Network.

The knowledge base and the feed back network are the novelties of this model. When a sound is to be produced, the parameters residing at specific locations in the knowledge base is transmitted to an intermediate unit called Impulse Generator. This unit will generate impulses at a required rate which control the air pressure and the rate of air flow. The output of this unit is a function of time and pressure, g(t,p), where t and p are parameters controlling the rate of impulses and pressure of air.

The vocal tract and nasal cavity are modelled as a time varying filter. The filter coefficients can be taken as a function of frequency and volume, h(f,v). The factors controlled by the filters are the shape and resonance frequencies. When the impulses pass through this time varying filter, these two functions are convolved and an output function is produced from the radiator, X(t,p,f,v).

The speech system is a closed loop system in which there is a feed back and correction network. These networks are shown as α and β in the model. α and β networks keep track of the sound output and each one checks the parameters in comparison with the knowledge base. α network traces g(t,p) and the change in this function g(Δt, Δp) is applied to the impulse generator which will correct the subsequently generated impulses.

Similarly β network traces the function h(f,v) and the correction factor h(Δf, Δv) is applied to the time varying filter. In an ideal case, the difference between knowledge base parameters and actual speech output parameters is zero.

$$X_{KB}\ (t,p,f,v) - X_{op}\ (t,p,f,v) = 0 \qquad (6.1)$$

then

$$g(\Delta t, \Delta p) = 0 \quad \text{and}$$

$$h(\Delta f, \Delta v) = 0$$

which means that there is no correction factor applied to the system. But this ideal condition never arises in a real life situation. In a speech synthesizer, when fixed impulses are sent to produce a sound, keeping all other factors unchanged, it always produces the same sound. There is no feed back mechanism in a speech synthesizer. But in human beings, the speech produced at different times are different, in spite of the same knowledge base. This is because of the dynamic nature of human speech. There is an error between the stored parameters and the actual sound parameters made available through the feed back network and on a dynamic basis, this error is corrected. The existence of the feed back mechanism is obvious from the fact that deaf persons are normally dumb. In general those with hearing impairment, particularly children, exhibit difficulties in speech production too, irrespective of whether the speech producing organs are in good order or not. In the case of deaf individuals, their feed back mechanisms do not help them either in learning or in correcting and hence the knowledge base regarding speech might be empty.

The error probability in the speech of human beings is a unique feature which is different for each speaker. The approximations and comparisons made by each person are different and hence their knowledge base and error probability are also unique. The existence of the feed back network is the main cause for the intra-speaker variability. This network tries to correct the speech dynamically, while it is being produced. So variation for the same speaker is possible. But the error pattern is unique for each speaker and varies only within certain limits. Thus recognition of speakers is possible if one can measure these error probabilities in the speech. The feed back network compares the error pattern with the patterns corresponding to each speaker stored in the knowledge base and the human system takes an appropriate decision. The recognition is assured only if the pattern is available in the KB; ie., only if the person is known earlier to the recognizer.

Nevertheless, it is not possible to measure the exact error parameters, which the human system generates. This is because, it is still unknown precisely what a human being measures internally. The knowledge base is not measurable directly, so are the impulses and other parameters introduced in this present model. One has to derive these parameters from the quantifyable and measurable features, which can be extracted from the actual speech output.

However, it stands to reason to believe that the human recognition is based on simple parametric comparison since recognition ability is inherent in human beings. Even without much intelligence, recognition is possible. A small baby can recognize his mother's voice. similarly, pet animals, whose intelligence is very low compared to human beings, also can recognize their masters' voice. Assuming that the recognition process in all these cases are same and only the size of knowledge base is increasing, one can contemplate on simple techniques. Considering the tremendous size of the knowledge base in human beings, if very complex measurements and techniques are involved, definitely the recognition cannot be so fast and natural as is experienced. Taking all these factors into consideration, simple and easily measurable features are used in this present method relating them to the model.

The relation between some of the speaker dependent features and the speech model are examined here. Pitch period is a largely used feature in many earlier speaker recognition methods [3-6,8,10,18,37,90]. The pitch period is a direct implication of the impulses and it's rate. This can be expressed by a function of time and related to the function g(t,p) in the model. Formant frequencies, power spectral density, spectrogram, phoneme spectra, filter averages, log area function etc., are measurements considered as speaker dependent and used in many methods [5,9,12,20]. These features are all frequency dependent and can be related to the filter function in the model, h(f,v). It is felt that these features are more physiological than Knowledge Base dependent. Only the variations presented during times of ill health, stress and emotional strain are possibly KB dependent. Energy, another widely used speaker dependent feature, is dependent on the pressure function generated by impulse generator. Zerocrossing rate, cepstrum and autocorrelation function are some of the time domain features available in speaker recognition [19-21,25,111]. These are also functions of time and controlled by the impulse generator. These features have a greater dependence on the impulses produced as input to the vocal tract and hence are probably generated by the block labelled 'Impulse Generator' and consequently less dependent on physiological aspects of the vocal tract.

From the above set of speaker dependent features a few features are selected for the present study. More attention is paid towards the simplicity of the measurement and ease of implementation. Thus, pitch by symmetry check, short-time energy, zerocrossing and autocorrelation function are selected. The phoneme rates and transition rates are also used in this method. These selected features are able to measure indirectly the knowledge base parameters. The difference between the knowledge base parameters and the parameters from the output speech is the error in the system. This error, which is taken as correction factor in the feed back system, is the cost to produce the actual speech from the original KB parameters. This error can be measured and compared with similar error patterns stored in the system.

6.4 CONCLUSION

A new model for speech production has been suggested. The necessity of a knowledge base and a feed back system is established by this model. The learning feature with the help of the feed back network is introduced. The inter-speaker and intra-speaker variation is discussed. This model also explains the recognition possibility using simple techniques and measurements.

CHAPTER 7

SPEAKER IDENTIFICATION ALGORITHMS

Chapter 7

SPEAKER IDENTIFICATION ALGORITHMS

7.1 INTRODUCTION

There are many Speaker Identification methods based on different models for speech. The identification performance mainly depends on the features selected and the way in which they are used. Most of the speaker recognition methods employ a pattern recognition approach, in which the template formulated from different speaker dependent features being the pattern descriptor. A decision or pattern classification is based on the different distance measures.

The three stages involved in a speaker recognition experiment are:

1. Selection of appropriate speaker dependent features and their extraction from the original speech.
2. Template creation from the features extracted
3. Decision making after finding distance measures between the templates.

The recognition methods differ either in all these stages or in any one of the steps.

In this method the features are selected on the basis of their contribution to the performance of the system. A few features are selected based on their established speaker dependency. Then each one of these are used in the experiment and performance is evaluated. The features which contribute less to the performance are knocked out. The measurement of each feature is also different in this method. Instead of taking the pitch period, which is a known speaker dependent feature, four intermediate measurements are taken in this method. Similarly some new measurements are also introduced.

Two methods are tried for speaker identification. One is, the Fixed Text Phoneme Model (FTPM) and the other is Similarity Measure Method. In the first approach an entropy measure of the phonemes is made use of. In the second approach, the feature measures are formed into a template and the similarity between templates are determined.

7.2 FIXED TEXT PHONEME MODEL

According to the speech model discussed in the previous chapter, there is an error pattern due to approximations and feed back systems. This error pattern must be unique for every speaker and must be determinable from measurable quantities. Since a fixed text approach is used in this system, the probability of occurrence of the phonemes are known a priori. From this probability of occurrence, entropy of each phrase can be determined using

$$H = - \sum_{i} p_i \log_2 p_i \qquad (7.1)$$

Where p_i's are the probability of occurrence of each phoneme. Eqn.7.1 is a standard expression of entropy of discrete symbol source. p_i stands for probability of occurrence of the symbol. Here however, the significance of p_i and the evaluation of p_i are different. A complete phoneme is treated as a symbol from a source phrase. Thus the contribution towards the entropy of the source (phrase set) due to the phoneme is averaged and determined. For reasons already explained in the preceding chapter, this is speaker dependent.

Similarly entropy is determined for all the four phrases and a phoneme entropy vector is obtained. This phoneme entropy vector is standard for all the speakers. From the actual phoneme timings and transition timings, a Feature Entropy Vector is formed. The ratio of individual phoneme duration to the total phrase duration is calculated for the selected phonemes. Since the phonemes are distributed almost equally throughout the phrases, an average of the ratios of the similar phonemes is determined and substituted for those phonemes. This phoneme matrix is similar to the probability matrix obtained from the text. Similarly the transition matrices corresponding to the phonemes at the rising and falling regions are also determined. A linear combination of these three matrices yield a single feature matrix. Entropy is now determined for each row in the matrix and feature entropy vector is obtained. The difference between a speaker's feature entropy vector and the common phoneme entropy vector generates a difference vector, which is stored as the reference template for that speaker. Similar templates are created for all the speakers. The test template is also created in the similar fashion and compared with the reference templates stored in the system. The reference template showing minimum deviation from the test template is identified.

7.3 SIMILARITY MEASURE METHOD

In this approach measurements are taken from the actual speech and the similarity between two patterns are determined. The phrases are segmented into phonemes and further to smaller segment of 600 samples. From this phoneme segments, feature measurements are extracted. The time measurements viz., individual phoneme times and two transition times for each phoneme, are determined during the automatic segmentation of the phrases. The other measurements are also taken from each individual phoneme segments. The final set of measurements used for identification is fixed by trial and error method based on the performance. Different combinations of the extracted features are tried and performance is evaluated. Each set of measurements give a matrix corresponding to the four phrases. Similar matrices are taken for all the speakers and the comparison is made based on a similarity measure. Each element in the matrix is compared with another matrix and weightage is given for every measurement if that falls within a particular limit. The limits for the measurements are also fixed by trial and error method. The weightages are based on the consistency of the measurements. The measurement showing maximum consistency gets more weightage and measurement showing minimum consistency gets least weightage. The weightages are so given that if all the measurements agree for an individual phoneme, the similarity factor is 1. In an ideal case the similarity factor for two similar phrase set is 16. The measurements contributing least to the performance are removed and other measurements are tried. Thus an effective measurement set is fixed after several trials. This gives a good performance.

7.4 CONCLUSION

Two approaches for speaker identification are explained in this chapter. These two approaches are rather simple and easy to implement.

CHAPTER 8

IMPLEMENTATION OF SPEAKER IDENTIFICATION ALGORITHMS

Chapter 8

IMPLEMENTATION OF SPEAKER IDENTIFICATION ALGORITHMS

The implementation of the recognition algorithm is discussed in this chapter. Care has been taken to make the system as simple as possible without sacrificing the performance.

8.1 TRAINING OF SPEAKERS

Very informal introduction is given to each speaker before the training session. They are briefed about the purpose of the training and demonstrated how to read out the phrases. Enough time is given to them to familiarize the phrase set. The speakers are asked to utter or read the phrases in their most natural way. They are also instructed not to speak either very slowly or very fast. Certain trials have to be repeated mainly to avoid overlap of words. Overlap and long delay between words created problems when segmenting the phrase. So these two situations are avoided. However, the speakers are instructed not to lose their naturalness in speech. In each training session three or four sets of phrase data are collected. Sessions are arranged on different days. Interval between some sessions is too long, whereas for some sessions it is short.

8.2 TEST CONDITIONS

Speech data are collected using the PC based speech digitizer. A dynamic microphone is used as input to the system. There are no constraints regarding the distance between the microphone and the speaker, gain of the input amplifier etc., to make the system work at a most natural environment. The speech data are collected from an ordinary A/C room. The system noise is adjusted to a minimum at the beginning of each session.

8.3 TYPICAL FUNCTIONING OF THE SYSTEM

Each phrase consisting of four words is uttered into the dynamic microphone. This speech is bandlimited to 4 KHz and sampled at 8 KHz. Samples are first stored in the system RAM and then transferred and stored as files on disks. This sampled data of a phrase is then fed to a boundary detection and segmentation algorithm.

The boundary detection and segmentation algorithm makes use of an energy envelope method. Energy of short duration segment (8 samples) is calculated and then the energy envelope for the total phrase is plotted. From this energy envelope, broad peaks are detected first. Here the assumption is that the voiced region has more energy than the unvoiced region and the main large peaks correspond to voiced sounds in that phrase. After detecting the peaks, each peak is assigned to respective words. Since the text is known a priori, the possible number of main peaks in a word can be easily assigned. Smaller peaks which are in the vicinity of the main peaks are also assigned to the same word. Inter peak distances and total word length are also checked while detecting the word boundary. The intra-word silence and inter word silence are also to be considered in boundary detection. Since words with intra-word silence are known a priori, care has been taken to skip over this silence and segment the whole word. For example, words like 'AFTER' and 'PARTY' have distinct long silence in them. This silence sometimes can be confused as inter word silence. So during segmentation this factor is also considered. This automatic segmentation algorithm worked with more than 90% accuracy. The complete algorithm is shown in Fig.8.1 (flow chart).

After the word boundary detection, maximum energy region within each word is detected. This region corresponds to the prominent voiced sound in that word. Since the recognition system extracts features of voiced sounds, this region has to be segmented. During segmentation of phonemes, the time measurements corresponding to words, phonemes and phrases are also taken. Transition from 20% of maximum to 80% of maximum

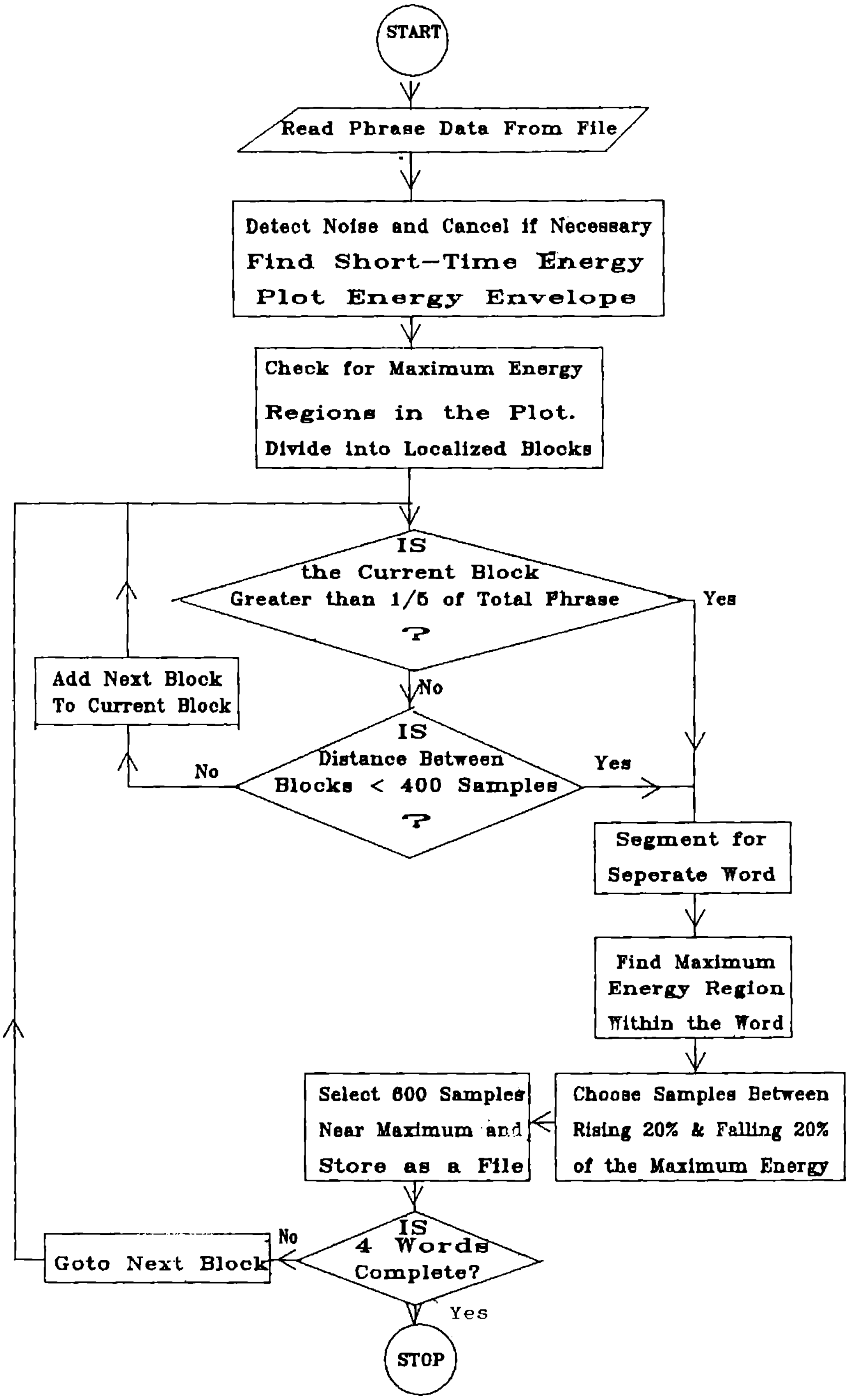


Fig.8.1 Flow Chart of Automatic Phoneme Segmentation.

at rising edge of the phoneme and 80% to 20% at the falling edge are taken as transition region within each phoneme. The samples between the 20% at rising region and 20% at the falling region are taken as the phoneme length. 600 samples from the maximum energy region of each word are stored as separate files for further feature extraction. The same process is repeated for all the four phrases.

Time normalized data files are also created from the same original phrase data. To generate time normalized data, all words are standardized to 4000 samples during segmentation. Words having more than 4000 samples are compressed and words with less than 4000 samples are expanded to 4000 samples. A linear interpolation method is made use of for this linear time normalization.

Let

N_o be the number of samples originally available for a word

N_f be the number of samples fixed as standard (4000)

T_s be the original sampling frequency, and

T_n be the new sampling frequency

Then

$$N_f * T_n = N_o * T_s$$

$$T_n = N_o * T_s/N_f$$

$$\text{Turn Ratio } Tr = (T_n - T_s)/T_s$$

During normalization, new samples are generated from the original samples using

$$S_n = S_i + (S_i - S_{i+1}) * Tr$$

where S_n is the new interpolated samples and S_i and S_{i+1} are the original samples.

During time normalization, the ratio of compression or expansion is also recorded, which is later used for determination of the original pitch period.

The experiment is conducted using both time-normalized and non-normalized data. The segmented files are processed for feature extraction. Among the selected features, only energy measurement requires amplitude normalization. Other measurements are not affected by the variation in the amplitude.

Pitch is determined using the symmetry check approach. In the case of time normalized data, the compression ratio is multiplied with the obtained pitch to get the actual pitch. The number of samples corresponding to the pitch period, which is called pitch number is taken as one measurement. The average symmetry coefficient, rms coefficient and rate of change coefficient are also extracted from this pitch measurement.

Zerocrossings in sample number corresponding to 4 times the pitch number are determined. Number of positive peaks within this sample length is also taken as a measurement. Samples are now normalized in amplitude and energy of the segment (600 samples) is determined. The Auto Correlation Function of 512 points from the segment is determined using an FFT method. From the ACF, slope between two main peaks, distance between two peaks and number of smaller peaks between two main peaks are calculated. Thus from each individual phoneme, 10 different measurements are extracted and a 16x10 element matrix is obtained corresponding to a phrase set.

A knock out method is employed for feature selection based on the performance of the features. Time measurements and feature measurements are both used in the system and their performance is evaluated. Based on this evaluation, a final set of features is selected.

In the experiment 6 speakers (5 males and 1 female) have participated and for each speaker 10 trials are taken at different sessions. Interval between sessions varied from speaker to speaker. Table 8.1 shows the time interval between sessions for each speaker and the number of trials in each session.

Table 8.1

Interval between each session

Sessions	Aj	BA	JK	AK	Na	TT
I-II	4.75 Months	5 Months	2 Days	13 Days	4.5 Months	2.75 Months
II-III	5.25 Months	5 Months	5 Months	5 Months	18 Days	2 Months
I-III	10 Months	10 Months	5 Months	5.5 Months	5 Months	4.75 Months

Two approaches are employed for speaker identification. In the first method a Fixed Text Phoneme Model (FTPM) approach is used. From the text of the phrases the probability of occurrence of each phoneme in the phrase set is determined. Entropy is calculated for each phrase and a phoneme entropy is formed from the phrase set. The individual phoneme time stored during the segmentation of phrases are taken and ratio with the corresponding phrase is determined. Average of the ratio of similar phonemes are taken and substituted for that phonemes. From the resulting matrix, entropy is calculated for each row. Similarly the transition timings stored during segmentation are also used to find entropy. The phoneme ratio entropy vector is added to the transition entropy vectors and the vector obtained is a feature entropy vector. The difference between each individual's feature entropy vector and the phoneme entropy vector from the text gives a difference vector or a distortion vector. This distortion vector is stored as the reference template for each speaker.

The identification in the second approach is based on a similarity measurement. Each measurement is compared and if it falls within a limit, a particular weightage is assigned to it. The weightages are fixed by trial and error based on the consistency of the measure. This weightage assignment is repeated for all the measurements and phonemes. Finally a similarity measure is obtained for all the speakers. The reference template corresponding to the maximum similarity measure is identified for the speaker.

Identification of speakers are attempted in two ways. In the first way, one of the trials of a particular speaker is taken as the test template and the remaining trials are averaged and kept as the reference template. All the trials of other speakers are also averaged and kept as reference template of each speaker. Then a comparison is made between the test template and the reference template. These tests are called total session tests.

Speaker identification is attempted in another way, in which the trials are considered session-wise. One of the trials in a particular session is taken as the test template and average of other trials in that particular session is taken as the reference template for a speaker. This is called separate session test. This will enable study of the effects of time variation on the speaker identification.

8.4 EXPERIMENT AND RESULTS

8.4.1 Results of Fixed Text Phoneme Model (FTPM) approach

The universally accepted confusion matrix method is used to display results. The speaker names have been abbreviated to initials. For the FTPM approach using the timings and entropy measures, confusion matrices are obtained as shown in Figs.8.2 and 8.3.

		Speaker Identified as						
		Aj	BA	JK	AK	Na	TT	Error
Token (Speaker) Given as	Aj	1	2	3	2	2	0	(9)
	BA	0	5	0	1	4	0	(5)
	JK	3	0	6	1	0	0	(4)
	AK	1	4	0	3	2	0	(7)
	Na	0	0	0	2	8	0	(2)
	TT	0	0	1	0	0	9	(1)
						Total Error		(28)

Fig.8.2 Confusion Matrix in total session for FTPM

		Speaker Identified as						
		Aj	BA	JK	AK	Na	TT	Error
Token (Speaker) Given as	Aj	6	2	1	1	0	0	(4)
	BA	0	5	0	1	4	0	(5)
	JK	3	0	6	1	0	0	(4)
	AK	2	0	2	5	0	1	(5)
	Na	0	0	0	2	8	0	(2)
	TT	0	0	1	0	0	9	(1)
						Total Error		(21)

Fig.8.3 Confusion Matrix in separate session for FTPM

The error performance for each speaker in the FTPM method is shown in the Fig.8.4. The enhancement in performance after taking separate session is insignificant in this approach. The average identification accuracy in total sessions is only 53.33%. When taken as separate sessions, this has improved to 65%, which is not very encouraging result. It is observed that, when separate sessions are considered, error percentages of only two speakers are changed. For the remaining 4 speakers no change in performance is noticed. Individual errors in this approach is also very high. Though the misclassification between female and male speakers is negligible, the misclassification among male speakers is very high. Only TT and Na have got comparatively lesser error rates.

Speakers	Aj	BA	JK	AK	Na	TT	Total
Total Sessions	9	5	4	7	2	1	28
	90%	50%	40%	70%	20%	10%	46.67%
Separate Sessions	4	5	4	5	2	1	21
	40%	50%	40%	50%	20%	10%	35%

Fig.8.4 Speaker wise error performance in FTPM approach

8.4.2 Results of Similarity Measure approach

In this section, results of the experiment with similarity measure using different feature measurements are presented. Based on the results, a knock out method has been adopted to select more appropriate feature measurements. In certain tests, only the weightages are altered and the performance is checked. Following are the confusion matrices obtained for each test using a set of feature measurements.

Test 1

In the first test, only four feature measurements are considered. These are, pitch number from symmetry check method, number of zerocrossings in samples equal to 4 times the pitch number, energy of the segment and the distance between two main peaks from the Auto Correlation Function. All these measurements are given equal weightage (0.25). Fig.8.5 shows the confusion matrix for this test in total sessions. Fig.8.6 shows the confusion matrix at separate sessions.

Test 2

In this test, phoneme timings and transition timings are also incorporated. Here along with zerocrossings, pitch and energy, entropies corresponding to each phoneme time and transitions in the rising and falling regions within the phoneme are also used. The weightages given to the measurements are as follows:

Entropy of phoneme	0.15
Entropy of transition at the rising region	0.15
Entropy of transition at the falling region	0.15
Number of zerocrossings	0.15
Energy	0.20
Pitch number	0.20

Figs.8.7 and 8.8 show the confusion matrices of total sessions and separate session respectively.

Test 3

The weightages are changed for the same set of measurements as shown below and the test is repeated.

Entropy of phonemes	0.10
Entropy of transition in the rising region	0.20
Entropy of transition in the falling region	0.10
Number of zerocrossings	0.15
Energy	0.20
Pitch number	0.25

		Speaker Identified as						
		Aj	**BA**	**JK**	**AK**	**Na**	**TT**	**Error**
Token (Speaker) Given as	**Aj**	1	2	1	2	4	0	(9)
	BA	1	7	0	0	2	0	(3)
	JK	1	3	3	1	2	0	(7)
	AK	0	0	0	8	2	0	(2)
	Na	1	0	0	0	9	0	(1)
	TT	0	0	0	0	0	10	(0)
						Total Error		(22)

Fig.8.5 Confusion Matrix for total session

		Speaker Identified as						
		Aj	**BA**	**JK**	**AK**	**Na**	**TT**	**Error**
Token (Speaker) Given as	**Aj**	9	1	0	0	0	0	(1)
	BA	2	7	0	0	1	0	(3)
	JK	0	0	10	0	0	0	(0)
	AK	0	0	0	10	0	0	(0)
	Na	1	0	0	0	9	0	(1)
	TT	0	0	0	0	0	10	(0)
						Total Error		(5)

Fig.8.6 Confusion Matrix for separate session

		Speaker Identified as						
		Aj	BA	JK	AK	Na	TT	Error
Token (Speaker) Given as	Aj	1	0	4	1	4	0	(9)
	BA	2	6	0	0	2	0	(4)
	JK	0	3	2	0	5	0	(8)
	AK	0	0	0	6	4	0	(4)
	Na	1	0	0	1	8	0	(2)
	TT	1	0	0	0	0	9	(1)
						Total Error		(28)

Fig.8.7 Confusion Matrix for total sessions

		Speaker Identified as						
		Aj	BA	JK	AK	Na	TT	Error
Token (Speaker) Given as	Aj	10	0	0	0	0	0	(0)
	BA	4	5	0	0	1	0	(5)
	JK	0	0	10	0	0	0	(0)
	AK	0	0	0	10	0	0	(0)
	Na	1	0	0	0	9	0	(1)
	TT	0	0	0	0	0	10	(0)
						Total Error		(6)

Fig.8.8 Confusion Matrix for separate sessions

The confusion matrices for total sessions and separate sessions are shown in Figs.8.9 and 8.10.

Test 4

The weightages of the entropies for rising and falling transitions are interchanged and the confusion matrices are shown in Figs.8.11 and 8.12.

Test 5

In this test the entropy measurements are knocked out and other feature measurements are included. The weightages are also changed in this case. Consistency of measures are mainly taken into consideration. Measures showing maximum consistency are provided with more weightage and measures showing minimum consistency are provided with less weightages. Weightages are as shown below:

Pitch number	0.25
RMS coefficient	0.08
Number of zerocrossings	0.17
Distance between peaks in ACF	0.15
Slope between two peaks	0.05
Average coefficient	0.02
Rate of change coefficient	0.03

		Speaker Identified as						
		Aj	BA	JK	AK	Na	TT	Error
Token (Speaker) Given as	Aj	1	0	4	0	5	0	(9)
	BA	2	6	0	0	2	0	(4)
	JK	0	3	2	0	5	0	(8)
	AK	0	0	0	6	4	0	(4)
	Na	1	0	0	1	8	0	(2)
	TT	0	0	0	0	0	10	(0)
						Total Error		(27)

Fig.8.9 Confusion Matrix for total sessions

		Speaker Identified as						
		Aj	BA	JK	AK	Na	TT	Error
Token (Speaker) Given as	Aj	10	0	0	0	0	0	(0)
	BA	4	5	0	0	1	0	(5)
	JK	0	0	9	0	1	0	(1)
	AK	0	0	0	10	0	0	(0)
	Na	1	0	0	0	9	0	(1)
	TT	0	0	0	0	0	10	(0)
						Total Error		(7)

Fig.8.10 Confusion Matrix for separate sessions

					Speaker Identified as			
		Aj	BA	JK	AK	Na	TT	Error
	Aj	1	0	4	0	5	0	(9)
	BA	2	6	0	0	2	0	(4)
Token (Speaker) Given as	JK	0	3	2	0	5	0	(8)
	AK	0	0	0	7	3	0	(3)
	Na	1	0	0	1	8	0	(2)
	TT	0	0	0	0	0	10	(0)
						Total Error		(26)

Fig.8.11 Confusion Matrix for total sessions

					Speaker Identifies as			
		Aj	BA	JK	AK	Na	TT	Error
	Aj	10	0	0	0	0	0	(0)
	BA	4	5	0	0	1	0	(5)
Token (Speaker) Given as	JK	0	0	9	0	1	0	(1)
	AK	0	0	0	10	0	0	(0)
	Na	1	0	0	0	9	0	(1)
	TT	0	0	0	0	0	10	(0)
						Total Error		(7)

Fig.8.12 Confusion Matrix for separate sessions

Number of positive peaks	0.05
Number of peaks between two main peaks in ACF	0.05
Energy of the segment	0.15

This test is repeated for two sets of data, namely (i) Time normalized data, and (ii) Non normalized data. The confusion matrices are shown in Figs.8.13, 8.14, 8.15 and 8.16.

The results are summarized in Table 8.2.

Table 8.2

Error in different tests for similarity measure approach

Error in	Test 1	Test 2	Test 3	Test 4	Test 5 TN	 Normal
Total Session	22/60	28/60	27/60	26/60	23/60	21/60
Separate Session	5/60	6/60	7/60	7/60	11/60	4/60

From the above confusion matrices and result summary, it is seen that the performance of the system is rather poor when total sessions are considered for reference template. The performance is better when test and reference templates are taken

		Speaker Identifies as						
		Aj	**BA**	**JK**	**AK**	**Na**	**TT**	**Error**
Token (Speaker) Given as	**Aj**	4	0	0	0	6	0	(6)
	BA	2	6	0	0	2	0	(4)
	JK	1	1	7	0	1	0	(3)
	AK	0	1	0	2	4	3	(8)
	Na	0	1	0	0	9	0	(1)
	TT	0	0	0	0	1	9	(1)
						Total Error		(23)

Fig.8.13 Confusion Matrix of TN Data in total sessions

		Speaker Identified as						
		Aj	**BA**	**JK**	**AK**	**Na**	**TT**	**Error**
Token (Speaker) Given as	**Aj**	7	0	0	0	3	0	(3)
	BA	2	8	0	0	0	0	(2)
	JK	2	1	6	0	1	0	(4)
	AK	0	0	0	8	2	0	(2)
	Na	0	0	0	0	10	0	(0)
	TT	0	0	0	0	0	10	(0)
						Total Error		(11)

Fig.8.14 Confusion Matrix of TN Data in separate sessions

		Speaker Identified as						
		Aj	BA	JK	AK	Na	TT	Error
Token (Speaker) Given as	Aj	1	2	1	1	4	1	(9)
	BA	1	8	0	0	1	0	(2)
	JK	2	3	3	1	1	0	(7)
	AK	0	2	0	8	0	0	(2)
	Na	1	0	0	0	9	0	(1)
	TT	0	0	0	0	0	10	(0)
						Total Error		(21)

Fig.8.15 Confusion Matrix of Real Data in total sessions

		Speaker Identified as						
		Aj	BA	JK	AK	Na	TT	Error
Token (Speaker) Given as	Aj	9	1	0	0	0	0	(1)
	BA	1	9	0	0	0	0	(1)
	JK	0	0	9	1	0	0	(1)
	AK	0	0	0	10	0	0	(0)
	Na	1	0	0	0	9	0	(1)
	TT	0	0	0	0	0	10	(0)
						Total Error		(4)

Fig.8.16 Confusion Matrix of Real Data in separate sessions

from the same session. One reason for this performance deterioration is the effect of time on each speaker. Longer the interval between sessions, worse the identification performance. This variation due to time interval is an established fact [36]. In order to avoid this, whenever the system makes a correct identification, the reference template has to be updated with the test template which was correctly recognized.

Another reason is the change of systems during data collection. The speech digitizer circuit is used with different PCs in some sessions. Since the sampling frequency is generated by software in the digitizer using software delay, this delay is system dependent. Whenever, the system is changed, the system clock is different for that system while the software delay routine is the same and hence an approximation is to be used for each system depending on the system clock frequency. This approximation also has affected the data in each session. The data taken in the same system with small time interval between sessions show a better performance.

One of the speakers in the above experiment is a female speaker. The confusion between the female and male speakers is negligible. Mostly the male speakers are confused with other male speakers only. In all the tests, the female speaker was rarely misidentified. Thus the features selected clearly separate the male and female speakers.

It is also seen that the time normalization of the data makes the performance of the system poor. Better performance is obtained for the original data (without time normalization). From the above tests, the measurements that contribute less to the performance of the system are knocked out. The interesting point is that, the use of entropy in similarity measure method does neither deteriorate nor enhance the performance. Hence, it was felt that the simpler of the techniques can be used and the features selected have been listed below.

1. Pitch number by symmetry check.
2. Average coefficient
3. RMS coefficient
4. Rate of change coefficient
5. Number of zerocrossings in the segment corresponding to 4 times the pitch period
6. Number of positive peaks in the same segment
7. Energy of the segment
8. Distance between two peaks in the ACF
9. Slope between two peaks in ACF
10. Number of smaller peaks between the two peaks.

Using this set of feature measurements, with weightages as discussed in the experiment, an identification accuracy of 65% in total sessions and 93% in separate sessions are obtained.

8.5 DISCUSSION

The results show a good percentage identification, comparable to many other existing techniques. The system does not require special chips (hardware) and employs only very simple methods. The number of speakers included is six. This number might appear small. However, the technique used here does not really depend on number of speakers. More speakers can be added provided the templates are obtained and updated during training sessions. The system is not real time. After the speaker utters the phrases, the similarity measure is to be evaluated off line and a finite time is taken by the system to do this. Since most of the routines are software in high level language, except for the speech digitizer, it is entirely portable. Only a computer with reasonable size of memory is required.

There is not much use in trying to evaluate the speed and response. The system can be implemented on a PC/XT or PC/AT. A PC/AT with 386/387 (this was used in the laboratory) at 20 MHz may provide very good speed, while on 8088 machine it appears slow. All the parameters required for the analysis are extracted from the digitized sample set and no additional input is needed by the system.

Regarding the performance, another observation is that the system does not respond favourably to a speaker not included in the sample set. It neither misclassifies nor identify. It simply abandons the test. From this, artifacts like alarms can be raised. Even amongst the set of speakers included, only the machine 'knows' the consistent misclassification. For example, Aj and Na form a one way pair. Aj has been consistently misclassified as Na, while the reverse is not true. Though an attempt can be made to determine the main reason for this consistency in misclassification, it serves no purpose. The features which have been specially chosen and weighted, are simple enough and common enough. It is only a judicious linear combination of these features that is providing the dissimilarity vectors for classification. Also, the weightages are chosen to suit the general background of the speakers. The whole set of speakers are from one region and more or less adopt the same type of pronounciation. Further, the phrases though not nonsense words, are practically 'meaningless' to the speakers. This makes all samples as nearly similar as possible. The feature set selects a highly individual combination and is not amenable to simple analysis.

8.6 INTERPRETATION OF RESULTS

Over total sessions the speaker wise error committed is shown in Fig.8.17. The separate session wise results are shown in Fig.8.18. It can be seen that Aj consistently shows error and this is rather high. Na on the other hand is a 'good' speaker and shows only error of the order of 10-20%. As mentioned earlier, TT of course shows negligible error, approaching zero, being a female voice. Session wise, the errors are not that alarming and the maximum error is 40% in rare case.

Not going into the reasons for why Aj and JK contribute such high errors, if these two are eliminated from the speaker set (i.e., omit row 1 and row 3 separately or together) the percentage error on an average falls from 40.83 to 20.83. This is shown in Fig.8.19. This shows that the system has a poor adaptation vector for Aj and JK but pretty good for the rest. The performance can be improved by an additional input like for example, a different phrase. As per misidentification, which is shown in Fig.8.20 and Fig.8.21. Misidentification is inconsequential since in a large number of applications, a person asking for access, claims he is X. If he is not X, irrespective of what he is identified as, he is denied access. Nevertheless a look at the results is very interesting. A misidentified as B does not mean B is misidentified as A in all cases. Large misidentification arises only in the case of Aj who is mistaken for Na 45% of the trials. At the other extreme Na is misidentified as Aj only 8.75% of the trials. Similarly BA is never misidentified as JK while JK is mistaken for BA more than 20% of the time. Some examples of this irreversible relationship is depicted in Fig.8.22. This nonreciprocal relationship can only be explained on the basis of the heuristic model. The system is only identifying the learned aspects as modified by the individual. It can be established that under test conditions there is strong correlation between the human recognizer and the system developed in these aspects. But the experiment to prove this is not completely unbiased as all the subjects know each other over a few years. However, the model presented in chapter 6 appears to be as good or as bad as other similar models presented in literature.

The percentage of misidentification can be reduced to values as obtained in the case of TT and Na etc. by further tuning the system to strictly incorporate the learning-feedback path by collecting further information, which will have to be personal. This will enhance both the complexity and cost of the system, thus defeating the purpose of designing and building an inexpensive, PC based sample identification system.

Test	1		2		3		4		5a		5b	
Speaker	Error		Error		Error		Error		Error		Error	
	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
Aj	9	90%	9	90%	9	90%	9	90%	6	60%	9	90%
BA	3	30%	4	40%	4	40%	4	40%	4	40%	2	20%
JK	7	70%	8	80%	8	80%	8	80%	3	30%	7	70%
AK	2	20%	4	40%	4	40%	3	30%	8	80%	2	20%
Na	1	10%	2	20%	2	20%	2	20%	1	10%	1	10%
TT	0	0%	1	10%	0	0%	0	0%	0	0%	0	0%

Fig.8.17 Speaker wise error (Total Session)

Test	1		2		3		4		5a		5b	
	Error		Error		Error		Error		Error		Error	
Speaker	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
Aj	1	10%	0	0%	0	0%	0	0%	3	30%	1	10%
BA	3	30%	5	50%	5	50%	5	50%	2	20%	1	10%
JK	0	0%	0	0%	1	10%	1	10%	4	40%	1	10%
AK	0	0%	0	0%	0	0%	0	0%	2	20%	0	0%
Na	1	10%	1	10%	1	10%	1	10%	0	0%	1	10%
TT	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%

Fig.8.18 Speaker wise error (Separate Sessions)

	% of error for all speakers	% of error eliminating speakers with poor performance	Speakers eliminated
Test 1	36.67%	15%	Aj and JK
Test 2	46.67%	27.5%	Aj and JK
Test 3	45%	25%	Aj and JK
Test 4	43.33%	22.5%	Aj and JK
Test 5a	38.33%	22.5%	Aj and JK
Test 5b	35%	12.5%	Aj and JK
Average Error %	40.83%	20.83%	

Fig.8.19 Error performance by eliminating speakers

	Aj	BA	JK	AK	Na	TT	Total No.	Error %
Aj	-	4	14	4	28	1	51	85%
BA	10	-	0	0	11	1	21	35%
JK	4	16	-	2	19	0	41	68.33%
AK	0	3	0	-	17	3	23	38.33%
Na	5	1	0	3	-	0	9	15%
TT	1	0	0	0	1	-	2	3.3%

Fig.8.20 Misclassification and error for each speaker (Total)

	Aj	BA	JK	AK	Na	TT	Total No.	Error %
Aj	-	2	0	0	3	0	5	8.33%
BA	17	-	0	0	4	0	21	35%
JK	2	1	-	1	3	0	7	11.7%
AK	0	0	0	-	2	0	2	3.3%
Na	5	0	0	0	-	0	5	8.33%
TT	0	0	0	0	0	-	0	0%

Fig. 8.21 Misclassification and error for each speaker (Session)

Total sessions (in all the tests)

Aj identified as Na	46.67%	Aj identified as JK	23.33%
Na identified as Aj	8.33%	JK identified as Aj	6.67%
Aj identified as BA	6.67%	JK identified as BA	26.67%
BA identified as Aj	16.67%	BA identified as JK	0%
JK identified as Na	31.67%	AK identified as Na	28.67%
Na identified as JK	0%	Na identified as AK	5%
BA identified as Na	18.33%	Na identified as BA	1.67%

Fig.8.22 Example for non-reciprocal relation between speakers

CHAPTER 9

CONCLUSIONS

Chapter 9

CONCLUSIONS

The thesis presents work done by the author in the field of Speaker Identification. As detailed in chapter 2 there have been a large number of research workers interested in this problem. The author is benefitted greatly by trying to understand the reported literature. The methods are presented for speaker identification. The classical model for speech production has been used in many approaches to the problem. However, it can be noted (as shown in chapter 2) that the concept of using certain 'characteristics' (later called as features) existed even before the model was presented and universally accepted. Some of these features exist even now in the later works. Altogether, the differences between two speakers have been presented as differences between two networks. The intuitive feeling one gets at this stage is that the learning process, acquiring accents, securing imperfections etc., have not been reflected in the model at all. For instance, it was presented (in chapter 6) that children and pets learn to discriminate between voices. The capacity to discriminate grows with age and expanding areas of activity. It was the intention of the author to include this aspect in this study. This was possible only by proposing a model different from classical model. Also, the model has to depend on a knowledge base to enable gradual learning. The model, at the same time, has to retain the 'network' properties of the classical model not merely because the model has been extensively and successfully used, but also because certain features are dependent on 'filter' parameters and do not entirely need the knowledge base. It was with this aim that the FTPM model was postulated. Defining and measuring the parameters of the knowledge base turned out to be a tough problem. It was felt that certain irrelevance will creep into the work if greater attention is devoted to the knowledge base and its parameters. Indeed the work would possibly look like one on knowledge engineering and will definitely fall beyond the capability of the author. The model presented finally in chapter 6 is chosen after trying a number of such models proposed and discarded either because all the aspects of learning-knowledge base interaction could not be presented effectively or due to extreme complexity. However, the final model presented is not perfect since all the aspects presented in the heuristics have not found a place in it. Simplicity of implementation was another constraint which led to the choice of the model and methods. As a matter of fact, it was strongly felt that the implementation must be PC oriented. Therefore, a suitable, simple and reasonably good measures have to be devised. The measurements were hence localized on entropy and information content. The measurements proved to be multi-dimensional and the reduction was essential. Though certain similarity exist with Markov Model and particularly Hidden Markov Model, the difference lies in the definition of probabilities of occurrence of the phonemes. The FTPM model however, did not provide results from the second and near conventional method. The reason for this is due to the fact that the entropy measure defined and used is probably as good or as bad as a speaker dependent measure like Pitch, ACF, Energy etc. As was pointed out in the chapter 8, the performance did not deteriorate too. The model therefore, needs certain changes to include measurable, speaker dependent entropies. The changes are mainly in defining the rate of change (corresponding to transition probabilities in HMM) more precisely but without losing the speaker dependence. This is very difficult to achieve. One has to precisely define and mark a unit within the speech samples. The author has chosen a set of phrases which try to reflect as far as possible the naturalness of the speaker without constraining the actual mode of delivery. Thus it becomes a trial and error method and has taken many trials and time. As a matter of fact, even the choice of phrases took considerable time and as pointed out in chapter 4, some changes had to be made from a set of selected phrases, after study. These two areas, it is felt, can be further explored. Though the FTPM and the new knowledge base model did not show remarkable results, there is potential in it to do better than the other models. A fine tuning of the learning parameters and the way of measurements will improve the performance. Similarly the phrase set chosen is in English. The subjects are from a background in different language but all belonged to one region. It is likely that if the phrases were in the language where the influence of learning is much more pronounced, for example Malayalam, the performance would have been much better. The reason for this is the fact that English is acquired by all the subjects as a spoken language in school and college thus providing similarity in learning for all the subjects. The error percentages and cross-identification are likely to improve. However, a phrase set independent of drawbacks is almost always impossible to choose. Choice of a phrase set, as of is not based on rigid mathematics. Further work will be interesting.

An attempt was also made to use the concept of eigen frequencies. However, this was not found universal and hence could not be incorporated into the model. The possibility of including this is an open area for further work.

Regarding the system fabricated as front end, it was attempted to duplicate as nearly as possible, toll quality speech. The aim can be furthered to consider systems accessible via telephone.

With regard to the set of new features defined, for example, Symmetry Check method for the determination of pitch, it was found very effective and consistent. In fact all measurements on voiced portions were found to be consistent not only by this author but also by many others. Why an unvoiced sound should not show consistency in some measure is not known. The author has tried to utilize some unvoiced sound portions also in this work and the performance is not unduly deteriorated. It would be interesting to define some measures for unvoiced sounds, at least a selected few, and explain the lack of consistency in the measures or otherwise.

Yet another interesting aspect is cross-identification. Even in the simple system presented in this work there is a one way consistency of cross-identification. This suggests a possible similarity in the learning environments of the two speakers as measured by the system. However, the lack of reciprocity suggests that the measures are measuring only one aspect of the feature. It is likely that this vector (space) is anisotropic. Many analyses were attempted but were not acceptable. This aspect of non-reciprocity in identification is also an open problem for further work.

The thesis attempts therefore, to present a low cost PC based speaker identification system and in implementing this a few new concepts have been introduced. The performance is not exemplary but neither is it too poor to be discarded.

For a simple system, indeed the performance is comparable to many more sophisticated systems. It may appear that too many features have been selected, but, by a judicious knock-out method and by monitoring the performance it was possible to maintain a good score. However, the knock-out procedure is one of trial and error. There has not been a rigorous analysis of the technique in literature and from the author's experience it is felt that, it is not possible to provide a rigorous analysis or a procedure for this. It is for this reason that one cannot claim that the features selected by the author are the optimum features. It would an interesting further work if optimality in feature selection is defined and proved for a set of features. In the present work the single set of measurements on the samples yields all the features defined and used in this work. To this extent, that is minimizing the data base and computations, this work is near optimal, eventhough the system is not in real time.

APPENDIX A

SYMMETRY CHECK METHOD FOR PITCH DETECTION

APPENDIX A

SYMMETRY CHECK METHOD FOR PITCH DETECTION

A.1 INTRODUCTION

Pitch or fundamental frequency is a vital feature in speech processing. It is popularly used in most of the speech recognition as well as speaker recognition methods. The main reason is that pitch is dependent on both speech and speaker. The essential principle of any pitch detection method is to determine the periodicity observed in the voiced sounds. Many standard methods are available for pitch detection. Most of them are employing very complex methods.

In the present method a simple algorithm for pitch detection is presented. This method is based on a purely time domain window operation. The main advantage is that the algorithm is mathematically simple and easy to implement in any small systems.

A.2 ALGORITHM

The principle of the algorithm is discussed in chapter 5. The flow of the algorithm is as follows. Data samples are read into the files first. To start with checking, an initial window of 32 samples is considered. The logic behind this is that pitch periods will not be less than 2 milliseconds. If the sampling time of speech is 125 microseconds, then the minimum size of the window should be more than 32 samples. As a first step, correlation coefficient between two halves of the samples in the window is determined. If the correlation coefficient is less than 0.75 then the window size is increased by adding 8 samples and again correlation coefficient is checked. This process is repeated till window shows a correlation coefficient greater than 0.9. There are three possible outputs at this stage; procedure in each case is explained in the following paragraphs.

Case 1: **Correlation coefficient is greater than 0.99**

On this condition, the window is considered to be completely symmetric and the period corresponding to half the number of samples in the window is taken for pitch period without checking any other measurements.

Case 2: **Correlation coefficient between 0.9 and 0.99**

At this instance, the three other symmetry coefficients and total symmetry are determined using Eqns.5.2, 5.3, 5.4 and 5.5. If the total symmetry is less than 0.66 then the corresponding window is considered symmetric and pitch period is determined as before.

Case 3: **Correlation coefficient is between 0.9 and 0.75**

This condition means that the window is not completely symmetric and it needs some alteration. As a first step, the window size is reduced by 4 samples and the symmetry is checked as explained earlier. If the window is giving a correlation coefficient between 0.9 and 0.75 then again the window size is reduced by 2 samples and same steps are repeated to determine the symmetry. Still if it is not showing symmetry then the window is augmented by another 8 samples and symmetry check is performed.

If a segment showscorrelation coefficient greater than 0.9 and does not show symmetry in other measurements, then 8 samples are added to the window and the procedure is repeated. In this manner one can adjust the window size by expanding and reducing the size and exactly locate the periodicity of a speech segment. The result is confirmed only after consistency check. This is done by shifting the window through 32 samples in the segment and reiterating the procedure.

The error possibility in this method is either showing a periodicity in half of the actual pitch period or in double the pitch. In either case, the consistency check eliminates the error possibility. A tolerance of 2 samples are allowed while checking the consistency. The maximum window size is restricted to

320 samples corresponding to 40 milliseconds. This is because, 50 Hz (20 milliseconds) is the minimum fundamental frequency in normal cases. If the algorithm cannot find a symmetry in this maximum window size, the window is reset to initial 32 samples and shifted with 32 samples through the segment. Then the whole scheme is repeated until it checks the complete segment.

A.3 EXPERIMENT AND RESULTS

The experiment is carried out in an IBM compatible PC/AT. The speech data are digitized with the speech digitizer described in chapter 3 which is interfaced to this PC, sampling frequency is set to 8 KHz. A set of 30 words in 'Malayalam' language is digitized and stored as separate files. The words chosen for the experiment have /CVC/ and /CVCV/ structure. The same set of words are spoken by two speakers. The voiced region of the words are segmented manually and used for the experiment. Pitch is determined in two methods namely, Cepstrum method (standard) and Symmetry check method (new) and results are compared. Cepstrum is determined by FFT method. Table A.1 shows the some of the comparative results from these two methods.

From the table it is observed that in many segments the results agree and in rare cases there is a maximum difference of 3 samples (375 microseconds). Checking the samples manually it is found that the symmetry check method gives more closer pitch periods. The cepstrum method always take the same amount of time in a fixed sample length, whereas in symmetry check approach the computation time vary from segment. Even in the worst case, the time taken for pitch determination in symmetry check method is less than that of cepstrum method.

Table A.1

Comparison of the pitch period in Symmetry and Cepstrum methods

WORDS	SPEAKER I			SPEAKER II		
	Sym.	Cpt.	Dif.	Sym.	Cpt.	Dif.
K/A/L	47	46	1	49	50	1
P/A/L	47	48	1	49	51	2
V/A/L	47	47	0	47	49	2
T/I/T	37	39	2	44	46	2
P/I/P	39	40	1	47	48	1
V/I/V	38	40	2	42	44	2
M/I/M	37	39	2	41	42	1
K/U/K	33	34	1	43	42	1
P/A/S	47	48	1	43	45	2
V/A/S	47	47	0	51	52	1
M/A/S	47	47	0	51	54	3
T/I/S	42	43	1	45	47	2
P/I/S	43	45	2	48	49	1
T/U/S	44	42	2	49	48	1
V/U/S	41	41	0	46	46	0
M/U/S	43	45	2	50	50	0

APPENDIX B

USE OF ENERGY ENVELOPE AND ZEROCROSSING RATE IN ISOLATED DIGIT RECOGNITION

APPENDIX B

USE OF ENERGY ENVELOPE AND ZEROCROSSING RATE IN ISOLATED DIGIT RECOGNITION

B.1 INTRODUCTION

Digit recognition is a comparatively easier task among the speech recognition problems. Isolated digit recognition does not need much complicated pattern recognition approaches, due to its limited vocabulary size. A simple approach for the isolated digit recognition for a set of trained speakers has been discussed in the section. Zerocrossing rate and energy envelope are the features used for classification. A tree structure has been implemented for decision-making.

B.2 SYSTEM DESCRIPTION AND DATA ANALYSIS

A 4 KHz band limited speech signal is sampled at a frequency rate of 8 KHz. The analog speech is digitized using a 12 bit speech digitizer explained in chapter 3. Data for each digit are stored in separate files for analysis.

End points of each digit is detected manually. This is possible by visually examining the plots of data and also listening back the reconstructed samples with the aid of speech digitizer. The number of samples varied for different recording sessions and speakers. A time normalization is also implemented in the system. Number of sample in each data file is fixed as 6000. Using linear interpolation, new samples are generated as follows:

Let N_o be the number of original samples available for a digit.

N_f be the number of samples fixed (6000).

T_s be the original sampling time (125 micro seconds)

T_n be the new sampling time.

Then

$$N_f * T_n = N_o * T_s \tag{B.1}$$

$$T_n = \frac{N_o * T_s}{N_f} \tag{B.2}$$

$$\text{Turn ratio } T_r = \frac{(T_n - T_s)}{T_s} \tag{B.3}$$

If S_1 and S_2 are the two consecutive original samples, then the new sample,

$$S_n = S_1 + (S_1 - S_2) * T_r \tag{B.4}$$

These newly generated speech samples are segmented into 8 milliseconds and zerocrossing rate energy for each segment are determined. Energy envelope is obtained by plotting these individual segment energy. Then their envelope is smoothened. Typical energy envelope for different digits are shown in Fig.B.1.

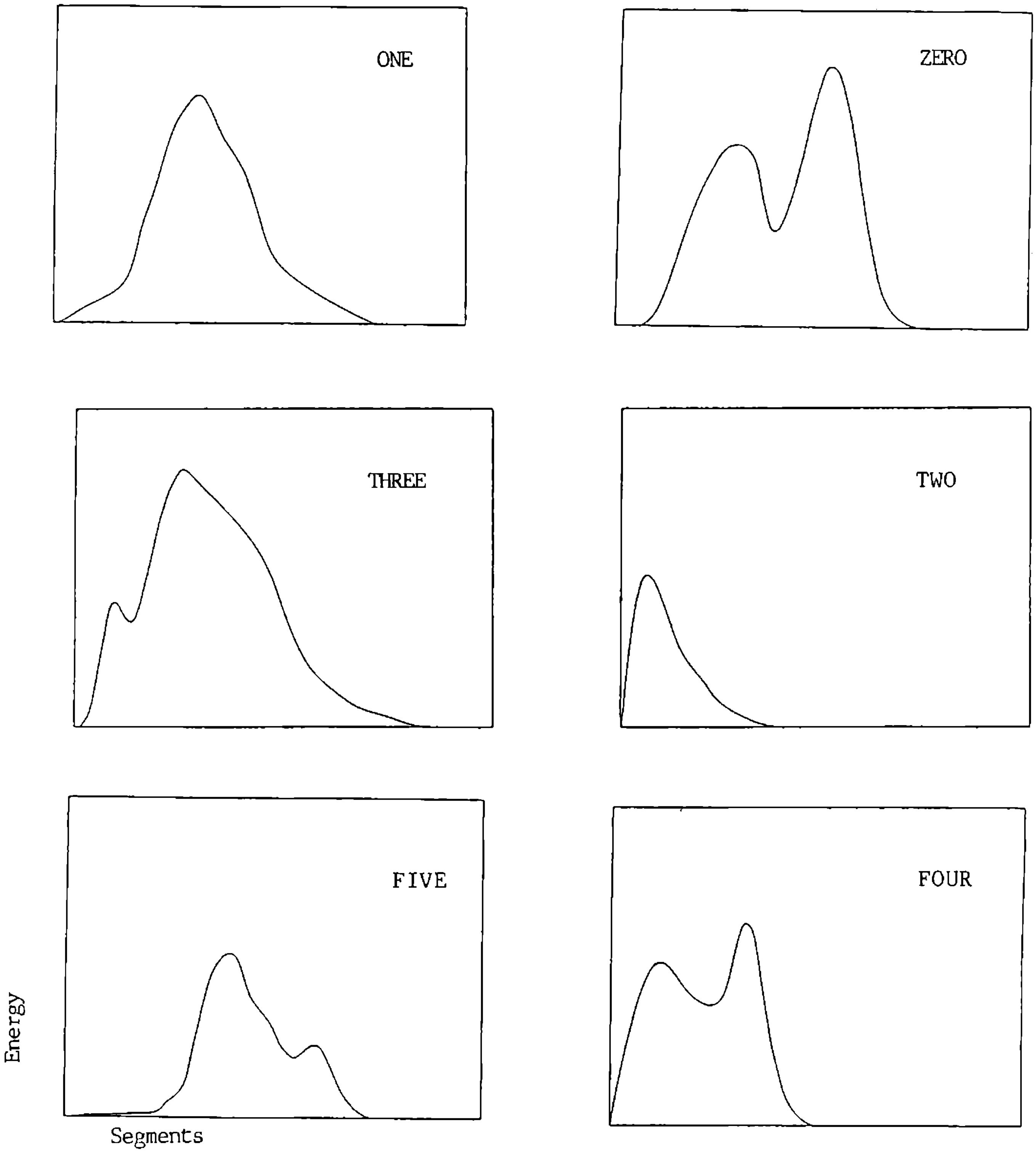


Fig.B.1 Typical Energy Envelope of Digits.

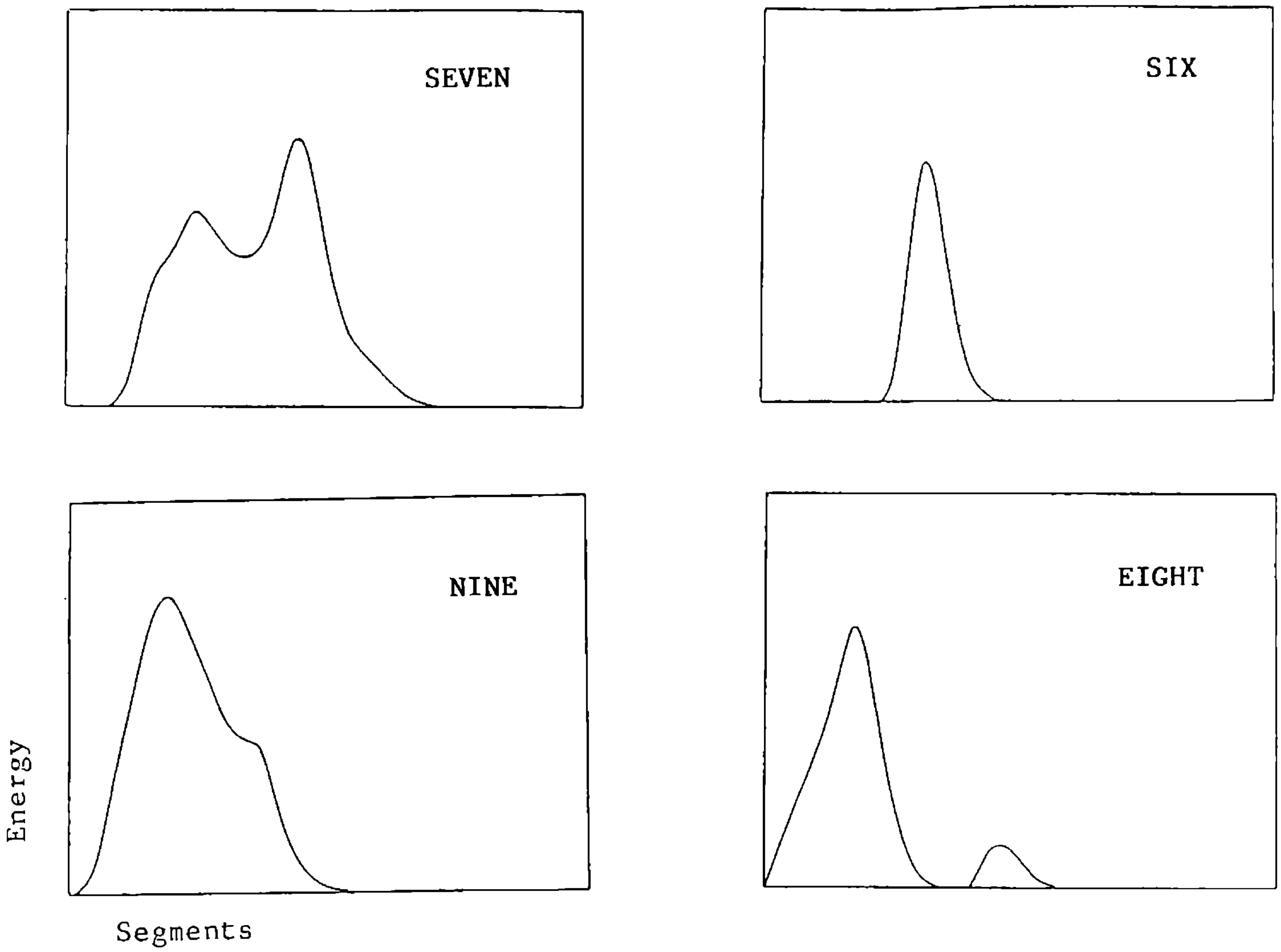


Fig.B.1(Cont.) Typical Energy Envelope of Digits.

The zerocrossing rate obtained for the segments are again super segmented into 5 regions for the purpose of classification. Normalized energy and total zerocrossing rate (TZCR) are also determined.

B.3 CLASSIFICATION ALGORITHM

The classification procedure has a tree like structure. The preliminary classification is performed by counting the number of peaks in the energy envelope. The digits are classified into 3 main groups, viz., Digits having single peak, two peaks and three peaks. Digits 'one', 'two', 'four', 'five', 'six' and 'nine' have shown single peak. Digits 'zero', 'three', 'seven' and 'eight' have 2 peaks. In certain cases digits 'four' and 'five' also have shown 2 peaks. The digits 'seven' and 'eight' have 3 peaks in some cases.

In the group having single peak, further classification is based on the peak position. If the peak position is less than 20 segments, then the digits are either 'two' or 'six' To distinguish between 'two' and 'six' one has to check normalized energy. Digit 'six' has low normalized energy. Moreover, for digit 'two', total zerocrossing rate is less than 7. If peak position is not less than 20 segments, then next classification is based on normalized energy. In the case of digit 'six' the normalized energy is low and ZCR is high. If these two conditions are satisfied, then digit 'six' is confirmed. If the normalized energy is higher than 20, then the possible digits are 'one' 'nine' or 'five' If the total ZCR is less than 9, then it is either digit 'one' or 'four'. But for digit 'one' ZCR at super segments 2 or 3 is higher. In some cases, total ZCR is greater than 10 for digits 'four' 'nine' and 'five' In such cases first super segment has higher ZCR for digit 'four' or else the digits are 'five' or 'nine'. During classification, digits 'five' and 'nine' were confused many times. The digit 'nine' usually come under this classification of 'five'. In most cases of digit 'nine' the TZCR is greater than 9 and less than 10 and also some of the first three super segments have ZCR higher than 6.

The classification among the group having two peaks is performed by searching the peak positions. If both the peaks are at less than 47th segment and first peak position is at less than 15th segment (ie., energy peak at the very initial stage) then the possible digits are 'three' and 'five' But for digit 'three' total zerocrossing rate is very low, whereas for digit 'five' total zerocrossing rate is comparatively higher. If the second peak position is at the end region (ie., beyond 65th segment) and the differences between first and second peak position is greater than 35 segments (due to the silence between /ai/ and /t/) then the digit is 'eight'. If the first peak position is at less than

20th segment (ie., energy peak at initial position) and total zerocrossing rate is greater than 6, then it is digit 'four' If ZCR is higher in first super segment and total ZCR is higher than 9 and second peak position is at less than 60th segment then the digit is 'seven'. Else, if zerocrossing rate is higher in the fourth super segment, it is digit 'zero' or in some cases it is digit 'four' But, in the case of digit 'four', second peak position is beyond 60th segment.

In the case of group having 3 energy peaks, the possible candidates are digits 'seven' and 'eight'. If the last peak position is at the end region and differences between the last and previous peak position is larger, then it is classified as digit 'eight' For digit 'seven' the zerocrossing rate is higher in the first super segment.

The complete tree structure for classification is given in Fig.B.2.

B.4 EXPERIMENT AND RESULTS

The experiment is conducted with 3 male speakers. Ten recordings of each digit for all speakers are taken in a noisy room using a dynamic microphone attached to the speech digitizer. The decision algorithm is not designed for the characteristics of each person, so as to give a true test of the speaker independent

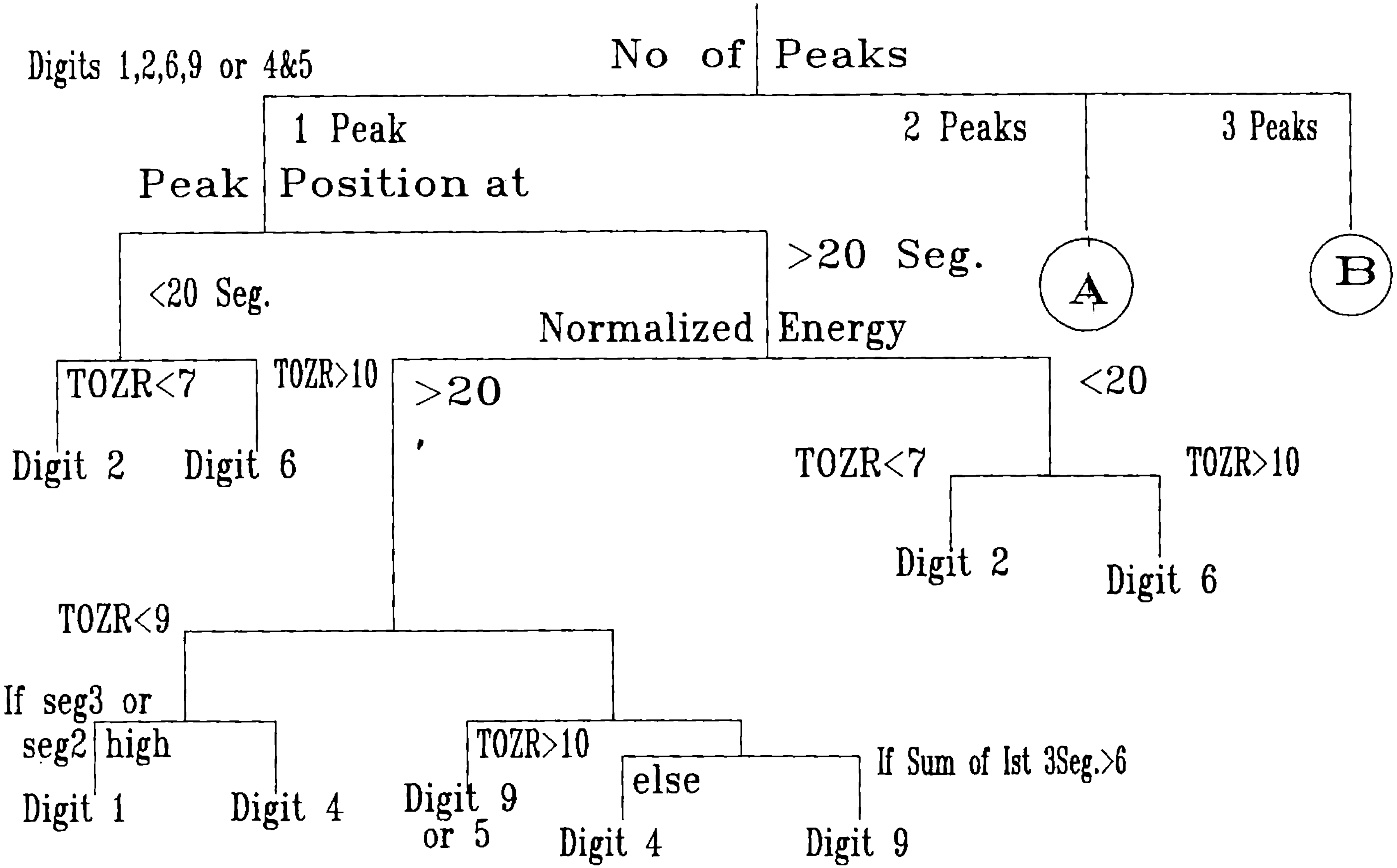


Fig.B.2 Tree Classification of Digits

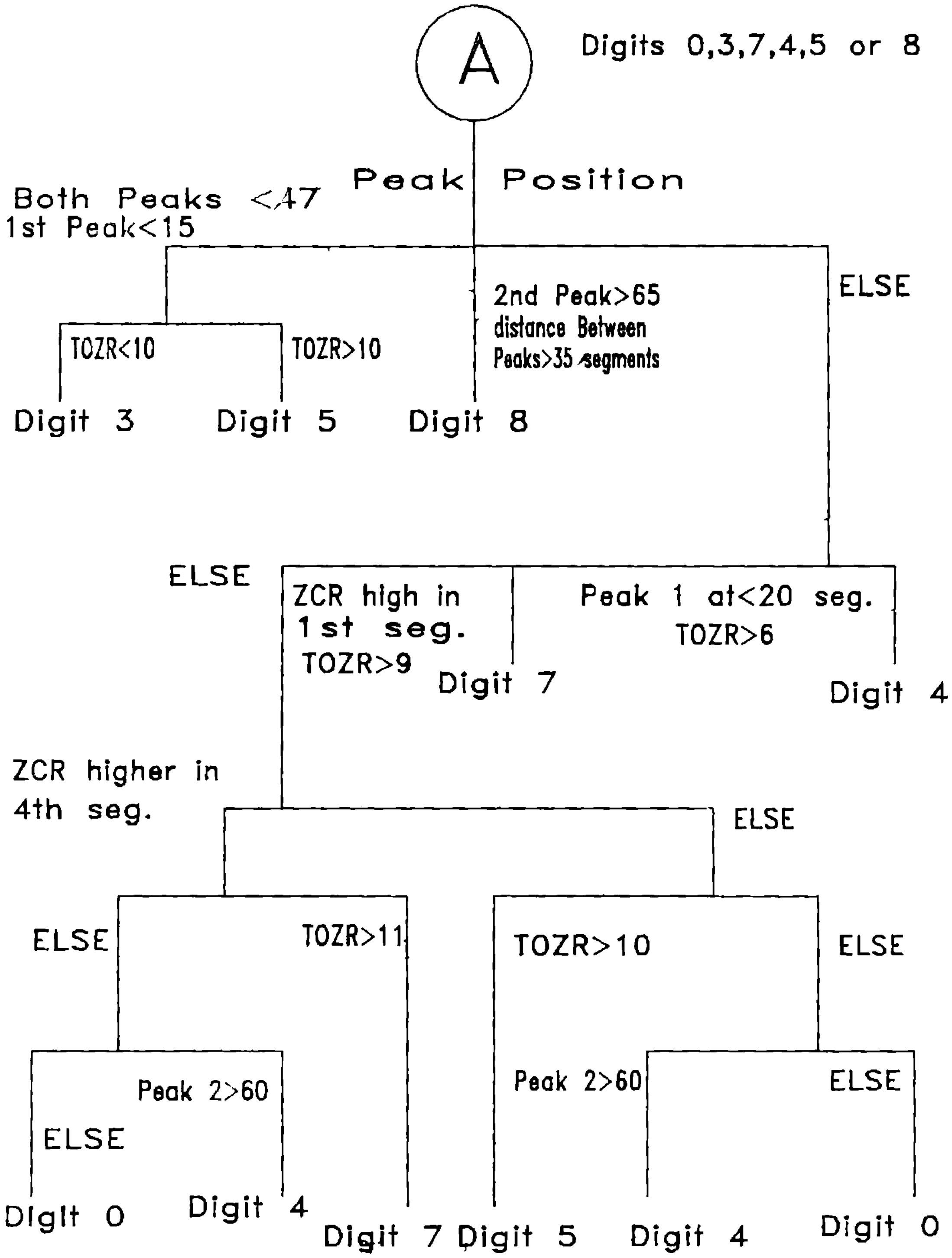


Fig.B.2(Cont.) Tree Classification of Digits

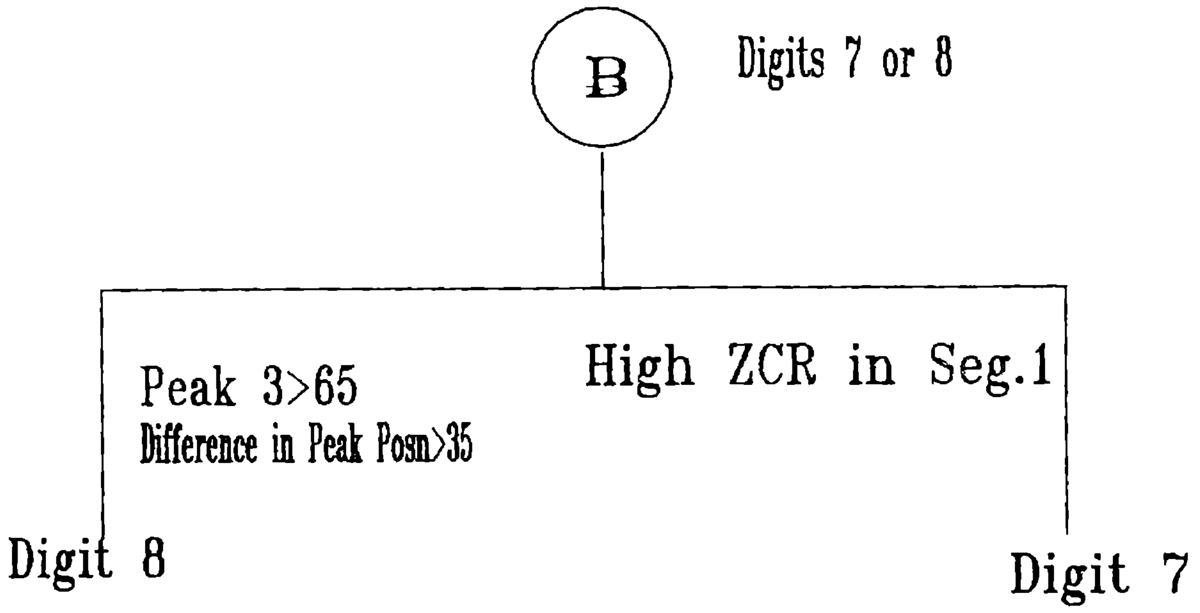


Fig.B.2(Cont.) Tree Classification of Digits nature of speech. A confusion matrix given in Fig.B.3 shows the complete result. A recognition accuracy of 81.60% has been achieved by this method. The time taken by a 386 based super PC/AT,for the recognition,is less than 20 seconds.

B.5 CONCLUSION AND DISCUSSION

A very simple approach for digit recognition using computer is introduced. The computational simplicity is the main attractive point of this method. This method makes use of an a priori knowledge about each digit and uses more or less a knowledge based decision approach. The extra hardware requirement is nil and this is purely a software technique. The features chosen are also simple and easily measurable.

There are some short comings in this work. The speech digitizer used is not of very high quality. The performance would have been improved if a better quality digitizer was used for digitizing the data. The method is not completely automatic, since the end point detection is done manually. Attempts are being made for automatic end point detection using the energy envelope.

Digit Recognized as

	Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Error
Zero	20	X	X	X	8	2	X	X	X	X	10
One	X	26	X	X	X	X	1	X	X	3	4
Two	X	X	28	X	1	X	1	X	X	X	2
Three	X	X	X	28	X	2	X	X	X	X	2
Four	2	X	X	X	27	X	X	1	X	X	3
Five	5	X	X	X	X	15	X	X	X	10	15
Six	X	X	X	X	X	X	28	X	2	X	2
Seven	X	2	X	X	X	X	X	27	X	1	3
Eight	X	X	X	X	4	X	X	X	26	X	4
Nine	X	7	X	X	X	X	X	X	X	20	10
									Total	Error	55

Fig.B.3 Confusion Matrix for Digits

REFERENCES

REFERENCES

1. J.Pollack, J.M.Pickett and W.H.Somby, "On the identification of speakers by voice", Journal Acoustical Society of America, Vol.26, No.3, May 1954, pp.403-406.

2. S.Pruzansky, "Pattern-matching procedure for automatic talker recognition", ibid., Vol.35, No.3, March 1963, pp.354-358.

3. R.C.Lummis, "Real time technique for speaker verification by computers", ibid., Vol.50, 1971, p.106(A).

4. A.E.Rosenberg, "Listener performance in a speaker verification task", ibid., Vol.50, No.1, part 1, 1971, p.106(A).

5. S.K.Das and W.S.Mohn, "A scheme for speech processing in automatic speaker verification", IEEE Transactions of Audio Electro Acoustics, AU-19, 1971, pp.32-43.

6. R.C.Lummis, "Implementation of an on-line speaker verification scheme", Journal Acoustical Society of America, Vol.52, 1972, p.181(A).

7. B.S.Atal, "Text-independent speaker recognition", ibid., Vol. 52, No.1, part 1, 1972, p.181(A).

8. C.E.Williams and K.N.Stevens, "Emotions and speech: Some acoustical correlates", ibid., Vol.52, 1972, pp.1238-1250.

9. O.Tosi et al., "Experiment on voice identification", ibid., Vol.51, 1972, pp.2030-2043.

10. J.J.Wolf, "Efficient acoustic parameters for speaker recognition", ibid., Vol.51, 1972, pp.2044-2056.

11. R.C.Lummis and A.E.Rosenberg, "Test of an automatic speaker verification method with intensively trained mimics", ibid., Vol.51, 1972, p.131(A).

12. G.D.Hair and T.W.Rekieta, "Automatic speaker verification using phoneme spectra", ibid., Vol.51, 1972, p.131(A).

13. D.Meltzer, "Vowel and speaker identification in natural and synthetic speech", ibid., Vol.51, No.1 (part 1), 1972, p.131(A).

14. S.Cort and T.Murry, "Aural identification of children's voice", ibid., Vol.51, No.1 (part 1), 1972, p.131(A).

15. T.W.Rekieta and G.D.Hair, "Mimic resistance of speaker verification using phoneme spectra", ibid., Vol.51, No.1 (part 1), 1972, p.131(A).

16. A.E.Rosenberg, "Listener performance in a speaker verification task with deliberate imposters", ibid., Vol.51, No.1 (part 1), 1972, p.132(A).

17. N.S.Jayant, "Decision-theoretic approach to speaker verification", ibid., Vol.51, No.1 (part 1), 1972, p.132(A).

18. B.S.Atal, "Automatic speaker recognition based on pitch contours", ibid., Vol.52, No.6 (part 2), 1972, pp.1687-1697.

19. R.C.Lummis, "Speaker verification by computers using speech intensity for temporal registration", IEEE Transactions on Audio Electro Acoustics, Vol.AU-21, April 1973, pp.80-89.

20. B.S.Atal, "Effectiveness of linear predictive characteristics of the speech wave for automatic speaker identification and verification", Journal Acoustical Society of America, Vol.55, 1974, pp.1304-1312.

21. L.S.Su, K.P.Li and K.S.Fu, "Identification of speakers by use of nasal coarticulation", ibid., Vol.56, 1974, pp.1876-1882.

22. A.E.Rosenberg and M.R.Sambur, "New techniques for automatic speaker verification", IEEE Transactions on Acoustic Speech and Signal Processing, Vol.ASSP-23, No.2, April 1975, pp. 169-176.

23. M.R.Sambur, "Selection of acoustic features for speaker identification", ibid., Vol.ASSP-23, April 1975, pp.176-182.

24. J.E.Paul,Jr., A.S.Rabinowitz, J.P.Riganati, J.M.Richardson, "Development of an analytical method for semi automatic speaker identification system", 1975 CARNAHAN Conf. on Crime and Counter Measures, May 1975, pp.52-64.

25. D.A.Wasson and R.W.Donaldson, "Speech amplitude and zerocrossings for automated identification of human speakers", IEEE Transactions on Acoustic Speech and Signal Processing, Vol. ASSP-23, August 1975, pp.390-392.

26. S.Furui, F.Itakura and S.Saito, "Personal information in the long-time averaged speech spectrum", Review of the Electrical Communication Labs., Vol.23, No.9-10, September-October 1975, pp.1133-1141.

27. B.S.Atal, "Automatic recognition of speaker from their voices", Proc.IEEE, Vol.64, No.4, April 1976, pp.460-475.

28. A.E.Rosenberg, "Automatic speaker verification: A review", ibid., Vol.64, No.4, April 1976, pp.475-487.

29. S.Furui and F.Itakura, "Analysis of talker differences in statistical properties of speech spectra", Review of Electrical Communication Labs., Vol.24, Nos.5-6, May-June 1976, pp.418-428.

30. G.R.Doddington, "Personal identity verification using voices", Proc. ELECTRO-76, 1-5, May 11-14, 1976, pp.22-24.

31. A.E.Rosenberg, "Evaluation of an automatic speaker verification system over telephone lines", The Bell System Technical Journal, Vol.55, No.6, July-August 1976, pp.723-744.

32. M.R.Sambur, "Speaker recognition using orthogonal linear prediction", IEEE Transactions on Acoustic Speech and Signal Processing, Vol.ASSP-24, No.4, August 1976, pp.283-289.

33. S.Furui, "Study on talker dependent features of speech and its application to speech recognition", Review of Electrical Communication Lab., NTT, No.149, October 1976, pp.1-15.

34. R.L.Kashyap, "Speaker recognition from an unknown utterance and speaker speech interaction", IEEE Transactions on Acoustic Speech and Signal Processing, Vol.ASSP-24, December 1976, pp.481-488.

35. V.V.S.Sarma and B.Yegnanarayana, "A critical survey of automatic speaker recognition systems", CSI-76 Convention Record, 1976, pp.282-296.

36. S.Furui, "Analysis of temporal variation of talker dependent features", Review of Electrical Communication Lab., Vol.25, No.3-4, March-April 1977, pp.231-244.

37. M.J.Hunt, J.W.Yates, J.S.Bridle, "Automatic speaker recognition for use over communication channels", IEEE Int.Conf.Rec. on Acoustic Speech and Signal Processing, May 9-11, 1977, pp.764-767.

38. E.Bunge, "Automatic speaker recognition system AUROS for security system and forensic voice identification", Proc. 1977 Int.Conf. on Crime Counter Measures--Science and Engineering, July 25-29, 1977, pp.1-7.

39. J.D.Markel, B.Oshika and A.H.Gray,Jr., "Long-term feature averaging for speaker recognition", IEEE Transactions on Acoustic Speech and Signal Processing, Vol.ASSP-25, No.4, August 1977, pp.330-337.

40. E.Bunge et al., "The AUROS project-automatic recognition of speakers by computers", Frequency, Vol.31, No.12, December 1977, pp.345-351.

41. L.Fasalo and G.A.Mian, "A comparison between two approaches to automatic speaker recognition", Proc. Int.Conf. on Acoustic Speech and Signal Processing, April 10-12, 1978, pp.275-278.

42. P.Jesorsky, U.Hofker and M.Talmi, "Extraction of speaker-specific features from spoken code sentences", Proc.Int. Conf. on Acoustic Speech and Signal Processing, April 10-12, 1978, pp.279-282.

43. L.L.Pfeifer, "New technique for text-dependent speaker identification", ibid., pp.283-286.

44. J.D.Markel and S.B.Davis, "Text independent speaker identification", ibid., pp.287-290.

45. H.Matsumoto and T.Minura, "Text independent speaker identification based on piecewise canonical discriminant analysis", ibid., pp.291-294.

46. H.M.Dante and V.V.S.Sarma, "Automatic speaker identification for a large number of speakers", ibid., pp.295-298.

47. C.Basztura and W.Majewski, "The application of long-term analysis of the zerocrossing of a speech signal in automatic speaker identification", Arch.Acoust., Vol.3, No.1, 1978, pp.3-15.

48. G.R.Doddington, "Voice authentication parameters and usage", 1978 WESCON Technical Papers, Los Angeles, USA, September 12-14, 1978, pp.28.3/1-6.

49. C.Basztura and J.Jurkeiwicz, "The zerocrossing analysis of a speech signal in the short-term method of automatic speaker identification", Arch.Acoust., Vol.3, No.3, 1978, pp.185-195.

50. C.A.McGonegal, L.R.Rabiner, B.J.McDermott, "Speaker verification by human listeners over speech transmission system", Bell Syst.Tech. Journal, Vol.57, No.8, October 1978, pp. 2887-2900.

51. R.S.Cheung and B.Eisenstein, "Feature selection via dynamic programming for text independent speaker identification", IEEE Transactions on Acoustic Speech and Signal Processing, Vol.ASSP-26, October 1978, pp.397-403.

52. T.Tran Cao, F.Spronck, "The zerocrossing method for vowel recognition and speaker identification", Rev.HF, Vol.10, No.11-12, 1978, pp.330-342.

53. S.Saito and S.Furui, "Personal information in dynamic characteristics of speech spectra", Proc. 4th Int. Joint Conf. on Pattern Recognition, Japan, November 7-10, 1978, pp.1014-1018.

54. Y.Granier, "Speaker identification from linear prediction", ibid., pp.1019-1021.

55. A.Ichikawa, A.Nakajima and K.Nakata, "Speaker verification from actual telephone voice", Journal Acoustic Society of Japan, Vol.35, No.2, February 1979, pp.63-69.

56. S.Nakagawa and T.Sakai, "Feature analysis of Japanese phonetic spectra and consideration on speech recognition and speaker identification", ibid., Vol.35, No.3, March 1979, pp.111-117.

57. V.V.S.Sarma, B.Yegnanarayana, T.V.Ananthapadmanabha, "A speaker recognition scheme on a mini computer based on a signal processing facility", Acoustica, Vol.41, 1978, pp.117-121.

58. G.A.Mian, "Some factors influencing the performance of speaker recognition system based on LPC", Proc.of Int.Conf. on Acoustic Speech and Signal Processing, Washington, USA, April 2-4, 1979, pp.781-784.

59. M.Shridhar and M.Baranieki, "Accuracy of speaker verification via orthogonal parameters for noisy speech", ibid., pp.785-788.

60. U.Hofker and P.Jesorsky, "Structure and performance of an on-line speaker verification system", ibid., pp.789-792.

61. W.D.Voiers, "Towards the development of practical methods of evaluating speaker recognizability", ibid., pp.793-796.

62. H.M.Dante, V.V.S.Sarma and G.R.Dattatreya, "Multistage decision schemes for speaker recognition", ibid., pp.797-800.

63. E.Bunge, "Automatic forensic speaker recognition", Proc. of 1979 Carnahan Conf. on Crime Counter Measures, LX, USA, May 16-18, 1979, pp.41-45.

64. U.Hoefker, P.Jesorsky, B.Kriener, M.Talmi and D.Wesseling, "A new system for authentication of voice", ibid., pp.47-52.

65. R.Geppert, M.H.Kuhn, H.Piotrowsky and H.Tomaschewski, "An operational system for personal authentication using long-term averaged voice spectra", ibid., pp.77-80.

66. R.Dubes, A.K.Jain and O.I.Tosi, "Speaker recognition studies on choral speech", ibid., pp.165-171.

67. P.Jesorsky, "Principles of automatic speaker recognition", in Speech Communication with Computers, Ed. L.Bolc, Macmillan, 1978, pp.93-137.

68. H.M.Dante and V.V.S.Sarma, "Automatic speaker identification for a large population", IEEE Transactions on Acoustic Speech and Signal Processing, Vol.ASSP-27, No.3, June 1979, pp. 255-263.

69. A.I.Menkiti, "Speaker identification", Acoust.Lett., Vol.3, No.2, August 1979, pp.45-53.

70. M.H.Kuhn, "Speaker recognition accounting for different voice condition by unsupervised classification (Cluster Analysis)", IEEE Transactions on Syst., Man and Cybernetics, Vol.SMC-10, No.1, January 1980, pp.34-37.

71. M.H.Kuhn, "Access control by means of automatic speaker verification", J.Phys.E., Vol.13, No.1, January 1980, pp.85-86.

72. M.Baraniecki and M.Shridhar, "A speaker verification algorithm for speech utterances corrupted by noise with unknown statistics", Proc.Int.Conf.on Acoustic Speech and Signal Processing, Denver, USA, April 9-11, 1980, Part III, pp. 904-907.

73. R.E.Wohlford, E.H.Wrench and B.P.Landell, "A comparison of four techniques for automatic speaker recognition", ibid., pp.908-911.

74. S.Furui and A.E.Rosenberg, "Experimental studies on a new automatic speaker verification system using telephone speech", ibid., pp.1060-1062.

75. C.E.Chafei, "A real time automatic speaker recognition system on mini computer", Proc. 1980 Carnahan Conf. on Crime and Counter Measures, KY, USA, May 14-16, 1980, pp.53-56.

76. M.H.Kuhn, "A low cost speaker verification device", ibid., pp.57-61.

77. H.M.Dante and V.V.S.Sarma, "A pattern recognition model of voice-based personal verification system for forensic application", IEEE Transactions on Syst.Man and Cybernetics, Vol. SMC-10, No.9, September 1980, pp.585-589.

78. M.H.Kuhn, R.Geppert and R.Frehse, "Online evaluation of user acceptance in speaker verification", Proc. III Int. Conf. Security - The Science and Engineering, KY, September 23-26, 1980, pp.161-167.

79. H.Ney and M.H.Kuhn, "Cluster analysis for telephone line speaker recognition", Proc. of EUSIPCO-80, Switzerland, September 16-18, 1980, pp.609-613.

80. R.E.Bogner, "On talker verification via orthogonal parameters", IEEE Transactions on Acoustic Speech and Signal Processing, Vol.ASSP-29, No.1, February 1981, pp.1-12.

81. A.E.Rosenberg and K.L.Shipley, "Speaker identification and verification combined with speaker independent word recognition", Proc. Int. Conf. on Acoustic Speech and Signal Processing, Atlanta, USA, March-April 1981, Vol.I, pp.184-187.

82. H.Ney, "Telephone line speaker recognition using clipped autocorrelation", ibid., pp.189-192.

83. E.H.Wrench,Jr., "A real time implementation of a text independent speaker recognition system", ibid., pp.193-196.

84. M.Shridhar, N.Mohanakrishnan, M.Baraniecky, "Text independent speaker recognition using orthogonal linear prediction", ibid., pp.197-200.

85. R.J.Fontana and M.S.Fox, "A composite source model for speaker and isolated word recognition", ibid., pp.741-745.

86. S.Furui, "Cepstral analysis techniques for automatic speaker verification", IEEE Transactions on Acoustic Speech and Signal Processing, Vol.ASSP-29, April 1981, pp.254-272.

87. M.De George, "Experiments in automatic speaker verification", Proc.of 1981 Carnahan Conference on Crime Counter Measures, KY, USA, May 13-15, 1981, pp.103-110.

88. A.Kohen and I.Froind, "Software package for interactive text independent speaker verification", Proc. of MELECON '81, Tel Aviv, May 24-28, 1981, pp.6.2.3/1-4.

89. S.Furui, "Comparison of speaker recognition methods using statistical features and dynamic features", IEEE Transactions on Acoustic Speech and Signal Processing, Vol.ASSP-29, No.3, June 1981, pp.342-350.

90. H.Ney, "Measuring of the contours of intensity and fundamental period of speech for automatic speaker recognition", Frequenz, Vol.35, No.10, October 1981, pp.265-270.

91. K.Itoh and S.Saito, "Effects of acoustical feature parametders of speech on perceptual identification of speakers", Trans. Inst. of Elect. & Commun. Engg. Jpn, Sect.E, Vol.E65, No.1, January 1982, pp.73-74.

92. S.Furui, "Speaker recognition by statistical features of cepstral parameters", Trans.Inst.Electr. & Commun.Engg. Jpn, Sect.E, Vol.E65, No.2, February 1982, p.130.

93. S.Furui, "Speaker recognition by statistical features and dynamic features", Review of Electrical & Commun.Lab., Vol.30, No.3, 1982, pp.467-482.

94. H.Ney and R.Gierioff, "Speaker recognition using feature weighting technique", Proc.Int.Conf.on Acoustic Speech and Signal Processing, Paris, May 3-5, 1982, pp.1645-1648.

95. R.Schwartz, S.Roucos and M.Berouti, "The application of probability density estimation to text-independent speaker identification", ibid., pp.1649-1652.

96. N.Mohanakrishnan, M.Shridhar and M.A.Sid-Ahmed, "A composite scheme for text-independent speaker recognition", ibid., pp.1653-1656.

97. M.Shridhar, M.Baraniecki and N.Mohanakrishnan, "A unified approach to speaker verification with noisy speech inputs", Speech Communication, Vol.1, No.2, August 1982, pp.103-112.

98. H.Ney, "Automatic speaker recognition using time alignment of spectrogram", ibid., pp.135-149.

99. M.Shridhar and N.Mohanakrishnan, "Test independent speaker recognition: A review and some new results", Speech Communication, Vol.1, No.3-4, December 1982, pp.257-267.

100. J.Wolf, M.Krasner, K.Karnofsky, R.Schwartz and S.Roucos, "Further investigation of probabilistic methods for text-independent speaker identification", Proc.Int.Conf. on Acoustic Speech and Signal Processing, Boston, USA, Vol.2, April 14-16, 1983, pp.551-554.

101. K.P.Li and E.H.Wrench,Jr., "An approach to text-independent speaker recognition with short utterances", ibid., pp.555-558.

102. M.Shridhar, N.Mohanakrishnan, and M.A.Sid-Ahmed, "A comparison of distance measure for text independent speaker identification", ibid., pp.559-562.

103. M.J.Hunt, "Further experiments in text-independent speaker recognition over communication channels", ibid., pp.563-566.

104. D.Helling and H.W.Strube, "Speaker recognition based on isolated words", Proc of EUSIPCO-83, Erlangen, Germany, September 12-16, 1983, pp.423-426.

105. S.Nakagawa and M.Sakamoto, "Evaluation of FFT cepstrum and LPC cepstrum for speech and speaker recognition", Trans. on Inst. Electr. & Commun. Engg. Jpn, Part A, Vol.J66A, No.12, December 1983, pp.1199-1206.

106. M.Krasner, J.Wolf, K.Karnofsky, R.Schwartz, S.Roucos and H.Gish, "Investigation of text independent speaker identification techniques under conditions of variable data", Proc. Int.Conf.on Acoustic Speech and Signal Processing, San Digo, USA, Vol.2, March 19-21, 1984, pp.18B.5/1-4.

107. P.E.Papamichialis and G.R.Doddington, "A speaker recognizability test", ibid., pp.18B.6/1-4.

108. H.Holein, "Natural speech vectors for speaker identification", Proc. of Speech Tech '85: Voice input/output application", Show and Conf., N.Y., USA, April 22-24, 1984, pp.331-334.

109. C.B.A.Shaw Cross, S.R.Parker and A.H.Karyan, "Isolated word recognition and speaker identification using two dimensional half plane lattice parameter of spectrogram", Digital Signal Processing-84, Proc.Int.Conf.Florance, Italy, September 5-8, 1984, pp.449-453.

110. D.E.Crabbs and D.P.Conrad, "Practical experience with a proto-type speaker verification system", Proc. of MIDCON-84 Electronic Show and Convention, TX, USA, September 11-13, 1984, pp.2/15/19-2/1-10.

111. H.Tanaka, H.Noda, S.Hikita, S.Kai and K.Hirakawa, "An application of time sequence matrix to speaker identification", Technology Report Kyushi Univ., Vol.57, No.5, October 1984, pp.613-620.

112. D.E.Crabbs and J.R.Clymer, "Operational evaluation modelling of automatic speaker verification system", 1984 Winter Simulation Conference Proc., Texas, USA, November 28-30, 1984, pp.753-761.

113. A.Schmidt Nielen and R.Stern, "Identification of known voice as a function of familiarity and narrow-band coding", Journal Acoustical Society of America, Vol.77, No.2, February 1985, pp.658-663.

114. H.Gish, K.Karnofsky, M.Krasner, S.Roucos, R.Scwartz and J.Wolf, "Investigation of text-independent speaker identification over telephone channels", Proc.Int.Conf.on Acoustic Speech and Signal Processing, Vol.1, Tampa, USA, March 26-29, 1985, pp.379-382.

115. S.S.Everett, "Automatic speaker recognition using vocoded speech", ibid., pp.383-386.

116. F.K.Soong, A.E.Rosenberg, L.R.Rabiner and B.H.Juang, "A vector quantization approach to speaker recognition", ibid., pp.387-390.

117. J.T.Buck, D.K.Burton and J.E.Shore, "Text-dependent speaker recognition using vector quantization", ibid., pp.391-394.

118. W.Flix and M.De Geoorge, "A speaker verification system for access-control", ibid., pp.399-402.

119. H.Kashiwagi, S.Nakamura and M.Takanashi, "Speaker identification by spectral envelope of linear prediction residual", Trans.Inst.Electr. & Commun.Engg. Jpn. Part A, Vol.J68A, No.7, pp.702-703.

120. H.Noda, "A method for computing decision threshold for use in forensic application of speaker verification", Proc. 1985 Carnahan Conf. on Security and Technology, KY, USA, May 15-17, 1985, pp.159-162.

121. A.E.Rosenberg and K.L.Shipley, "Talker recognition in tandem with talker-independent isolated word recognition", IEEE Transactions on Acoustic Speech and Signal Processing, Vol. ASSP-33, No.3, June 1985, pp.574-586.

122. Z.Uzdy, "Human speaker recognition using performance of LPC voice processors", ibid., pp.752-753.

123. M.Ganesan, S.S.Agrawal, A.M.Ansari and K.D.Pavate, "Recognition of speaker based on acoustic parameters of vowel sounds", Jour.Acous.Soc.India, Vol.XIII, No.4, October 1985, pp.202-212.

124. G.R.Doddington, "Speaker recognition identifying people by their voices", Proc.IEEE, Vol.73, No.11, November 1985, pp.1651-1664.

125. G.Audiso and A.Caramella, "Speaker verification algorithm using Parzen estimate", CSELT Tech.Rep., Vol.13, No.6, November 1985, pp.389-390.

126. A.L.Higgins, "Speaker recognition by template matching", Official Proc. of SPEECH TECH '86, Voice Input/Output Application Show and Conf., NY, USA, April 28-30, 1986, pp.273-275.

127. M.C.K.Yang, H.Hollien and R.Huntley, "A speaker identification system for field use", ibid., pp.277-280.

128. M.De George and W.Fiex, "A speaker verification system for access-control", ibid., pp.281-286.

129. M.Burnbaum, R.W.Bossulmeyer, L.A.Cohen and F.X.Welsh, "Using cepstral features in speaker verification", ibid., pp.287-290.

130. S.Furui, "Research on individuality features in speech waves and automatic speaker recognition techniques", Speech Communication, Vol.5, No.2, June 1986, pp.183-197

131. N.Fakotakis, E.Dermatas and G.Kokkinakis, "Optical decision threshold for speaker verification", Signal Processing III, Theory and Application, Proc.EUSIPCO-86, Hagen, Netherlands, September 2-5, 1985, Vol.1, pp.585-587.

132. D.O'Shaughnessy, "Speaker recognition", IEEE ASSP Magazine, Vol.3, No.4, October 1986, pp.4-17.

133. D.K.Burton, "Text-dependent speaker verification using vector quantization source coding", IEEE Transactions on Acoustic Speech and Signal Processing, Vol.ASSP-35, No.2, February 1987, pp.133-143.

134. A.Michailov and D.Milev, "A method for automatic speaker recognition with personal computer", Fifth Symp. on Microcomputer and microprocessor application", Budapest, Hungary, September-October 1987, pp.659(A).

135. H.Noda, "Frequency-warped spectral distance measures for speaker verification in noise", Proc.Int.Conf.on Acoustic Speech and Signal Processing, Vol.V, NY, USA, April 11-14, 1988, pp.576-579.

136. Y.C.Zheng and B.Z.Yuan, "Text-dependent speaker identification using circular hidden Markov models", ibid., pp.580-582.

137. G.Velius, "Variants of cepstrum based speaker identity verification", ibid., pp.583-586.

138. J.Wilbur and F.J.Taylor, "Consistent speaker identification via Wigner smoothing techniques", ibid., pp.591-594.

139. J.B.Attili, M.Savic and J.P.Campbell,Jr., "A TMS32020-BASED real time, text independent automatic speaker verification system", ibid., pp.599-602.

140. L.R.Rabiner and R.W.Schafer, "Digital processing of speech signals", Prentice Hall Inc., 1978.

141. S.Saito and K.Nakata, "Fundamentals of speech signal processing", Academic Press Inc., 1985.

142. T.Tsukamoto and Y.Tohkura, "Perceptual units of the infant cry", 117th ASA Meeting, Syracuse, May 1989.

143. J.Kitler, "Feature selection and extraction", in Handbook of pattern recognition and image processing, Academic Press Inc., 1986.

144. A.M.Noll, "Cepstrum pitch determination", Journal Acoustical Society of America, Vol.41, February 1967, pp.293-309.

145. M.M.Sondhi, "New methods of pitch extraction", IEEE Transactions on Audio Electroacoustics, Vol.AU-16, June 1968, pp.262-266.

LIST OF PUBLICATIONS

1. "A Memory Backed Non-Linear Speech Processor", International Symposium on Computers, Systems and Signal Processing, IISc., Bangalore, Dec.1984.

2. "Use of Zerocrossing Informations to Extract Speech Parameters", Journal Acoustical Society of India, Vol.XV, No.1, Jan.1987.

3. "A New Algorithm for Pitch Detection by Symmetry Checking", Proc. of International Symposium on Electron Devices, Computers and systems, IIT Kharagpur, Dec.1987.

4. "Feature Matrix Method for Speaker Identification", Proc. of Symposium on Signals, Systems and Sonar, NPOL, Cochin, Mar.1988.

5. "A Low Cost PC Interfaceable Speech Digitizer", Proc. of Symposium on Signals, Systems and Sonar, NPOL, Cochin, Mar.1988.

6. "Study of Pitch Consistency for Speaker Recognition", Journal Acoustical Society of India, Vol.XVI, Nos. 3&4, pp.270-72.

7. "A Speech Workstation with 12 bit Digitizer", Proc. of National Symposium on Instrumentation(NSI-13), JCE College, Mysore, Nov.1988.

8. "A Universal Fully Programmable 16-Channel Data Acquisition System", Proc. of National Symposium on Instrumentation(NSI-13), JCE College, Mysore, Nov.1988.

9. "A Phonemic Model Based Speaker Verification System", Proc. of National Conference on Electronics, Circuits and Systems, Univ. of Roorke, Nov. 1989.

10. "Digit Recognition Using Energy Envelope and Zerocrossing Rate", Proc. of TENCON'89: IEEE Region 10 International Conference, Bombay, Nov.1989.

11. "A Microprocessor Based Position Controller for Laser Optical Grating", Proc. of TENCON'89: IEEE Region 10 International Conference, Bombay, Nov.1989.

12. "Hardware Implementation of FFT 8086 Based System", Proc. of TENCON'89:IEEE Region 10 International Conference, Bombay, Nov.1989.

13. "Speaker Verification- Problems in Small Systems", Accepted for presentation in 120th Meeting of Acoustical Society of America, to be held at San Diego, CA, USA, during 26-30 Nov.1990.