Text-independent speaker recognition.

N. Mohankrishnan

University of Windsor

Follow this and additional works at: https://scholar.uwindsor.ca/etd

Recommended Citation
https://scholar.uwindsor.ca/etd/3741
NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us a poor photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

THIS DISSERTATION HAS BEEN MICROFILMED EXACTLY AS RECEIVED
TEXT-INDEPENDENT SPEAKER RECOGNITION

by

N. Mohankrishnan

A Thesis
presented to the University of Windsor
in partial fulfillment of the
requirements for the degree of,
Doctor of Philosophy
in
Electrical Engineering

Windsor, Ontario, 1984
ABSTRACT

This work is an investigation of text-independent speaker recognition schemes making use of acoustic characterizations obtained from short-duration reference and test speech segments of the various speakers in the population.

Initially, various parametric representations of speech, both in the time and frequency domains, were tested to ascertain their relative potential for speaker recognition. These were the linear prediction coefficients, reflection coefficients, cepstrum coefficients, log area ratio coefficients, speech spectrum parameters, inverse filter spectral coefficients, and the direct FFT coefficients. The speaker classification was performed using the Mahalanobis minimum distance classifier. The spectral parameters, in general, were found to perform better than the time domain parameters. The highest identification accuracy of 88.54% for a 5 second test utterance duration was obtained using the inverse filter spectral coefficients which also yielded the lowest total error verification rate of 3.82%. The recognition accuracies improved with use of increased test utterance lengths.
The potential for feature selection among the elements of each feature set in the orthogonal space obtained by principal component analysis was investigated through the use of a dynamic programming algorithm. The feature selection resulted in improved accuracies for most of the parameter sets. This improvement was between 2% to 4% for the various parameter sets in the identification task. Smaller decreases in the total error rate were obtained in the case of the verification task.

The performance of the nearest neighbour classification rule in speaker identification was investigated. The results reveal a performance comparable to that of the NN. By incorporating blocking strategies into the decision logic significant improvements in recognition accuracy could be obtained.

Finally, a composite scheme for speaker recognition has been proposed. This scheme is based on logical combination of recognition decisions obtained from two different feature sets. The use of this decision scheme led to identification accuracies of over 95% and verification total error rates as low as 2%. It is felt that the composite scheme is very promising as the basis of a high-accuracy speaker recognition system.
ACKNOWLEDGEMENTS

I wish to express my sincere thanks to Dr. K. Shridhar for his guidance and supervision during the course of this work. Working with him was an inspiration indeed. I also wish to thank all the other faculty members for their interest and advice, in particular Dr. G.A. Jullien, who gave freely of his time in helping to resolve the numerous hardware problems that arose periodically.

I gratefully acknowledge the assistance given by my fellow graduate students throughout this work. I also wish to express my sincere gratitude to those extremely patient people who helped in the creation of the speech data base for this study by consenting to having their voices recorded.

Finally I wish to thank our technicians Mr. Joe Novosad and Mr. Alan Thibert and our secretaries Mrs. Shirley Ouelette and Mrs. Linda Kennedy for all the help they rendered during this work.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>ii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iv</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>viii</td>
</tr>
<tr>
<td>I. INTRODUCTION</td>
<td></td>
</tr>
<tr>
<td>Preview</td>
<td>1</td>
</tr>
<tr>
<td>Problem Statement</td>
<td>3</td>
</tr>
<tr>
<td>Outline of Thesis</td>
<td>4</td>
</tr>
<tr>
<td>II. SPEAKER RECOGNITION</td>
<td>5</td>
</tr>
<tr>
<td>Introduction</td>
<td>5</td>
</tr>
<tr>
<td>General Procedure for Speaker Recognition</td>
<td>6</td>
</tr>
<tr>
<td>Classification of Speaker Recognition</td>
<td>9</td>
</tr>
<tr>
<td>Speaker Verification</td>
<td>9</td>
</tr>
<tr>
<td>Speaker Identification</td>
<td>11</td>
</tr>
<tr>
<td>Distance Measures</td>
<td>16</td>
</tr>
<tr>
<td>Sources of Inter-speaker Variability</td>
<td>23</td>
</tr>
<tr>
<td>Sources of Intra-speaker Variability</td>
<td>24</td>
</tr>
<tr>
<td>Literature Review</td>
<td>25</td>
</tr>
<tr>
<td>III. SPEAKER RECOGNITION POTENTIAL OF SPEECH PARAMETER SETS</td>
<td>40</td>
</tr>
<tr>
<td>Introduction</td>
<td>40</td>
</tr>
<tr>
<td>Speech Data Collection and Preprocessing</td>
<td>41</td>
</tr>
<tr>
<td>Extraction of Speech Parameters</td>
<td>42</td>
</tr>
<tr>
<td>Time Domain Parameters</td>
<td>42</td>
</tr>
<tr>
<td>Frequency Domain Parameters</td>
<td>43</td>
</tr>
<tr>
<td>Orthogonalization of Parameter Sets</td>
<td>46</td>
</tr>
<tr>
<td>Convergence Properties of the Covariance Matrix</td>
<td>48</td>
</tr>
<tr>
<td>Speaker Recognition Experiments</td>
<td>53</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>55</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>IV. FEATURE SELECTION</td>
<td></td>
</tr>
<tr>
<td>Introduction</td>
<td>62</td>
</tr>
<tr>
<td>Feature Selection through Dynamic Programming</td>
<td>64</td>
</tr>
<tr>
<td>Choice of PEC and Speaker Recognition Experiments</td>
<td>67</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>71</td>
</tr>
<tr>
<td>V. NEAREST NEIGHBOURHOOD TECHNIQUE</td>
<td>83</td>
</tr>
<tr>
<td>Introduction</td>
<td>83</td>
</tr>
<tr>
<td>Modified Nearest Neighbour Method</td>
<td>84</td>
</tr>
<tr>
<td>Speaker Recognition Experiments</td>
<td>85</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>86</td>
</tr>
<tr>
<td>VI. A COMPOSITE SCHEME FOR SPEAKER RECOGNITION</td>
<td>91</td>
</tr>
<tr>
<td>Introduction</td>
<td>91</td>
</tr>
<tr>
<td>Composite Recognition Scheme</td>
<td>93</td>
</tr>
<tr>
<td>Speaker Recognition Experiments</td>
<td>95</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>95</td>
</tr>
<tr>
<td>VII. DISCUSSION</td>
<td>100</td>
</tr>
<tr>
<td>Summary</td>
<td>100</td>
</tr>
<tr>
<td>Conclusions</td>
<td>103</td>
</tr>
<tr>
<td>Contributions of this work</td>
<td>103</td>
</tr>
<tr>
<td>Suggestions for future work</td>
<td>104</td>
</tr>
<tr>
<td>APPENDIX I</td>
<td>107</td>
</tr>
<tr>
<td>APPENDIX II</td>
<td>109</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>114</td>
</tr>
<tr>
<td>VITA AUCTORIS</td>
<td>118</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Fig. no.</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>General procedure for speaker recognition</td>
<td>7</td>
</tr>
<tr>
<td>2.2</td>
<td>Error rates in speaker identification and speaker verification vs. population size $N$ (after Doddington[9])</td>
<td>13</td>
</tr>
<tr>
<td>2.3</td>
<td>Summary of differences between speaker identification and speaker verification (after Flandam[3])</td>
<td>15</td>
</tr>
<tr>
<td>3.1</td>
<td>Generation of speech spectrum parameters</td>
<td>45</td>
</tr>
<tr>
<td>3.2</td>
<td>Convergence properties of covariance matrix of speech parameters</td>
<td>52</td>
</tr>
<tr>
<td>6.1</td>
<td>The composite scheme algorithm</td>
<td>94</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table no.  | Description                                                                 | Page |
---------|-----------------------------------------------------------------------------|------|
3.1  | Identification scores (error %) for various test data segment lengths - Time domain parameters | 56   |
3.2  | Identification scores (error %) for various test data segment lengths - Frequency domain parameters | 57   |
3.3  | Verification scores (total error %) for various test data segment lengths - Time domain parameters | 58   |
3.4  | Verification scores (total error %) for various test data segment lengths - Frequency domain parameters | 59   |
4.1  | Feature subsets selected through dynamic programming using FEC 1 - Cepstrum parameters (6 out of 12) | 73   |
4.2  | Feature subsets selected through dynamic programming using FEC 1 - IFSC parameters (16 out of 32) | 74   |
4.3  | Identification scores (error %) based on feature selection obtained using FEC 1 - Time domain parameters | 75   |
4.4  | Identification scores (error %) based on - vili -
Chapter I
INTRODUCTION

1.1 PREVIEW

Human beings have long had the ability to recognize a person solely from his voice. This property of speaker recognition possessed by human beings has been the subject of several investigations over the years[1],[2]. This remarkable ability of human beings does, however, have certain limitations. For example, human beings tend to forget voices as time passes, there is a limit on the number of voices one can remember, and mimics can sometimes make one believe one is listening to someone, when, in fact, one is not.

Since the early sixties efforts have been underway to investigate the feasibility of developing objective computer-based voice recognition algorithms. While the digital computer cannot, as yet, duplicate the complex perceptual base of speaker recognition as performed by the human mind, it can overcome some of its limitations mentioned above.

Man's most natural form of communication with others is through speech. It would be extremely convenient if voice communication could also be used to interact with machines.
Computerized voice recognition systems would serve as one mode in a three-mode man-machine speech communication system as visualized by Flanagan[3]. The three modes are:

1) Voice response systems
2) Speaker recognition systems
3) Speech recognition systems

According to Flanagan[3], "Research coming to fruition over the past several years has made clear that the techniques of man-machine communication by voice constitute a whole new range of communication services - services that can extend man's capabilities, serve his social needs, and increase his productivity. To insist that machines communicate with man on a level he finds effective, namely, by natural voice, seems a fitting requirement for machines - as the servants of man. The ability to speak makes the immense sophistication of modern computers immediately available over the ubiquitous voice channel, and from a terminal no more complicated than an ordinary telephone set."

There are several potential areas of application for speaker recognition. These include automated commercial business and banking systems for telephone-based transactions, regulated access to confidential information or restricted premises to authorized individuals and in the forensic sciences[4].
1.2 Problem Statement

The main objective of this thesis was to examine the feasibility of text-independent speaker recognition in a controlled setting characterized by the following features:

a) A population consisting of 12 cooperative speakers. This choice of population size was based on a consideration of system constraints which, for the main part, was the size (32K) of main memory available, and, to a lesser extent, computation time.

b) A relatively noise-free speech data base recorded in a normal quiet room environment.

c) Use of short-duration speech segments for obtaining the acoustic characterization of the reference and test speakers. This last feature is one that is desirable from the point of view of practicality in most, if not all, common applications of speaker recognition systems.

An answer to the question of feasibility was sought in terms of the following aspects:

a) What is the relative effectiveness of the various possible parametric representations of the speakers in the two tasks of speaker identification and speaker verification?

b) Is there a potential for feature selection among the elements of each parameter set as a means of improving recognition scores?
c) Is it possible to improve recognition scores through the use of a sequential or multi-level decision strategy?

1.3 OUTLINE OF THESIS

This thesis is organized in the following manner. Chapter II carries a discussion of the terminology and methodology of speaker recognition and provides a brief survey of the literature outlining the techniques used and the results obtained in the past. The results of a preliminary evaluation of the ability of several different sets of speech parameters in the time and frequency domain to discriminate between different speakers forms the subject of Chapter III. The possibility of improving recognition scores through feature selection of the elements of a parameter set is discussed in Chapter IV. Speaker recognition schemes based on the nearest neighbour classification rule are investigated in Chapter V. This technique calls for a multi-reference characterization of each speaker in the population. In Chapter VI a composite scheme for speaker recognition is proposed. This scheme involves logical combination of decisions from two independent tests performed with two different speech parameter sets to make a recognition decision. Finally, the contributions of this work, the results obtained and the conclusions arrived at are summarized in Chapter VII.
Chapter II
SPEAKER RECOGNITION

2.1 INTRODUCTION

This chapter carries a detailed description of the mechanics of speaker recognition - how it is performed, the problems peculiar to it and the results obtained in the past by earlier research workers in the area. The steps involved in performing the speaker recognition task are listed in Section 2.2. The speaker recognition task can be thought of in terms of the two possible sub-tasks of speaker identification and speaker verification and these are discussed in Section 2.3. An important step in speaker recognition is to be able to find a measure of dissimilarity (commonly known as distance) between the acoustic features characteristic of the reference and test speakers. Section 2.4 lists the various distance measures in use. The sources of inter-speaker variability which make two different people sound different even when reading the same text are discussed in Section 2.5. Section 2.6 carries a parallel discussion of the sources of intra-speaker variability. While inter-speaker variability simplifies the task of speaker recognition, intra-speaker variability makes it more difficult. The chapter concludes with a survey of
the efforts of earlier workers in the area in Section 2.7, highlighting the techniques used and the results obtained in the past.

2.2 GENERAL PROCEDURE FOR SPEAKER RECOGNITION

Speaker recognition is essentially a pattern recognition problem. The pattern in question in this case is the speech or voice pattern of a particular speaker. Most speaker recognition systems that have been proposed or developed utilize the following procedure:

a) Derivation of a set of acoustic features that are characteristic of the voice pattern of each speaker in the population from his training speech database.

b) Derivation of the acoustic features characteristic of the unknown test speaker's voice pattern from an analysis of his test speech.

c) Derivation of a quantitative measure of similarity or dissimilarity (generally known as distance) between the reference speaker's acoustic features and the test speaker's acoustic features.

d) A recognition decision based on the similarity/dissimilarity measure.

The flow chart of Figure 2.1 illustrates the steps described above. The following geometric interpretation enables better understanding of the general procedure for speaker recognition. The extraction of acoustic features
Figure 2.1. General procedure for speaker recognition.
that are representative of a speaker's identity from a segment of his speech could be thought of as mapping the segment into a point in a multi-dimensional parameter space[5]. Different segments of the same speaker's speech would, as a consequence of intra-speaker variability (this will be discussed in greater detail later on), generate points distinct from each other. Hence each speaker would be characterised by, ideally, a closely-grouped cluster of points in the parameter space. Ideally we would also like the clusters corresponding to different speakers to be concentrated at different locations without any overlap. In practice there is an overlap and this results in errors in the speaker recognition task.

Wolff[6] outlined a set of desirable properties that speaker-characterizing acoustic parameters should possess. He suggested that, "ideally, the speech characteristics measured should:

i) occur naturally and frequently in normal speech.

ii) be easily measurable.

iii) vary as much as possible among speakers, but be as consistent as possible for each speaker.

iv) not change over time or be affected by the speaker's health.

v) not be affected by reasonable background noise nor depend on specific transmission characteristics."
not be modifiable by conscious effort of the speaker, or, at least, be unlikely to be affected by attempts to disguise the voice."

Steps (c) and (d) of the general procedure above require some elaboration but before proceeding further it is necessary here to discuss the two different ways in which speaker recognition can be applied.

2.3 CLASSIFICATION OF SPEAKER RECOGNITION

Speaker recognition can be classified into the distinct subareas of speaker verification and speaker identification.

2.3.1 Speaker Verification

In speaker verification the unknown test speaker provides a sample of his voice and claims an identity from among the n-speaker population. The verification system has to extract the set of acoustic features that relate to the unknown speaker's identity from his test speech sample and compare them with the stored reference acoustic features of the speaker whose identity is being claimed. If the two patterns are similar enough the identity claim must be accepted, otherwise rejected. This binary nature of the decisions that a speaker verification system is called upon to make automatically suggests that two types of error probabilities would characterize the performance of such a system, viz., false acceptance and false rejection. False
acceptance is the error resulting from acceptance of a claim that is not genuine. False rejection is the error resulting from rejection of a bona fide claim.

It is possible to express the problem of speaker verification in more mathematical terms [3], [7], [8]. Let $p_i(X)$ represent $p(X|i)$, the conditional probability density function of the vector of measurements for the $i$th class (speaker). Let $p_{av}(X)$ represent the average joint probability density function over the entire population of users. Physically it signifies the average proximity of the patterns of the user population and can be obtained as follows:

$$p_{av}(X) = \sum_{i=1}^{N} p_i(X)p_i$$   \hspace{1cm} (2.1)

where $p_i$ is the a priori probability of occurrence of a sample of class 'i'.

The decision rule for verification can be stated as,

Accept speaker 'i' if $p_i(X) > c_i p_{av}(X)$   \hspace{1cm} (2.2)

Reject speaker 'i' if $p_i(X) < c_i p_{av}(X)$

c_i is a threshold parameter for the $i$th speaker whose value enables control of the error mix of false acceptance and false rejection. The exact value of $c_i$ chosen would be determined by the relative penalties attached to false acceptance and false rejection in the event of failure of
the verification system in its specific area of application.
It is obvious that there will be a certain value of the
threshold for which the rate of 'false alarms' is equal to
the rate of 'misses'. This is called the equal-error


criterion and is often used as an index of performance when
testing a speaker verification system.

2.3.2 Speaker Identification

In speaker identification the unknown test speaker
provides a sample of his speech but makes no identity claim.
The identification system has to answer the question, "WHO
IS HE?". It must do this by extracting the set of
acoustic features that characterise the unknown speaker's
voice from his test speech sample and comparing them with
the stored reference feature sets of each of the N speakers
in the population to find the speaker with the most similar
reference pattern.

The joint probability density of $X$, the vector of
measurements for the unknown speaker, occurring and being a
member of the $i$th class is given by,

$$p(i,X) = p_i p_i(X)$$

(2.3)

The decision function that minimizes the risk of
misclassification assigns the measurement vector $X$ to class
' $i$ ' if [7][8],

$$p(i,X) > p(j,X) \text{ for all } i \neq j$$

(2.4)
Under the assumption that the a priori probabilities of occurrence of a sample from each of the \( N \) classes are equal, the above decision rule takes on the form:

\[
\text{Choose speaker 'i' such that } p_i(X) > p_j(X) \quad (2.5)
\]

\[j=1,2,...,N; \; j \neq i\]

It is to be noted that while a verification test involves only one comparison of a pair of feature sets an identification test involves \( N \) comparisons between pairs of feature sets. In terms of our earlier geometric visualization of feature extraction as a mapping of a segment of speech from a speaker into a point in a multidimensional parameter space, it can be seen that as the population of users \( N \) increases and becomes very large "the probability of error must tend to one since an infinite number of distributions cannot remain distinct in a finite parameter space"[7]. As opposed to this, in the case of speaker verification, it can be shown that the probability of the two types of errors is essentially independent of the population size. Doddington[9] carried out a computer simulation to obtain expected error rate as a function of the population size for the two tasks of speaker verification and speaker identification. It was assumed that all the speaker measurements were normally distributed. The results of his study are presented in Figure 2.2 and essentially show that the error rate in speaker verification
Fig. 2.2 Error rates in speaker identification and speaker verification vs. population size N (after Doddington [9])
flattens out as \( N \) increases whereas in speaker identification it continues to increase.

The essential differences between the tasks of speaker identification and speaker verification discussed above are summarised in Figure 2.3 which is from the paper by Flanagan[3].

As mentioned earlier, the third step in the general procedure for speaker recognition is to obtain a distance measure quantifying the degree of dissimilarity between the reference feature vector and the test feature vector. In the next section the most common distance measures that are used in speaker recognition are discussed.
Fig. 2.3. Summary of differences between speaker identification and speaker verification (after Planagan [3])
2.4 DISTANCE MEASURES

The simplest distance metric than can be conceived of is perhaps the Euclidean measure given by [10], [11]:

\[ d_i(X) = (M_i - X)^T (M_i - X) \]  \hspace{1cm} (2.6)

where,

- \( X \) is an L-dimensional column vector representing the unknown test speaker.
- \( M_i \) is an L-dimensional column vector representing the mean of the reference feature vectors of the \( i^{th} \) speaker and can be obtained as follows:

\[ M_i = \frac{1}{R} \sum_{r=1}^{R} M_{ir} \]  \hspace{1cm} (2.7)

where, \( M_{ir} \) is the \( r^{th} \) vector in the training set of speaker \( y \) and \( R \) is the total number of such vectors in the training set of speaker \( i \).

\( d_i(X) \) is the distance between the unknown speaker's feature vector and the mean reference feature vector of speaker \( i \).

The Euclidean distance metric is the distance between the points represented by \( M_i \) and \( X \) in the multi-dimensional feature space.
A distance measure performs the function of yielding one number that represents the similarity/dissimilarity between two feature vectors the components of which may be a varied set of measurements which have different physical significance. For instance, a pattern vector may be composed of pitch, energy, spectral parameters and linear prediction coefficients. Some form of weighting (or normalization) is necessary in order to assign each dimension of the feature vector just the right importance. It can be seen that the Euclidean distance metric assigns equal weightage to each component of the feature vector.

We will now discuss a more general weighted Euclidean (sometimes known as non-Euclidean) distance metric. This measure turns out to be a logical consequence of the maximum likelihood principles used in the formulation of the decision rules for identification along with the assumption of Gaussian statistics for the observation vectors. This can be seen as follows[7]. The cluster of points in the multi-dimensional parameter space that characterise the feature vector observations for a particular speaker in the population can be modelled by a multivariate conditional probability density function. In principle either a parametric or non-parametric pdf can be used. However reliable estimation of the non-parametric pdf requires a large number of observations on the feature vector even in the case of a small parameter set. It is more convenient to
choose a parametric representation of the multi-dimensional conditional probability density function such as the Gaussian which is completely specified by the mean and covariance matrix of the elements of the feature vector.

The $1$-dimensional Gaussian conditional probability density function of the $i$th speaker's observations is given by,

$$p_i(x) = (2\pi)^{-N/2} |W_i|^{-1/2} \exp \left[ -\frac{1}{2} (M_i - x)^T W_i^{-1} (M_i - x) \right]$$  \hspace{1cm} (2.8)

where,

$W_i$ is the covariance matrix of the elements of the $i$th speaker's feature vector computed over the training set and is given by,

$$W_i = \frac{1}{R} \sum_{r=1}^{R} M_{ir} M_{ir}^T - M_i M_i^T$$  \hspace{1cm} (2.9)

$|W_i|$ is the determinant of $W_i$.

$M_{ir}$ and $M_i$ have the same significance as discussed in connection with equation (2.6).

The decision rule of equation (2.5) can now be applied to the density function of equation (2.8). By taking the natural logarithm of both sides of the inequality of equation (2.5) and noting the negative sign of the exponent
of equation (2.9), the decision function can be manipulated to the following form under the assumption that the a priori probability of occurrence of each class is the same:

\[ d_i(X) = \frac{1}{2} (M_i - X)^T W_i^{-1} (M_i - X) + \frac{1}{2} \log |W_i| \]  \hspace{1cm} (2.10)

\[ \forall i \neq j \]

The last term of the above equation is dependent on \('i'\) alone and is not a function of \(X\). The inclusion of \(|W_i|\) does not, in practice, appear to influence the decision process. Hence the form of the distance measure used in the decision function is given by,

\[ d_i(X) = (M_i - X)^T W_i^{-1} (M_i - X) \]  \hspace{1cm} (2.11)

This expression for distance is the well-known Mahalanobis distance metric. This distance metric has the important and convenient property that \(d_i(X)\) is invariant with respect to any arbitrary non-singular linear transformation of the feature set[5].
It can be seen that if the covariance matrix $W_i$ is replaced by an identity matrix, the Mahalanobis distance metric is identical to the Euclidean distance metric of equation (2.6). Thus the Euclidean distance measure is a special case of the more general Mahalanobis distance metric, where equal weightage is given to each dimension of the feature vectors and an implicit assumption is made that the components of the feature vectors are not correlated with each other.

Several variations of the Mahalanobis distance metric have also been used in the literature. In one form the individual intra-speaker covariance matrices $W_i$ are replaced by a pooled (over the entire speaker population) intra-speaker covariance matrix $W$. The latter can be obtained as follows:

$$W = \frac{1}{N} \sum_{i=1}^{N} W_i$$  \hspace{1cm} (2.12)

A second variation involves a further computational simplification of the Mahalanobis distance metric by replacing $W_i$ by a matrix constructed from just the diagonal elements of $W_i$ as follows:

$$A = \text{DIAG}(W_i)$$  \hspace{1cm} (2.13)
This has the effect of weighting the components of the feature vector by the inverse of the calculated variances of each component over the training set of each speaker thus giving those components which are strongly clustered a greater say in determining the overall distance. As in the case of the Euclidean distance measure of equation (2.6), use of this form is justified when the off-diagonal elements of $W_i$ are small, signifying small correlations between different components of the feature set.

A third form is a pooled version of $A_i$ of equation (2.13) obtained as follows:

$$A = \frac{1}{N} \sum_{i=1}^{N} A_i$$  \hspace{1cm} (2.14)

Another distance measure uses the crosscorrelation $\rho_i(X)$ between the reference feature vector $M_i$ and the test feature vector $X$ as a measure of similarity/dissimilarity between them \[5\]. It can be computed as follows:

$$\rho_i(X) = \frac{(X^T M_i)}{\left[ (X^T X) (M_i^T M_i) \right]^{\frac{1}{2}}}$$  \hspace{1cm} (2.15)

As can be seen from the above equation $\rho_i(X)$ is also the cosines of the angle between the vectors $M_i$ and $X$ in the multi-dimensional parameter space.
The classification rule represented by the inequality of equation (2.10) is known as the minimum distance classifier. Each speaker is represented by the mean of the feature vectors in the training set and the distance between the unknown speaker's feature vector and the mean of the reference speaker's feature vector is used to make a decision. As mentioned earlier, the minimum distance classifier is a parametric method and is a consequence of modelling the distribution of each speaker class by means of a multivariate Gaussian probability density function. On the other hand, in the non-parametric methods, no assumption is made regarding the form of the distribution of each class. One such method is the nearest neighbour classification technique.[8],[12]. In this method each speaker in the population is represented by a cloud of reference points derived from his training speech database. The nearest neighbour of the unknown speaker's feature vector is found from amongst the entirety of pattern points categorizing each reference speaker. The unknown test speaker is identified as the reference speaker from whose cloud of reference points the nearest neighbour originated. The nearest neighbour decision rule can be expressed mathematically as follows. Let \( s_i \) represent the collection of \( L \)-dimensional column vectors \( (M_{ir}, r=1,2,\ldots,R) \) in the \( i \)th speaker's training set. The decision rule for identification is then given by,

Choose speaker 'i' such that \( \min_{i,r} d(M_{ir}, X) \in s_i \) (2.16)
where \( d(M_{ir}, X) \) is the distance between the unknown feature vector \( X \) and the \( r \)th reference feature vector of the \( i \)th speaker, \( M_{ir} \) and the symbol \( \varepsilon \) is used to indicate 'belonging to'.

2.5 SOURCES OF INTER-SPEAKER VARIABILITY

The speech produced by two different people even when reading the same text sounds different to the ear. This is known as inter-speaker variability. There are two basic factors that cause variations in the speech patterns of different individuals\([5],[6] \):

i) Structural differences in the vocal tract
The size and shape of the vocal tract, which is the speech producing mechanism, differs from one person to another.

ii) Differences in the speaking habits
The speaking style differs from person to person and is generally a product of the environment in which one learns how to speak.

Closely related to the above issue is the question of whether one should use time-varying acoustic features (commonly known as dynamic features) or speech or time-averaged features (statistical features) to represent a speaker. The former is meaningful only in the case where the same prescribed text is used for all speakers. The use of dynamic acoustic features as they vary over standard
sentence-long utterances enables one to exploit differences in the speaking habits of different individuals for speaker recognition. However in this study we concentrated on the problem of text-independent speaker recognition and consequently we are constrained to using statistical parameters of speech. These parameters are inherently suitable for text-independent speaker recognition since the phonetic or message content is removed by extensive averaging over long sections of speech.

2.6 SOURCES OF INTRA-SPEAKER VARIABILITY

Because of the random nature of the speech process there is a scatter in the feature vector observations from different segments of speech of the same speaker. Some of the sources of this intra-speaker variability are differences in the emotional state, health and speaking rate of the speaker while recording different segments of reference or test speech. Intra-speaker variability could also be caused by differences in the recording conditions during different recording sessions and differences in the texts of the different reference and test speech segments. This latter source of intra-speaker variability can be reduced by imposing a requirement that the texts of the different reference segments and the test segment be the same. This possibility allows for a basic dichotomy in the speaker recognition task into text-dependent speaker
recognition and text-independent speaker recognition. In text-dependent speaker recognition the texts of the reference and test segments of speech are constrained to be the same and are normally sentence-long utterances. In text-independent speaker recognition these texts are different and, in general, are totally random.

In the next section a review of the relevant literature in the area of speaker recognition is presented in order to highlight some of the approaches adopted in the past and the unique problems associated with speaker recognition. This also serves to bring into focus the state of the art.

2.7 LITERATURE REVIEW

A survey of the literature reveals the existence of quite a few text-dependent speaker recognition systems with accuracies better than 99%. However, there are very few high accuracy text-independent speaker recognition systems which do not have to resort to extensive averaging of feature vectors over extremely long segments of speech.

A wide range of voice characteristics have been used in the past for speaker recognition. These include such features as pitch, intensity, linear prediction coefficients, vocal tract resonances and area functions, formants, spectral parameters, cepstral coefficients, etc. Atal[5] and Rosenberg[11] have conducted an extensive survey of the parameters used and their performances in speaker
recognition tasks. This survey will list only a few of the more conspicuous research efforts in this area which are pertinent to our study.

Atal (13) investigated the effectiveness of pitch contours in text-dependent speaker recognition. The fundamental frequency of vocal-cord vibrations as a function of time over a short utterance was used to characterize each speaker. The performance of several different recognition schemes based on the non-Euclidean (with pooled intra-speaker covariance matrix), Euclidean and crosscorrelation distance metrics was evaluated. While the Euclidean and crosscorrelation measures yielded accuracies of 68% and 70% respectively, the non-Euclidean measure performed considerably better with an identification accuracy of 97% for a speech utterance about 2 seconds in duration. Speech synthesis techniques using the linear prediction parameters have shown a remarkable ability to reproduce the original voice. This observation has resulted in several studies which have examined the performance of the linear prediction parameters of speech in speaker recognition experiments.

Atal (14) performed a comparative assessment of the relative effectiveness of the linear prediction characteristics of the speech wave and several other parametric representations derived from them through non-linear transformations, using the same set of data as in the earlier study based on pitch contours. The parameter sets obtained by transforming the
predictor coefficients included the impulse response corresponding to the transfer function \( H(z) \) of the vocal tract, the autocorrelation function of the impulse response, the area function and the cepstrum coefficients. The non-Euclidean distance metric with a pooled intra-speaker covariance matrix was used in the identification scheme. The identification scores were computed as a cumulative function of the number of frames of the prescribed utterances included in computing the distances. The highest accuracies were obtained for the cepstrum coefficients and yielded scores better than 98\% for speech durations greater than 0.5 seconds. The pitch contours yielded an identification accuracy of 97\% for a sentence about 2 seconds long. Based on these results Atal[5] suggested that the vocal tract characteristics may be more effective for speaker identification than the source characteristics. The cepstrum parameters were also tested in a speaker verification experiment which yielded an accuracy of 90\% for a speech duration of 0.2 seconds rising to 98\% for a duration of 1 second.

Sambar[15] proposed the use of orthogonalised speech parameters for speaker recognition. This technique is basically very similar to the techniques of principal component analysis[16] and the Karhunen-Loève expansion[17]. The method uses the eigenvectors of the covariance matrix of the elements of the feature vector estimated over the
reference utterances of a given speaker to orthogonally transform both his reference feature set and the test feature set of the unknown speaker to orthogonal space. The distance between the test and reference parameters is then evaluated in the orthogonal space using the eigenvalues of the covariance matrix as a normalizing factor for each dimension in the orthogonal space. It is appropriate to mention here that it can be shown that when all the orthogonal parameters are included in computing the distance measure the result is identical to using the Mahalanobis distance metric of equation (2.17) in the original feature space. Text-dependent speaker recognition experiments were conducted on a population of 21 male speakers using a database consisting of repetitions of a short sentence recorded over a period of three weeks. The orthogonal reflection and orthogonal log area ratio coefficients outperformed the orthogonal linear prediction coefficients yielding scores better than 99% in identification and verification experiments. On the basis of his results Sambur surmised that the least significant orthogonal parameters, which exhibit small variances across the analyzed utterance, were indicative of the talker's identity, whereas the most significant parameters with the largest variances, were reflective of the linguistic content of the utterance. An important observation that can be made from Sambur's results is that inclusion of the orthogonal parameters with large
variances in the distance measure, while not improving recognition accuracy, certainly did not cause any deterioration in performance.

Several recognition schemes have used the short-time power spectrum of speech to discriminate between speakers. The work of Pruzansky and Mathews[18] and Das and Mohr[19] were some of the earliest efforts in this area. In the former the spectrogram of single words excerpted from sentences of 10 talkers was obtained by passing them through a 17-channel filter bank covering the frequency range from 100 Hz to 10 KHz and sampling every 10 msec. Elemental energies in one frequency channel and one time interval of the spectrogram were used as features. The distance metric used was a crosscorrelation measure between the spectrograms of the reference and test words. A maximum recognition accuracy of 88% was obtained when all the features were used. This recognition accuracy was maintained even when only 25% of the features selected on the basis of their F-ratios[5] (ratio of inter-speaker to intra-speaker feature variance) were used in making a decision. The effect of averaging groups of elemental energies over a rectangular area of the spectrogram formed from several adjacent frequency channels and time intervals was also investigated. It was shown that while averaging of elemental energies over time improved performance, averaging over frequency worsened it.
Das and Nohr[19] used an extensive data base of phrases recorded from 50 real speakers and 68 impostors over a period of about 5 weeks. The spectrum of the speech was obtained by passing it through a bank of 20 bandpass filters ranging from 188-8203 Hz and sampling every 20 msec after full-wave rectification and smoothing. An elaborate segmentation procedure was used to isolate various phonetic speech events in the utterances. These events were used to align the different utterances exactly such that identical sections or these utterances could be compared. Based on the segmentation points, 405 features were calculated for each utterance, such as filter averages combining the outputs of several filters around segmentation points, filter averages of linear time-normalized versions of the same data performed in order to compensate for the differences in length of the various repetitions of the same utterance, formant information from both the original and time-normalized data, time differences between adjacent pairs of segmentation points, the mean and variance of fundamental frequency, etc. Using the F-ratio criterion a subset of 200 or the 405 features was selected. A mean error rate of misclassification of 1% was obtained. The authors also demonstrated through experiments the advisability of constructing the reference features of each speaker from speech data collected over an extended period of time.
Furui et al.[20] conducted recognition experiments using the cepstrum of the long-time average of the power spectrum of a short paragraph (10 seconds long) as speaker-characterizing parameters. They investigated the variation in the long-time average spectrum pattern over an extended period of time of more than a year. In addition to the effect on the recognition rate of various factors such as the distance measure used, the time distribution of the database used in formulating a reference feature set for each speaker, and the time interval between the last known sample used in constructing the reference feature set and the test sample were examined. The distance measures considered were the Euclidean metric of equation (2.6), the different forms of the non-Euclidean metric represented by equations (2.11), (2.12), (2.13) and (2.14), and the nearest neighbour method of equation (2.16). The broad conclusions arrived at were as follows:

i) The long-time averaged spectrum pattern exhibits a tendency of gradual unidirectional variation which is significant for a period of about 3 months for almost all the speakers considered in the study and for even longer periods for some of them.

ii) The use of weighted distances obtained from suitably time-spaced data compensates for this spectrum variation and enables recognition scores to be maintained even as the time interval between
the last known reference sample and the test sample increases to about a year.

iii) Tests conducted using the long-time average spectrum taken four times at 3-month intervals to construct the reference feature set yielded scores of 91% in identifying 1 out of 9 talkers. Verification scores of 93% to 95% were obtained, corresponding to the use of either 8 registered or 26 unregistered talkers in the population.

Another technique based on the location and isolation of acoustic events with a powerful potential for speaker discrimination is the nasal coarticulation scheme proposed by Su et al [21]. It is based on the premise that in connected speech, as a result of the inertia of the articulators, the vocal tract shape is not only a function of the current phoneme but also of neighbouring phonemes (coarticulation). The difference between the mean spectrum of a nasal consonant followed by a front vowel, and that of the same consonant followed by a back vowel, was used as an acoustic measure of nasal coarticulation in a consonant-vowel context. Identification scores of 100% were obtained with a group of trained speakers, but the authors did suggest that with a larger untrained population, this was not to be expected. They also concluded that the nasal coarticulation measure is superior to measures based directly on nasal spectra. It is appropriate to mention
here that most schemes involving location of nasal events rely on visual observation to perform the segmentation, an impractical proposition.

Bunge[22] developed a system called AUROS as a research tool for investigating the effect of various processing aspects on speaker recognition performance. Tests conducted using a large speech data base yielded identification and verification error rates of less than 1% for both text-dependent and text-independent speaker recognition. From his results Bunge concluded that parameters for speaker recognition derived from spectral features are superior to time domain parameters. It was also claimed that a minimum length of 11 seconds of test speech data was essential for reliable text-independent speaker recognition.

One of the earliest investigations of text-independent speaker recognition was conducted by Atal[14]. The text-independent data base for this study was created in an artificial manner from the utterances for the text-dependent investigation mentioned earlier. The 40 segments into which each utterance was divided, were recombined in a random fashion to destroy the alignment in the texts of the utterances. The features used were 12 cepstral coefficients for each of the 40 segments. The distance measure used was the non-Euclidean metric with pooled intra-speaker covariance matrix of equation (2.12). The contributions from each one of the segments of the utterances was combined
to obtain a composite measure of dissimilarity between utterances. An identification accuracy of 93% was obtained for speech 2 seconds in duration.

Sambur[15] conducted a preliminary investigation, using a rather limited text consisting of six different sentences, to investigate the suitability of using the technique of orthogonal linear prediction for text-independent speaker recognition. The reference feature set was derived from data obtained from five out of the six utterances with the sixth used as test data. The identification scores varied depending on which sentence was used as the test data. This perhaps indicated the presence of residual linguistic effects due to inadequate averaging in obtaining the reference feature set from just five utterances. An overall identification accuracy of 94% was obtained.

R.E. Wohlford et al[23] used a common speech data base to rank the effectiveness of four techniques of automatic speaker identification proposed by earlier workers in a text-independent mode. These techniques made use of four different feature sets such as the correlation of short and long-term spectral averages of Pruzansky and Mathews[18], cepstral measurements of long-term spectral averages of Furui et al[20], orthogonal linear prediction of Sambur[15] and long-term average LPC reflection coefficients, pitch and overall gain of Markel et al[24]. They found that the two techniques based on cepstral and spectral data did not
perform as well as the LPC-based systems of Sambur and
Markel which yielded recognition accuracies of 93% and 95%
respectively with 10 minutes of reference speech and 13
seconds of unknown test speech.

Several workers have investigated the potential of
spectral parameters of speech for text-independent speaker
recognition. Li and Huqmes[25] attempted to quantify inter-
speaker and intra-speaker differences based on correlation
matrices derived from continuous speech spectra by using
measures of difference between matrices. They found that a
minimum of about 30 seconds of test was needed to ensure
stability of the correlation matrices. Identification and
verification tests using measures of difference between the
spectral correlation matrices yielded a minimum error rate
of between 1% and 3%.

Markel and Davis[26] have made a significant
contribution to the area of text-independent speaker
recognition. They used an extensive data base of over 36
hours of conversational speech recorded from 17 speakers
over a period of over 3 months. The reference parameters
for each speaker were obtained by extensive averaging of the
feature vectors over approximately an hour's time-spaced
speech data for each speaker. An individual feature vector
was itself obtained by averaging a fundamental acoustic
parameter set over 1000 voiced frames each of length 20
msec. On the average about 39 seconds of speech were
required in order to obtain 1000 voiced frames. The
distance metric used in recognition tests was the
Mahalanabis distance metric or equation (2.11). Using a
22-feature reference parameter set composed of the mean and
standard deviation of the fundamental frequency and the mean
and standard deviations of 10 reflection coefficients, an
identification accuracy of 98% and an equal-error
verification rate of 4.25% was obtained. It was concluded
that averaging or acoustic parameters over long intervals of
speech of duration 30 seconds or more was sufficient to
eliminate the phonetic content of speech from these
parameters thus rendering them suitable for text-independent
speaker recognition.

Schwartz et al [27] compared the performance of several
probabilistic classifiers to the Mahalanabis minimum
distance classifier in a text-independent speaker
identification task. In the probabilistic classification
techniques each speaker's voice is modelled by a probability
density function estimated from the training data base of
that speaker. Of the three PDF estimation techniques
investigated, two were parametric and one non-parametric.
In the parametric methods the distribution of each class was
modelled by a Gaussian PDF and a Gaussian PDF modified by a
soft clipping function respectively. The non-parametric
method which is distribution-free was a variation of the K-
nearest neighbour technique, wherein the shape of the
distribution for each speaker is represented by the entire collection of reference points in his training set. The results obtained using a minute of speech data collected in a single session from each of 21 speakers indicated that the probabilistic classifiers performed considerably better than the Mahalanobis distance classifier. Among the probabilistic classifiers the non-parametric technique did better than the parametric methods.

In a follow-up effort Wolf et al. [28] applied the same techniques to a speech data base recorded over a radio channel. This data was gathered over a period of time and enabled multi-session experiments to be performed. The results again revealed that the probabilistic classifiers performed better than the Mahalanobis classifier.

Two large-scale functional text-dependent speaker verification systems operating under "real-world" conditions are in existence. These are the Texas Instruments Entry Control System [11] and the Bell Labs Automatic Speaker Verification System [11]. Both these systems have undergone several phases of evolution as a consequence of performance testing, evaluation and tuning based on extensive amounts of data from fairly large speaker populations gathered over an extended period of time. In fact they were the outcome of the developmental efforts of several different workers whose work is part of the published literature in the area of speaker recognition, viz., [29],[30],[31].
The Texas Instruments System makes use of spectral amplitude information on certain specific sections of the 4-word phrases used as input data. Readers are referred to the references mentioned above for details. Only the salient features will be mentioned here. One of the main features of this work was the incorporation of a sequential multiple-phrase decision strategy that made use of as many phrases as were necessary to establish a certain level of confidence. This technique resulted in considerable improvement in performance. Using a 4-phrase sequential decision strategy a practical implementation yielded a false rejection rate of 0.3% and a false acceptance rate of 1% with an average of 1.3 utterances being required per transaction to meet the desired confidence level. The need for periodic updates of the speaker reference files was also established.

The speech parameters used in the Bell Labs System were dynamic contour measurements of pitch, intensity, formant frequencies and linear prediction coefficients as they varied over sentence-long utterances. An important feature of this system was the selection of speaker-dependent subsets of segments of utterances and measurements for the calculation of overall distance that ensured maximal separation of the overall distance distribution of each genuine speaker's utterance from the impostor utterances. The evaluation of this system was done both with high-
quality speech recordings and recordings made over dial-up telephone lines. In the case of high-quality speech utterances an equal-error rate of 1.5% was obtained. For telephone bandwidth utterances the equal-error rate was initially quite high (about 10%) but stabilized to about 4% after a period of adaptation.

From the review presented above it is quite evident that, in general, the accuracy obtainable in text-independent speaker recognition systems, without having to resort to extensive averaging (an impractical proposition) is rather limited with the existing speaker-characterizing parameters.
Chapter III

SPEAKER RECOGNITION POTENTIAL OF SPEECH PARAMETER SETS

3.1 INTRODUCTION

This chapter discusses the results of a comparative evaluation of the potential for speaker recognition of several different feature sets derived from both spectral and time domain information. Section 3.2 provides details of the acquisition of the speech data base and the preprocessing performed on the speech prior to implementation of the recognition algorithms. The speech parameter sets investigated for their speaker recognition potential are listed in Section 3.3. Principal component analysis was used to transform the speech parameter sets to orthogonal space in which each element of a set was uncorrelated with every other element of the set. The process of orthogonalization of the parameter sets is described in Section 3.4. It will be seen from this section that a critical step in the orthogonalization process is to obtain a reliable estimate of the covariance matrix of the elements of the feature set. This amounts to a proper choice of the minimum duration of the speech utterances across which the covariance matrix is to be evaluated.
Section 3.5 attempts to answer the above question by presenting the results of an investigation of the convergence properties of the covariance matrix of the elements of the speech parameter set. The recognition tests conducted to evaluate the performance of the speech parameter sets are described in Section 3.6. The chapter concludes with a presentation of the results obtained and a discussion of the results in Section 3.7.

3.2 SPEECH DATA COLLECTION AND PREPROCESSING

The speaker population consisted of 12 speakers. Each participant was asked to read a magazine or newspaper or a passage from a text with which he was familiar. Speech recordings were made in a normal quiet room environment on a conventional tape recorder. The data base used in this study consisted of about 60 seconds of speech for each speaker from each of eight sessions spaced at least a week apart. The eight sessions spanned a period of about four months. The speech was band-limited between 150 Hz and 4.5 KHz and digitized at a sampling rate of 10 KHz using an A/D converter with 14 bit resolution. All the processing was done on a Data General NOVA 840 minicomputer equipped with an Array Processor (AP120B) for high speed signal processing.
3.3 EXTRAVATION OF SPEECH PARAMETERS

Two classes of parameters were used in this study:

a) Time domain parameters
b) Frequency domain parameters

3.3.1 Time Domain Parameters

The time domain parameters consisted of the following:

i) Linear Prediction Coefficients (LPC)
ii) Reflection Coefficients (REF)
iii) Log Area Ratio Coefficients (LAR)
iv) Cepstrum Coefficients (CEP)

The feature sets represented by (ii), (iii) and (iv) can be obtained through nonlinear transformations of (i). These transformations are discussed in Appendix I.

The entire speech was divided into 20 msec (200 sample) segments called frames. The frames corresponding to the silent portions, pauses and unvoiced segments in the digitized speech data were removed through the use of a simple energy threshold criterion[7]. The speech samples in the remaining frames were digitally pre-emphasized by a first order filter with transfer function of \((1 - 0.9375z^{-1})\) and subsequently multiplied by a Hamming window function. A 12th order linear prediction analysis using the autocorrelation method[10] was applied to the remaining frames amounting to a minimum of 1000 frames (20 sec) per session for each speaker.
3.3.2 Frequency Domain Parameters

Three different sets of frequency domain parameters were used. They were:

i) Inverse Filter Spectral Coefficients (IFSC)

These coefficients represent the magnitude spectrum of the linear prediction inverse filter. The inverse filter is characterized by,

$$\Lambda(Z) = 1 + \sum_{i=1}^{P} a_i Z^{-i}$$

(3.1)

where \( \{ a_i, i=1,2,\ldots,P \} \) represents the linear prediction coefficients and \( P \) is the chosen order of the linear prediction model.

The IFSC were obtained as follows[10]:

a) Compute the FFT of the sequence \( \{ 1, a_1, a_2, \ldots, a_p, 0, 0, 0, \ldots, 0 \} \) where 51 zeroes were appended to the sequence to obtain the required frequency resolution by increasing its length to 64.

b) Compute the spectral magnitudes of the 32 unique FFT coefficients.

c) Evaluate the logarithm of the magnitudes.

The above procedure essentially evaluates the spectral characteristics of the LPC model.

ii) Speech Spectrum Parameters (SSP)
These were obtained through a direct spectral analysis of speech as illustrated in Figure 3.1. The same speech utterances on which linear prediction analysis was performed, were passed through a bank of 16 bandpass filters. The filters were centred at frequencies ranging from 160 Hz to 5 kHz and were spaced a third of an octave apart. The outputs of the bandpass filters were rectified, smoothed and sampled every 10 msec (which is the frame size for this feature set). The frames corresponding to silent portions, pauses and unvoiced segments of speech were once again removed, this time making use of an energy threshold criterion in the spectral domain. The sampled values for the remaining frames represent the Speech Spectrum Parameters.

iii) Direct FFT of Speech (DFFT)

Short-time Fourier analysis was performed through the use of the FFT algorithm on the very same frames (of length 200 samples or 20 msec) on which LPC analysis was carried out earlier. Zeroes were appended to the speech samples to realize a 256-point sequence. The magnitudes of the Fourier coefficients were then computed and grouped into 32 4-point averages as follows:

$$X_n = \frac{1}{4} \sum_{k=1}^{4} |F_{n+k}|$$

$$n=1, 2, \ldots, 32$$
where,

\[ n_k = 4(n-1) + k \]  \hspace{1cm} (3.3)

and \( F_n \) is the \( n_{th} \) Fourier coefficient. The logarithm of the resultant set of 32 average magnitudes was taken to yield 32 DFT parameters.

### 3.4 Orthogonalization of Parameter Sets

The technique of principal component analysis for deriving uncorrelated parameters (also known as "orthogonal parameters") through a transformation of the original feature space consists of the following steps\([15],[16],[17]\):

a) Let \( \{x_{ik} : i=1,2,\ldots,M; k=1,2,\ldots,NF\} \) be the parameter set where \( x_{ik} \) is the \( i_{th} \) parameter of the \( k_{th} \) frame, \( M \) is the total number of parameters in the feature set and \( NF \) is the total number of frames in the training set of the speaker under consideration.

b) Compute the covariance matrix \( [C] \) of the parameter set where \( \{[C] : (c_{ij} : i=1,2,\ldots,M; j=1,2,\ldots,M)\} \) is given by,

\[ c_{ij} = \frac{1}{NF} \sum_{k=1}^{NF} (x_{ik} - \overline{x_i})(x_{jk} - \overline{x_j}) \]  \hspace{1cm} (3.4)

and,

\[ \overline{x_i} = \frac{1}{NF} \sum_{k=1}^{NF} x_{ik} \]  \hspace{1cm} (3.5)
is the average value of the \( i \)th parameter.

c) Compute the eigenvalues \( (\lambda_i, i=1,2,...,M) \) and the eigenvectors \( (T_{ij}, i=1,2,...,M) \).

d) Normalize \( T_{ij} \) to unit length.

e) Evaluate the orthogonal parameters \( (\phi_{ik}, i=1,2,...,M; k=1,2,...,NF) \) through the following linear transformation:

\[
\phi_{ik} = \sum_{j=1}^{M} t_{ji} x_{jk}
\]  

(3.6)

where \( \phi_{ik} \) is the \( i \)th orthogonal parameter in the \( k \)th frame and \( t_{ji} \) is the \( j \)th element of the \( i \)th eigenvector \( T_{ij} \). The orthogonal parameters thus calculated from each frame of speech data for a particular speaker constitute the elements of the new training set in the orthogonal domain.

f) The reference feature vector is the mean of the feature vectors in the training set and is given by,

\[
\bar{\phi}_i = \frac{1}{NF} \sum_{k=1}^{NF} \phi_{ik}
\]  

(3.7)

where \( \bar{\phi}_i \) is the average value of the \( i \)th orthogonal parameter.

q) The distance between the reference feature vector of speaker 's' and the test feature vector of the unknown speaker 'q' is given by,
\[ d_{sq} = \frac{M}{i=1} \frac{[\bar{\phi}_i^s - (\tilde{\phi}_i^s)]^2}{(\lambda_i^s)^2} \]  

(3.8)

where \((\cdot)_s\) and \((\cdot)_q\) refer to arguments for speaker 's' and speaker 'q' respectively, \((\bar{\phi}_i)_s\) is the mean value of the \(i^{th}\) orthogonal parameter for speaker 's', \((\tilde{\phi}_i)_q\) is the mean value of the \(i^{th}\) orthogonal parameter for speaker 'q' and is defined by,

\[ (\tilde{\phi}_i)_q = \frac{1}{NF} \sum_{k=1}^{NF} \sum_{j=1}^{M} [(t_{ji})^s (x_{jk})^q] \]  

(3.9)

\[ i=1,2,\ldots,M \]

3.5 **CONVERGENCE PROPERTIES OF THE COVARIANCE MATRIX**

It can be seen from the previous section that a crucial step in the computation of orthogonal parameters is to obtain a stable and accurate estimate of the covariance matrix of the speech parameters under consideration. This amounts to a suitable choice of the minimum duration of the speech utterance across which the covariance matrix is to be evaluated. The choice is governed by the convergence properties of the covariance matrix as computed using equations (3.4) and (3.5) with the number of frames being
the independent variable. The convergence properties were investigated by evaluating the change in the computed covariance matrix (as more and more frames of speech were included in its estimation) through the use of the 'Average Absolute Difference' (AAD) measure [25]. The AAD is defined as,

$$\text{AAD}(n) = \frac{2}{M(M+1)} \sum_{i=1}^{M} \sum_{j=1}^{M} |\gamma_{ij}(n+1) - \gamma_{ij}(n)|$$  \hspace{1cm} (3.10)$$

for \( n=1,2,\ldots,L-1 \)

where \( \gamma_{ij}(n) \) is the \( n \)th estimate of the \( ij \)th element of the normalized covariance matrix \([\Gamma]\).

The normalized covariance matrix, also known as the correlation coefficient matrix, is obtained by applying the following transformation [17] to the covariance matrix as computed using equations (3.4) and (3.5):}

$$\Gamma = A^{-1}CA^{-1}$$  \hspace{1cm} (3.11)

where \( A \) is a diagonal matrix whose elements are the standard deviations of the elements of the speech parameter set under consideration. This normalization causes the elements of the matrix to be less than 1 in magnitude, since any element \( \gamma_{ij} \) of \([\Gamma]\) represents the correlation coefficient between the \( i \)th and \( j \)th elements of the feature set.
In formulating the AAD the correlation matrix is considered to be a point in a feature space of \( M \times (M+1)/2 \) dimensions. Since \([C]\) and hence \([R]\) is a symmetrical matrix, only the upper diagonal elements need to be considered. The AAD measure is a function of both distance and angle between the two matrices being compared. It can be shown that its value ranges between 0 and 2 for correlation coefficient matrices.

In their work Li and Hughes[25] used the AAD as a measure of intra-speaker and inter-speaker difference between the correlation matrices of the speech spectra of 30 talkers. Data for the analysis of intra-speaker differences was based on recordings of a different segment of text on two separate occasions by 8 speakers while data for the analysis of inter-speaker differences was obtained from a single recording of the same segment of text by 30 speakers. The results obtained indicated the ability of the AAD measure to discriminate between spectral correlation matrices of different speakers reading the same text and spectral correlation matrices of the same speaker reading different texts.

A separation of 200 frames (4 seconds) was maintained between two consecutive estimates of \([R]\). Thus \( \gamma_{ij}^{(4)} \) was computed over \( 4 \times 200 \) (=300) frames of speech while \( \gamma_{ij}^{(5)} \) was computed over \( 5 \times 200 \) (=1000) frames of speech. In this study a total of 8000 frames per speaker (1000 frames/session
X 8 sessions) were utilized and this resulted in L being 40 (=8000/200).

Figure 3.2 illustrates the behaviour of the AAD measure as a function of 'n' when the covariance matrix was computed for the LPC feature set. Though this figure shows the results only for 3 speakers and the LPC feature set, similar results were obtained for the entire population and other parameter sets. It can be seen that the AAD is initially high but converges to 'mad level' after about 4000 frames of speech (corresponding to n=20) are utilized in the estimation of the covariance matrix. This indicates that if one goes by the AAD measure as a test for convergence of the covariance matrix, about 60 seconds of speech are required to obtain a stable estimate of [C]. However, in a recent study Shridhar et al.[32] used the technique of cluster analysis through functional mapping to show that in the case of some of the speakers in the population, an even larger database of 7000 frames (140 seconds) or more of speech may be necessary in order to obtain a consistent estimate of the covariance matrix. As a result it was decided to use 7000 frames of speech as the reference or training data with the test data coming from the remaining 1000 frames.

It can also be seen from Figure 3.2 that the AAD plot for speaker JD exhibits sharp fluctuations (for example at n=15 and n=30) perhaps reflecting a certain lack of consistency in this speaker's voice pattern. This
Figure 3.2. Convergence properties of covariance matrix of speech parameters.
conclusion appears to be borne out by an analysis of results of speaker recognition experiments, which seem to show a higher incidence of errors for such speakers. The details of the speaker recognition experiments that were conducted are provided in the next section.

3.6 SPEAKER RECOGNITION EXPERIMENTS

The reference feature set of each speaker in the population, consisting of the average values of the orthogonal speech parameters, was computed using equation (3.7). There were eight sessions of speech data in all. In the case of the LPC, REF, LAR, CEP, IFSC and DFFT parameter sets there were 1000 frames per session with each frame being 20 msec in length. In the case of the SSP parameter set a session consisted of 2000 frames, each of length 10 msec. Seven out of the eight available sessions were used as reference or training data. Consequently, the value of NF used in equation (3.7) was 7\times 1000 (=7000) frames or 140 sec of speech for the LPC, REF, LAR, CEP, IFSC and DFFT parameter sets and 7\times 2000 (=14000) frames or 140 sec of speech for the SSP parameter set.

The test data was obtained from the eighth session not used in computing the reference feature set. Each of the eight sessions was in turn used as a source of test data. Different lengths of speech were used to compute the test speaker's feature set using equation (3.9). This was done
in order to investigate the effect of the duration of the
test speech segment on recognition scores. The lengths of
the test speech segments used in various speaker recognition
experiments for all except the SSP parameter set were 100,
250, 500 and 1000 frames corresponding to speech of duration
2, 5, 10 and 20 seconds respectively. For the SSP parameter
set they were 250, 500, 1000 and 2000 frames corresponding
to speech of duration 2.5, 5, 10 and 20 sec respectively.

In identification experiments the distance between the
unknown test speaker and each reference speaker was
calculated using equation (3.8). The test speaker was
identified as the 'closest' reference speaker in accordance
with the decision rule of equation (2.10). In the case of
verification the distance between the test speaker and the
claimed reference speaker was compared with a threshold
value. Depending on whether the distance was smaller or
greater than the threshold, the test speaker's identity
claim was either accepted or rejected. Since there were 11
inter-speaker distances for every intra-speaker distance,
the data available for computing the false acceptance rate
was one-eleventh the data available for computing the false
rejection rate. In order to ensure statistical reliability
of the results, the threshold was chosen a posteriori such
that the total error (the sum of both the false acceptance
and false rejection errors) was minimized instead of the
usual equal-error criterion. This appears to weight the
verification test against false acceptance, causing FA to be less than FA. For many practical applications, for example banking, this appears to be an acceptable situation. The results obtained are discussed in the next section.

It is appropriate to mention here that tests were also conducted using the Euclidean distance measure of equation (2.6) and the correlation distance measure of equation (2.15) in the original untransformed (non-orthogonal) feature space. The results obtained were consistently bad and yielded an identification error rate in excess of 30%. It became evident that meaningful results could be achieved only when the parameter sets were orthogonalized prior to calculation of distances.

3.7 RESULTS AND DISCUSSION

The results of the speaker recognition experiments are presented in Tables 3.1 to 3.4. Table 3.1 and Table 3.2 deal with the identification test results for the time domain and frequency domain parameter sets respectively. The corresponding verification scores are presented in Table 3.3 and Table 3.4 respectively.

It is appropriate to mention here that preliminary tests in which reference and test data were drawn from the same recording session without overlap yielded extremely high scores of better than 99% for both identification and verification. But results of such tests are misleading and
<table>
<thead>
<tr>
<th></th>
<th>NO. OF TEST DATA FRAMES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100 (2 Sec.)</td>
</tr>
<tr>
<td>LPC</td>
<td>20.42%</td>
</tr>
<tr>
<td>CEPSTRUM</td>
<td>17.92%</td>
</tr>
<tr>
<td>REFLECTION</td>
<td>21.46%</td>
</tr>
<tr>
<td>AREA</td>
<td>22.50%</td>
</tr>
</tbody>
</table>

Table 3.1. Identification scores (error %) for various test data segment lengths – Time domain parameters.
<table>
<thead>
<tr>
<th>NO. OF TEST DATA FRAMES</th>
<th>250 (2.5 Sec.)</th>
<th>500 (5 Sec.)</th>
<th>1000 (10 Sec.)</th>
<th>2000 (20 Sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>22.79%</td>
<td>18.49%</td>
<td>18.75%</td>
<td>15.63%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NO. OF TEST DATA FRAMES</th>
<th>100 (2 Sec.)</th>
<th>250 (5 Sec.)</th>
<th>500 (10 Sec.)</th>
<th>1000 (20 Sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPSC</td>
<td>15.10%</td>
<td>11.46%</td>
<td>11.98%</td>
<td>12.50%</td>
</tr>
<tr>
<td>DPFT</td>
<td></td>
<td>10.42%</td>
<td></td>
<td>7.29%</td>
</tr>
</tbody>
</table>

Table 3.2: Identification scores (error %) for various test data segment lengths - frequency domain parameters.
<table>
<thead>
<tr>
<th>NO. OF TEST DATA FRAMES</th>
<th>100 (2 Sec.)</th>
<th>250 (5 Sec.)</th>
<th>500 (10 Sec.)</th>
<th>1000 (20 Sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPC</td>
<td>5.32%</td>
<td>4.04%</td>
<td>4.00%</td>
<td>3.30%</td>
</tr>
<tr>
<td>CEPSTRUM</td>
<td>4.94%</td>
<td>4.12%</td>
<td>3.82%</td>
<td>4.16%</td>
</tr>
<tr>
<td>REFLECTION</td>
<td>5.60%</td>
<td>4.74%</td>
<td>4.16%</td>
<td>4.00%</td>
</tr>
<tr>
<td>AREA</td>
<td>5.68%</td>
<td>4.90%</td>
<td>4.16%</td>
<td>4.00%</td>
</tr>
</tbody>
</table>

Table 3.3 Verification scores (total error %) for various test data segment lengths - Time domain parameters.
<table>
<thead>
<tr>
<th>NO. OF TEST DATA FRAMES</th>
<th>250 (2.5 Sec.)</th>
<th>500 (5 Sec.)</th>
<th>1000 (10 Sec.)</th>
<th>2000 (20 Sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSP</td>
<td>6.52%</td>
<td>5.56%</td>
<td>5.04%</td>
<td>4.86%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NO. OF TEST DATA FRAMES</th>
<th>100 (2 Sec.)</th>
<th>250 (5 Sec.)</th>
<th>500 (10 Sec.)</th>
<th>1000 (20 Sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IFSC</td>
<td>4.58%</td>
<td>3.82%</td>
<td>3.64%</td>
<td>3.82%</td>
</tr>
<tr>
<td>DFFT</td>
<td>-</td>
<td>4.78%</td>
<td>4.42%</td>
<td>4.16%</td>
</tr>
</tbody>
</table>

Table 3.4 Verification scores (total error %) for various test data segment lengths - Frequency domain parameters.
do not reflect the effect of variations in the speakers' voices over an extended period of time.

From Tables 3.1 and 3.2 it can be seen that for all of the parameter sets the trend indicated is a decrease in identification error as the length of the test data segment increases. For example, in the case of the cepstrum parameter set (Table 3.1) the error decreases from 17.92% with a 2 second test data segment to 13.54% with a 20 second test data segment. In the case of the IFSC parameter set (Table 3.2) the corresponding decrease is from 15.1% to 7.29% - the error rate is more than halved. The largest decrease is recorded in the case of the area coefficients with the error rate dropping from 22.5% to 12.5%.

In general it can be concluded that the time domain parameter sets, when considered over the entire range of test data segment lengths, have comparable performance. However for the minimum test data segment length of 2 seconds the CEP seems to perform better than the other time domain parameter sets. It appears that, in general, the frequency domain parameters, with the exception of the SSP, perform better than the time domain parameter sets. The dramatic decrease in the error rate of about 5-6% with the use of the IFSC parameter set over the LPC parameter set is particularly significant. The IFSC is obtained from the LPC through a transformation (Section 3.3.2) which apparently has the effect of reducing overlap between the clusters of
different speakers in the multi-dimensional feature space. The disappointing performance of the SSP parameter set as compared to the other frequency domain parameter sets is to be noted. A possible cause could be that the distribution of the centre frequencies of the bandpass filters of Figure 3.1 was not effective in highlighting inter-speaker differences.

The verification results of Tables 3.3 and 3.4 show a similar trend of declining total error rates with increased test data segment lengths though the reductions are much smaller as compared to the identification case. A typical example is the verification result with the LPC parameter set, which decreases from 5.32% for a 2 second test data segment length to 3.3% for a 20 second test data segment length. When considered against the backdrop of a system performing with an accuracy in the range of 94% to 96% this reduction is perhaps as significant as in the case of the identification task. The SSP parameter set performs poorly for short test data segment lengths but the performance improves with increased lengths. The verification task does not seem to be as problematic as the identification task and both the time and frequency domain parameter sets exhibit comparable performance.
Chapter IV
FEATURE SELECTION

4.1 INTRODUCTION

The results of the previous chapter have revealed the potential for speaker recognition of the various speech parameter sets. The identification scores ranged from about 77% to 93% whereas the verification total error rate ranged from 3.2% to 6.6%. In examining these results it became evident that the scores obtained, particularly in the case of the identification task, were too low to admit of practical application. It was then decided to investigate the possibility of improving these scores through feature selection. It has been reported in the past that some subset of the speech parameters have higher speaker selectivity as compared to others. In this connection the work of Sambur[15] and Cheung and Eisenstein[33] are worthy of mention. In Sambur's work, principal component analysis was used to transform parameters obtained through linear prediction analysis to orthogonal space, where they were uncorrelated with each other. It was hypothesized that those uncorrelated parameters having the smallest variances were indicative of the speaker's identity. Recognition accuracies of the order of 99% were achieved by using a
subset of the six (out of twelve) least-variant parameters in
the computation of distance. The author's recognition
experiments performed with different random subsets of the
orthogonal parameters used in evaluating distances yielded
some interesting results. In several instances a feature
subset composed of the six (out of twelve) least-variant
parameters did not perform as well as a subset made up of
parameters 4 through 9 which generally included most of the
high dispersion parameters. The performance of the latter
was at times better than the performance obtained with all
twelve parameters included in computing the distance
measure. The need for a thorough and methodical
investigation of the speaker selectivity of the orthogonal
parameters to arrive at the best feature subset for each
speaker became evident.

Cheung and Eisenstein[33] used the technique of dynamic
programming together with a criterion function called
divergence to obtain feature selection in a text-independent
speaker recognition experiment. The procedure adopted here
in this thesis is a modification of the work of Cheung and
Eisenstein to realize an algorithm for feature selection in
the orthogonal space.

The technique of dynamic programming as adapted to the
problem of feature selection in the orthogonal domain for
text-independent speaker recognition is discussed in Section
4.2. The speaker recognition experiments performed are
described in Section 4.3. The results of the study and a
discussion of the results are presented in Section 4.4.

4.2 FEATURE SELECTION THROUGH DYNAMIC PROGRAMMING

The problem of feature selection can be defined as follows. From each speaker’s training speech data base a
set of N acoustic features characterizing the speaker have
been extracted. Not all of these features have the same
speaker discriminating potential. The task at hand is to
select the best n-feature subset from the N-feature
characterization and to compare their performance in speaker
recognition experiments.

The technique of dynamic programming can be used to
select the "best" n-feature subset. Dynamic
programming [34], [35] is a multistage optimization technique
which allows the decomposition of an n-stage problem into n
single-stage subproblems. It is based on the Principle of
Optimality which states that whatever the initial state and
decisions are, the remaining decisions must constitute an
optimal policy with respect to the state resulting from the
first decision.

The steps involved in the dynamic programming feature
selection procedure are as follows: Let the set of N
available features be represented by $\Phi = (\phi_1, \phi_2, \ldots, \phi_N)$ and
let $S_n^j = (s_1^j, s_2^j, \ldots, s_n^j), j = 1, 2, \ldots, N$ be one of the N
possible subsets chosen after n stages, where
$s_k^j, k=1,2,\ldots,n$ represents a feature in $\phi$. For every $\phi_j, j=1,2,\ldots,N$ at the $n^{th}$ stage, the subset $s_n^j$ is selected such that

$$\text{FEC}_n(s_n^j) = \text{MAX OF FEC}_n(s_{n-1}^i, \phi_j), i=1,2,\ldots,N$$

where the symbol $\not\in$ is used to indicate 'not belonging to' and $(s_{n-1}^i, \phi_j)$ is a subset formed by augmenting $s_{n-1}^i$ with $\phi_j$ as shown below:

$$(s_{n-1}^i, \phi_j) = (s_1^i, s_2^i, \ldots, s_{n-1}^i, \phi_j)$$

FEC is a chosen feature effectiveness criterion.

The optimum subset $s_n^k$ at the end of $n$ stages is obtained as shown below:

$$\text{FEC}_n(s_n^k) = \text{MAX OF FEC}_n(s_n^i), j=1,2,\ldots,N$$

Whether the subset chosen at the end of the procedure listed above is optimal depends on whether it is possible to break up the $n$-stage problem into $n$ subproblems. It has been shown[35] that if the FEC satisfies the conditions listed below this is guaranteed.
i) The FEC is monotonic with respect to \( n \), the stage number.

\[
\text{FEC}_n(S^j_n) > \text{FEC}_{n-1}(S^j_{n-1}) \quad \text{for any } S^j_n.
\] (4.4)

ii) The FEC is a separable function of the history of the process till the \((n-1)\)st stage and the behavior of the process at the \(n\)th stage.

\[
\text{FEC}_n(S^j_n) = f[\text{FEC}_n(S^j_n), \text{FEC}_{n-1}(S^j_{n-1})]
\] (4.5)

If the chosen FEC does not satisfy both of the above conditions the chosen subset is not necessarily the optimum. The FEC chosen by Cheung and Eisenstein in their work was a measure called divergence which, while satisfying the condition of monotonicity, was not separable. It is important to point out here that Cheung and Eisenstein used the technique of dynamic programming to obtain feature selection in the original feature space. The FEC chosen by them was not separable because of the correlation between the original measurement parameters. In this work the possibility of feature selection in the orthogonal space, where the features are uncorrelated with each other, is investigated.
The choice of a suitable FEC can be based on two different approaches. One approach is to use the actual error performance of the features in speaker recognition experiments as a criterion function and to select a feature subset that minimizes the error. The second approach is to work with the probability density functions characterizing the collection of feature vectors in each speaker's training set. The philosophy and justification for these approaches together with the attendant problems is dealt with in detail by Marill and Green[36] and by Kullback[37]. The feature effectiveness criteria used in this work are discussed in the next section.

4.3 CHOICE OF FEC AND SPEAKER RECOGNITION EXPERIMENTS

Two different feature effectiveness criteria were considered in this investigation. They are defined below:

i) Approach I: The FEC was formulated as the difference between the sum of all the inter-speaker distances and the sum of all the intra-speaker distances.

\[
FEC = \sum \text{Inter-speaker distances} - \sum \text{Intra-speaker distances} \tag{4.6}
\]
This is a logical measure to use in assessing feature effectiveness since one would like to select a feature subset that maximizes inter-speaker distance and minimizes intra-speaker distance. It should be noted that the criterion function represented by equation (4.6) is not monotonic but satisfies the condition of separability.

It is evident that depending upon how the feature effectiveness criterion of equation (4.6) is evaluated, it is possible to select one subset of features for the entire population as a whole or to obtain a unique speaker-dependent set for each speaker in the population. In this work the latter approach, with its obvious advantages, was adopted.

The inter and intra-speaker distances required in equation (4.5) were evaluated for the feature subset under consideration with the summation being performed over all the distances obtained from using data from each one of the eight recording sessions in turn as trial data. The details are explained below.

Let \[ \{(\tilde{\phi})_{s}: i=1,2,...,M; \ell=1,2,...,8; s=1,2,...,12\} \] represent the \(i\)th reference orthogonal parameter obtained from all but the \(\ell\)th session of data of the \(s\)th speaker in the
population. Let \( (\lambda_{il})_s \) represent the corresponding eigenvalue. Let \([(\phi_{ikl})_q q=1,2,\ldots, M; i=1,2,\ldots, 4; q=1,2,\ldots, 12)] \) represent the \( i \)th test orthogonal parameter of the \( k \)th test data block from the \( q \)th session of the \( q \)th test speaker.

The feature effectiveness criterion for the \( s \)th speaker, \( \text{FEC}(s) \), is then given by:

\[
\text{FEC}(s) = \frac{1}{11} \sum_{q=1}^{12} \sum_{q \neq s} \sum_{i=1}^{8} \sum_{k=1}^{4} \frac{[(\phi_{ikl})_s - (\phi_{ikl})_q]^2}{(\lambda_{ikl})_s} \tag{4.7}
\]

(INTER)

\[
- \sum_{q=1}^{12} \sum_{i \in S} \sum_{k=1}^{4} \frac{[(\phi_{ikl})_s - (\phi_{ikl})_q]^2}{(\lambda_{ikl})_s} \tag{INTRA}
\]

The symbol \( \in \) is used to indicate 'belonging to' and \( S \) is the subset of features selected at any given stage through the dynamic programming algorithm. The first term in the above expression for the FEC represents the sum of the inter-speaker distances while the second term represents the sum of the intra-speaker distances. Since there were
12 speakers in the population, corresponding to each intra-speaker distance contribution there were 11 inter-speaker distance contributions. The factor of 11 in the first term corresponding to inter-speaker distances was inserted in order to give equal weightage to the inter and intra-speaker distance terms. The upper limit of 4 for the summation index was a consequence of the fact that the length of the test data block was fixed at 250 frames (5 seconds). Each session of speech yielded \(1000/250=4\) separate blocks of test data.

The feature selection was performed individually for each speaker in the population and the various \(n\)-feature subsets selected for each speaker were used in speaker recognition experiments as described in Section 3.6. The results obtained are discussed in the next section.

ii) Approach II: The \(\text{FEC} \) was defined as given below.

\[
\text{FEC}(s) = \sum_{q \in S} \sum_{i \in S} \frac{[(\phi_i)_s - (\phi_q)_q]^2}{(\lambda_{i_s})}
\]  

(4.8)

In the above expression \((\phi_i)_s\) is the \(i^{th}\) orthogonal parameter of the \(s^{th}\) reference speaker computed over all the 8000 frames of his speech data base. \((\lambda_{i_s})\) are the corresponding
eigenvalues. \((\phi_i)\) represents the \(i\)th orthogonal parameter of the \(q\)th test speaker computed using the reference eigenvectors of the \(s\)th reference speaker in the same fashion as represented by equation (3.9).

The criterion function of equation (4.8) is both monotonic and separable. Once again a separate speaker-dependent subset was selected for each speaker 's' in the population.

It can be seen that while the FBC of Approach I is related to the error performance approach, the FBC of Approach II is based on the probabilistic approach. The results obtained are discussed in the next section.

4.4 RESULTS AND DISCUSSION

The results of the feature selection study are presented in Tables 4.1 to 4.6. The "best" 6-parameter subset (out of 12) selected for each speaker through the dynamic programming algorithm using the FBC of Approach I is shown in Table 4.1 for the orthogonal cepstrum parameter set. The asterisks (*) indicate additional parameters to be included in order to obtain a 8-parameter subset.

Examination of these results reveals the possible existence of a pattern in the feature selectivity of the orthogonal cepstrum parameters. The following comments are in order.
parameters it appears as though parameters 1, 2, 11 and possibly 5 are generally not selected for the different speakers in the population. If the 8-parameter subset is considered (include parameters with *) a good rule of thumb seems to be to discard parameters 1 and 2.

The "best" 16-parameter subset (out of 32) selected for the orthogonal IFSC parameter set is indicated in Table 4.2. The asterisks indicate additional parameters to be included in order to obtain a 20-parameter subset. In this case it appears as though parameters 1, 2, 18, 19, 21, 22, 24, 25 are generally not selected for the various speakers in the population. The feature subsets selected for some of the other parameter sets are presented in Appendix II.

The identification scores that were obtained using various n-parameter subsets selected using Approach I are presented in 4.3 and 4.4 for the time domain and frequency domain parameter sets respectively. The corresponding verification scores are shown in Tables 4.5 and 4.6.

For most of the time domain parameter sets (Table 4.3), with the exception of the LPC, there is an initial decrease in the error rate when n-parameter subsets (n<12) as selected by the dynamic programming algorithm of Approach I are used in the computation of the distance measure. Thus, for example, in the case of the cepstrum parameter set, the error rate decreases from 15.89% (all 12 parameters) to a minimum of 12.50% (with a 10-parameter subset). In the case
<table>
<thead>
<tr>
<th>SPEAKER INITIALS</th>
<th>PARAMETER NO.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1  2  3  4  5  6  7  8  9  10 11 12</td>
</tr>
<tr>
<td>AT</td>
<td>X  X  X  *  X  X  *  X</td>
</tr>
<tr>
<td>BN</td>
<td>X  *  X  X  *  X  X  X</td>
</tr>
<tr>
<td>CC</td>
<td>*  X  X  X  X  X  X</td>
</tr>
<tr>
<td>CS</td>
<td>X  *  X  X  X  X  X</td>
</tr>
<tr>
<td>DV</td>
<td>X  X  X  X  *  X  *</td>
</tr>
<tr>
<td>JD</td>
<td>X  X  X  X  *  X  *</td>
</tr>
<tr>
<td>JZ</td>
<td>X  X  *  X  X  *  X</td>
</tr>
<tr>
<td>KD</td>
<td>X  X  *  X  X  X  X</td>
</tr>
<tr>
<td>KT</td>
<td>X  X  *  X  X  X  *</td>
</tr>
<tr>
<td>MF</td>
<td>X  *  X  X  X  X  *</td>
</tr>
<tr>
<td>RW</td>
<td>X  X  X  X  *  *  X</td>
</tr>
<tr>
<td>TO</td>
<td>X  X  *  X  *  X  X</td>
</tr>
</tbody>
</table>

Table 4.1 Feature subsets selected through dynamic programming using FEC I - Cepstrum parameters (6 out of 12). The asterisks (*) indicate parameters to be added for an 8-parameter subset.
<table>
<thead>
<tr>
<th>SPEAKER INITIALS</th>
<th>PARAMETER NO.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT</td>
<td>X X X X X X X X X X * * X X * X * X X X X X</td>
</tr>
<tr>
<td>BN</td>
<td>X X * X X X X X X X X * X X * X X X X X</td>
</tr>
<tr>
<td>CC</td>
<td>X X X X X X X X * X X X X X X X X X X X</td>
</tr>
<tr>
<td>CS</td>
<td>X X X X X X X X * X X X X X X X X X X X</td>
</tr>
<tr>
<td>DV</td>
<td>X X X X X X X X X X X X X X * X X X X X</td>
</tr>
<tr>
<td>JD</td>
<td>X X X X X X X X * X X X X X X X X X X X</td>
</tr>
<tr>
<td>JZ</td>
<td>X X X X X X X X X X X * X X X X X X X X</td>
</tr>
<tr>
<td>KD</td>
<td>* X X X X X X X X X X X X X X X X X X X</td>
</tr>
<tr>
<td>KT</td>
<td>X X * X X X X X X X X X X X X X X X X X</td>
</tr>
<tr>
<td>MF</td>
<td>X X * X X X X X X X X X X X X X X X X</td>
</tr>
<tr>
<td>RW</td>
<td>X X X X X X X * X X X X X X X X X X X</td>
</tr>
<tr>
<td>TO</td>
<td>X X X X X X X X X X X X X X X X X X X</td>
</tr>
</tbody>
</table>

Table 4.2. Feature subsets selected through dynamic programming using FEC I - IFSC parameters (16 out of 32). The asterisks (*) indicate parameters to be added for a 20-parameter subset.
<table>
<thead>
<tr>
<th>NO. OF SELECTED PARAMETERS</th>
<th>12</th>
<th>10</th>
<th>8</th>
<th>6</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPC</td>
<td>16.41%</td>
<td>18.23%</td>
<td>17.45%</td>
<td>21.35%</td>
<td>27.34%</td>
</tr>
<tr>
<td>REFLECTION</td>
<td>17.19%</td>
<td>13.54%</td>
<td>14.58%</td>
<td>24.74%</td>
<td>29.17%</td>
</tr>
<tr>
<td>CEPSTRUM</td>
<td>15.89%</td>
<td>12.50%</td>
<td>16.41%</td>
<td>19.53%</td>
<td>27.34%</td>
</tr>
<tr>
<td>AREA</td>
<td>18.23%</td>
<td>17.19%</td>
<td>15.36%</td>
<td>20.83%</td>
<td>27.34%</td>
</tr>
</tbody>
</table>

Table 4.3. Identification scores (error %) based on feature selection obtained using FEC I - Time domain parameters.
<table>
<thead>
<tr>
<th>NO. OF SELECTED PARAMETERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 14 12 10 8 6 4</td>
</tr>
<tr>
<td>18.49% 19.01% 21.09% 21.88% 28.13% 37.76% 51.82%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NO. OF SELECTED PARAMETERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>32 28 24 20 16 12 8 4</td>
</tr>
<tr>
<td>11.46% 10.42% 10.16% 9.38% 9.90% 16.15% 23.18% 41.93%</td>
</tr>
<tr>
<td>11.98% 8.59% 10.68% 9.38% 11.72% 13.28% 21.35% 51.30%</td>
</tr>
</tbody>
</table>

Table 4.4. Identification scores (error %) based on feature selection obtained using FEC I - Frequency domain parameters.
<table>
<thead>
<tr>
<th>Feature</th>
<th>12</th>
<th>10</th>
<th>8</th>
<th>6</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPC</td>
<td>4.04%</td>
<td>4.26%</td>
<td>4.00%</td>
<td>5.08%</td>
<td>5.52%</td>
</tr>
<tr>
<td>Reflection</td>
<td>4.74%</td>
<td>4.34%</td>
<td>4.22%</td>
<td>5.42%</td>
<td>6.16%</td>
</tr>
<tr>
<td>Cepstrum</td>
<td>4.12%</td>
<td>3.86%</td>
<td>4.26%</td>
<td>4.68%</td>
<td>6.30%</td>
</tr>
<tr>
<td>Area</td>
<td>4.90%</td>
<td>4.86%</td>
<td>4.74%</td>
<td>5.34%</td>
<td>5.82%</td>
</tr>
</tbody>
</table>

Table 4.5 Verification scores (total error %) based on feature selection obtained using FEC I - Time domain parameters.
<table>
<thead>
<tr>
<th>NO. OF SELECTED PARAMETERS</th>
<th>16</th>
<th>14</th>
<th>12</th>
<th>10</th>
<th>8</th>
<th>6</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSP</td>
<td>5.56%</td>
<td>5.68%</td>
<td>5.78%</td>
<td>5.86%</td>
<td>6.52%</td>
<td>7.68%</td>
<td>10.6%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NO. OF SELECTED PARAMETERS</th>
<th>32</th>
<th>28</th>
<th>24</th>
<th>20</th>
<th>16</th>
<th>12</th>
<th>8</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>IFSC</td>
<td>3.82%</td>
<td>3.86%</td>
<td>3.68%</td>
<td>3.86%</td>
<td>4.26%</td>
<td>4.64%</td>
<td>5.64%</td>
<td>8.08%</td>
</tr>
<tr>
<td>DFFT</td>
<td>4.78%</td>
<td>4.04%</td>
<td>4.00%</td>
<td>3.94%</td>
<td>4.12%</td>
<td>4.48%</td>
<td>5.26%</td>
<td>9.02%</td>
</tr>
</tbody>
</table>

Table 4.6 Verification scores (total error %) based on feature selection obtained using FEC I - Frequency domain parameters
of the log area ratio coefficients the decrease is from 18.23% (all 12 parameters) to a minimum of 15.34% (with a 8-parameter subset). Similar results are obtained in the case of the frequency domain parameter sets (Table 4.4) with the exception of the SSP. Thus, for instance, in the case of the WFFT the error rate decreases by over 3% from 11.98% (all 32 parameters) to 8.59% (with a 2d-parameter subset).

The verification results presented in Table 4.5 and Table 4.6 reveal a similar trend with some exceptions. The improvement in accuracy is smaller as compared to the identification task.

The "best" 6-parameter subset selected for each speaker through the dynamic programming algorithm using the FEC of Approach II is presented in Table 4.7 for the LPC parameter set. The identification and verification scores obtained using the selected subsets in the computation of the distance measure are shown in Table 4.8 for the LPC parameter alone. The results obtained with the FEC of Approach II are markedly different from those obtained with the FEC of Approach I. In this case, for the selected n-parameter subsets, the error rate monotonically increases as n decreases. This behaviour was prevalent in all the parameter sets investigated, leading to the conclusion that the features selected using the FEC of Approach II are unsatisfactory as far as the recognition task is concerned.
No advantage is gained by using a subset of the features as selected through Approach II instead of all the features, since the error rate monotonically increases as \( n \) decreases. Similar behaviour was exhibited by the other parameter sets and hence those results are not being presented.
<table>
<thead>
<tr>
<th>SPEAKER INITIALS</th>
<th>PARAMETER NO.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>AT</td>
<td>x</td>
</tr>
<tr>
<td>BN</td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td></td>
</tr>
<tr>
<td>CS</td>
<td></td>
</tr>
<tr>
<td>DV</td>
<td></td>
</tr>
<tr>
<td>JD</td>
<td>x</td>
</tr>
<tr>
<td>JZ</td>
<td></td>
</tr>
<tr>
<td>KD</td>
<td></td>
</tr>
<tr>
<td>KT</td>
<td></td>
</tr>
<tr>
<td>MF</td>
<td></td>
</tr>
<tr>
<td>RW</td>
<td></td>
</tr>
<tr>
<td>TO</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.7 Feature subsets selected through dynamic programming using FEC II - LPC parameters (6 out of 12)
<table>
<thead>
<tr>
<th>NO. OF SELECTED PARAMETERS</th>
<th>12</th>
<th>10</th>
<th>8</th>
<th>6</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPC</td>
<td>16.4%</td>
<td>18.75%</td>
<td>19.49%</td>
<td>24.22%</td>
<td>28.91%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NO. OF SELECTED PARAMETERS</th>
<th>12</th>
<th>10</th>
<th>8</th>
<th>6</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPC</td>
<td>4.04%</td>
<td>4.12%</td>
<td>4.60%</td>
<td>5.08%</td>
<td>5.72%</td>
</tr>
</tbody>
</table>

Table 4.8 Identification and verification scores (error % and total error %) based on feature selection obtained using FEC II - LPC parameters
Chapter V
NEAREST NEIGHBOURHOOD TECHNIQUE

5.1 INTRODUCTION

The technique of speaker recognition adopted in the preceding two chapters has been based on the characterization of each reference speaker in the population by the mean of the feature vectors in his training set. In Section 2.4 another classification rule known as the nearest neighbour method was discussed in which each reference speaker was described by a cluster of points, each representing a feature vector obtained from his training speech data base. This chapter investigates the effect on recognition scores of a decision algorithm in which a multiple feature vector characterization for each reference speaker in the population was used. In addition it discusses the incorporation of a decision strategy that pools decisions from a block of several successive test data utterance segments to make an identification decision for the block as a whole.

Section 5.2 discusses the implementation of the proposed nearest neighbour method. The conducted speaker identification experiments are described in Section 5.3. The results obtained and the conclusions arrived at are dealt with in Section 5.4.
5.2 Modified Nearest Neighbour Method

The speaker recognition technique to be discussed in this section is a variation of the nearest neighbour method of equation (2.16). The training set consisted of a cluster of NR reference sets for each of the NSP speakers. The KNNearest neighbours of the parameter set representing the test utterance segment of the unknown speaker were obtained from amongst the (NR x NSP) available reference feature sets. The reference speaker with the highest 'votes' amongst the KNNearest neighbours was identified as the speaker of the unknown utterance. If KNN is chosen to be NR, in the ideal case we would expect all of the NR nearest neighbours to originate from the cluster of the correct speaker.

In an extension of the above method test utterance segments were grouped in a block of 3, 5, 10 or 20 segments and a majority decision was obtained from the individual decisions resulting from each test utterance segment to identify the unknown speaker. It should be noted that ties may result from use of this algorithm. This may, for example, occur with a block size of 5 test utterance segments when the distribution of the 5 separate decisions is 2, 2, and 1 in favour of three of the speakers in the population. Such ties can, however, be broken by ruling in favour of that speaker of the two tied speakers who has the minimum average distance from amongst the 5 x KNN distances used to make individual decisions for the 5 utterances in
the block. The details of the speaker recognition experiments performed to test the nearest neighbour technique are described in the next section.

5.3 **Speaker Recognition Experiments**

Once again it was found that the use of a Euclidean distance measure in the original feature space led to extremely poor results in the implementation of the nearest neighbour method[38]. Orthogonalization of the parameter sets was essential prior to derivation of the KNN nearest neighbours.

Each of the Na reference sets for a particular speaker was the average of the orthogonal parameters of that speaker over 100 frames (2 seconds) of data from his training speech data base. The eigenvector analysis was performed using 7000 frames (from 7 sessions) for each speaker in a manner identical to that described in Section 3.6. As a consequence the value of NR was 70 (=7000/100). The test utterance segment was made up of 100 frames extracted from the eighth session not used in creating the reference. Each of the 8 sessions of each speaker was in turn used to obtain the test data. As a consequence a total of 80 (=8000/100) identification tests were conducted for each speaker.

An initial experiment was conducted in order to decide on the optimum value of KNN to be used in the decision logic for identification. This was done by comparing the
identification scores obtained corresponding to different values of \( KNN \). The results obtained are presented in the next section.

5.4 RESULTS AND DISCUSSION

The variation of the identification error rate with the no. of nearest neighbours (\( KNN \)) for the cepstrum parameter set is shown in Table 5.1. The minimum error occurs for \( KNN \) in the range 20 to 35. Similar behaviour was exhibited in the case of the other parameter sets. On the basis of these results it was decided to use a value of \( KNN \) equal to 35 in the decision logic of the nearest neighbour scheme.

The results of the identification tests using the nearest neighbour classification rule are presented in Table 5.2 for the time domain parameters and in Table 5.3 for the frequency domain parameters. The results obtained earlier using the Mahalanobis distance metric and presented in Section 3.7 are also listed to enable comparisons to be made. It can be seen that in the absence of blocking (\( BS=1 \)) the results obtained using the nearest neighbour rule are comparable to the results obtained using the Mahalanobis distance metric. These are, for the cepstrum parameter set, 17.19% and 17.92% respectively. On the other hand, the incorporation of blocking results in a significant improvement in performance over the MDK. For example, in the case of the cepstrum parameter set, the identification
<table>
<thead>
<tr>
<th>KNN</th>
<th>ERROR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>21.35%</td>
</tr>
<tr>
<td>5</td>
<td>17.50%</td>
</tr>
<tr>
<td>10</td>
<td>17.50%</td>
</tr>
<tr>
<td>20</td>
<td>17.19%</td>
</tr>
<tr>
<td>30</td>
<td>17.19%</td>
</tr>
<tr>
<td>35</td>
<td>17.19%</td>
</tr>
<tr>
<td>40</td>
<td>18.02%</td>
</tr>
<tr>
<td>60</td>
<td>18.44%</td>
</tr>
<tr>
<td>70</td>
<td>19.58%</td>
</tr>
<tr>
<td>80</td>
<td>19.79%</td>
</tr>
</tbody>
</table>

Table 5.1 Variation of identification scores (error %) with No. of nearest neighbours (KNN) - Cepstrum parameters
error decreases from 17.19% for BS=1 to 13.54% for BS=10. Similar behaviour is observed in the case of the other parameter sets. However it is only fair to compare these results with those obtained using the MDM with increased test data segment lengths (Table 3.1). On this basis, keeping in mind the greater computational burden of the nearest neighbour method, it appears as though the MDM is to be preferred over the nearest neighbour classification rule.

The results of the implementation of the nearest neighbour classification rule discussed above appear to differ from those obtained by other workers[27]. It should be noted, however, that the implementation undertaken in this work and the implementation of Schwartz et al[27] represent two different variations of the nearest neighbour method and hence any direct comparison of results is not valid.
<table>
<thead>
<tr>
<th>BLOCK SIZE</th>
<th>MDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20.42%</td>
</tr>
<tr>
<td>3</td>
<td>20.42%</td>
</tr>
<tr>
<td>5</td>
<td>17.92%</td>
</tr>
<tr>
<td>10</td>
<td>17.92%</td>
</tr>
<tr>
<td>20</td>
<td>17.92%</td>
</tr>
<tr>
<td>LPC</td>
<td>19.79%</td>
</tr>
<tr>
<td>CEPSTRUM</td>
<td>17.19%</td>
</tr>
<tr>
<td>Reflection</td>
<td>22.81%</td>
</tr>
<tr>
<td>Area</td>
<td>23.02%</td>
</tr>
</tbody>
</table>

Table 5.2. Variation of identification scores (error %) with block size. Time domain parameters.
<table>
<thead>
<tr>
<th>BLOCK SIZE</th>
<th>SSP</th>
<th>MDM</th>
<th>IFSC</th>
<th>MDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22.66%</td>
<td>19.05%</td>
<td>15.31%</td>
<td>15.10%</td>
</tr>
<tr>
<td>3</td>
<td>14.54%</td>
<td>16.67%</td>
<td>10.68%</td>
<td>10.54%</td>
</tr>
<tr>
<td>4</td>
<td>8.33%</td>
<td>9.09%</td>
<td>6.25%</td>
<td>6.25%</td>
</tr>
<tr>
<td>8</td>
<td>6.25%</td>
<td>16.67%</td>
<td>10.68%</td>
<td>10.54%</td>
</tr>
<tr>
<td>16</td>
<td>6.25%</td>
<td>9.09%</td>
<td>8.33%</td>
<td>8.33%</td>
</tr>
</tbody>
</table>

Table 5.3: Variation of identification scores (error %) with block size.

Frequency domain parameters.
Chapter VI
A COMPOSITE SCHEME FOR SPEAKER RECOGNITION

6.1 INTRODUCTION

The possibility of feature selection of the orthogonal speech parameters as a means of improving recognition scores was discussed in Chapter IV. From the results presented there it could be seen that for most of the speech parameter sets considered in the study, significant improvements in accuracy result as compared to using the entire feature set, particularly in the case of the identification task. In Chapter V a variation of the nearest neighbour method was investigated. The incorporation of a decision scheme, that combined decisions from several successive test data utterance segments or a block to make an identification decision for the entire block, was shown to produce improvements in accuracy.

In the past, several techniques have been proposed for improving recognition scores by incorporating a sequential or deferred decision strategy into the recognition scheme. In such schemes an additional test utterance is called for when the recognition algorithm does not yield an unambiguous decision. In the case of speaker identification, such a situation would arise when the identity of the test speaker
cannot be established with confidence. In the case of verification, any uncertainty in the distance measure would cause the algorithm to reject the identity claim of a genuine speaker or accept the identity claim of a test speaker who is an imposter. The work of Doudington described in the paper by Rosenberg[11] and discussed in Section 2.7 of this thesis is a case in point. The speech data used by Doudington consisted of four-word phrases. The decision function resulting from successive phrases up to a maximum of four were combined and their effect on verification rate studied. The equal-error rate dropped from 4\% for just one phrase to 0.5\% with the inclusion of all four phrases. A strategy of using as many phrases as necessary to establish a certain degree of confidence was adopted. A false rejection rate of 0.3\% was obtained for a chosen false acceptance rate of 1\%. An average of 1.3 phrases was found to be required for making a verification decision at the chosen confidence level.

Close examination of the results of recognition experiments conducted by the author (the subject of Chapter 3) led to an interesting observation. In many instances the decisions obtained from recognition schemes based on two different speech parameter sets were not identical in the case of test utterances that were incorrectly identified or verified. It was then decided to investigate the effect of logically combining decisions from two independent tests involving two different parameter sets.
Section 6.2 discusses the steps in the proposed composite recognition scheme. Section 6.3 lists the various combinations of speech parameter sets chosen for evaluation of the composite scheme and the recognition experiments performed. The results obtained and a discussion of the results are presented in Section 6.4.

6.2 **COMPOSITE RECOGNITION SCHEME**

The steps in the proposed composite recognition scheme involving logical combination of decisions obtained from two different parameter sets are as follows:

a) If the decisions from the two independent tests are identical, then accept this decision.

b) If the two decisions are different, repeat the recognition test using a different test utterance. If the two decisions resulting from the repeated test are now identical, accept this decision.

c) If the two decisions are different even after two such repeats, reject the identity claim of the test speaker in the case of a verification test. In the case of an identification test classify the decision as "unknown identity".

It is evident that the above verification scheme is weighted against false acceptance and in favour of false rejection. Figure 6.1 represents a flow chart of the composite scheme algorithm discussed above.
Fig. 6.1. The composite scheme algorithm.
6.3 Speaker Recognition Experiments

The following combinations of speech parameter sets were chosen for implementation and evaluation of the proposed composite scheme:

1) The 32-parameter IFSC with the 32-parameter DFFT.
2) The 12-parameter REP with the 16-parameter SSE.
3) The 12-parameter LPC with the 12-parameter CFF.
4) The best 20-parameter IFSC subset with the best 20-parameter DFFT subset. The subsets were obtained through the process of feature selection described in Chapter IV.

The above choice was governed by a criteria of combining parameter sets with comparable recognition accuracies to form the composite scheme. The speaker recognition experiments conducted were identical to those described in Section 3.6 with the following exception. Only a test utterance of length 250 frames (5 seconds) was considered in the implementation of the composite scheme.

6.4 Results and Discussion

The identification results for the four combinations of feature sets considered in the implementation of the composite scheme are presented in Table 6.1. The results obtained when using each parameter set individually have also been reproduced to enable comparisons to be made. The
corresponding results of verification tests are presented in Table 6-2.

It can be seen from Table 6-1 that for all combinations of parameter sets considered there is a significant increase in identification accuracy as compared to the results obtained by basing the decision on each parameter set alone. For example, in the case of the 32/32 IFSC with the 32/32 DFPT, a success rate of 95.11% is obtained as compared to 88.54% and 88.02% with the individual parameter sets. The failure rate drops from 11.46% (with 32/32 IFSC) and 11.98% (with 32/32 DFPT) to 3.00% with a composite scheme combining the decisions from these two parameter sets. The "UNIDENTIFIED" category (UI) accounted for the remaining 1.83%. About 8% of all the identification tests conducted required either one or two repeats before the identity could be established.

The combination of the 12/12 LPC with the 12/12 CEP, while also resulting in significant accuracy improvements, is not entirely satisfactory since the overall failure rate of 11.05% is still short of acceptable levels for acceptance as a viable practical implementation scheme.

The results obtained with the combination of the 12/12 REP with the 16/16 SSP are interesting. From the results of Section 3.7 and Section 4.4 the SSP parameters, as derived in this work, were found to be rather unsuitable for the task of identification, in view of the large error rates
<table>
<thead>
<tr>
<th>PARAMETER SET COMBINATION</th>
<th>COMPOSITE SCHEME</th>
<th>INDIVIDUAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SUCCESS</td>
<td>FAILURE</td>
</tr>
<tr>
<td>32/32 IFSC + 32/32 DFFT</td>
<td>95.11%</td>
<td>3.06%</td>
</tr>
<tr>
<td>12/12 REF + 16/16 SSP</td>
<td>91.45%</td>
<td>3.29%</td>
</tr>
<tr>
<td>20/32 IFSC + 20/32 DFFT</td>
<td>95.88%</td>
<td>2.94%</td>
</tr>
<tr>
<td>12/12 LPC + 12/12 CEP</td>
<td>88.08%</td>
<td>11.05%</td>
</tr>
</tbody>
</table>

Table 6.1. Identification results with composite scheme.
<table>
<thead>
<tr>
<th>PARAMETER SET COMBINATION</th>
<th>INDIVIDUAL SUCCESS</th>
<th>INDIVIDUAL ERROR</th>
<th>TOTAL SUCCESS</th>
<th>TOTAL ERROR</th>
<th>REPEAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>32/32 ESC + 32/32 DEFT</td>
<td>97.18%</td>
<td>2.83%</td>
<td>95.22%</td>
<td>4.78%</td>
<td>6</td>
</tr>
<tr>
<td>12/12 REF + 16/16 SSP</td>
<td>97.96%</td>
<td>2.04%</td>
<td>95.26%</td>
<td>4.74%</td>
<td>6</td>
</tr>
<tr>
<td>20/32 ESC + 20/32 DEFT</td>
<td>97.69%</td>
<td>2.31%</td>
<td>96.44%</td>
<td>3.56%</td>
<td>6</td>
</tr>
<tr>
<td>12/12 LPC + 12/12 CEP</td>
<td>96.96%</td>
<td>3.04%</td>
<td>95.96%</td>
<td>4.04%</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 6.2: Verification results with composite scheme.
obtained. However, in combination with the 12/12 REF (whose individual performance was also unsatisfactory), they produce the largest improvement in the success rate of about 9% and result in a failure rate of only 3.29%. The price paid is a large percentage of repeated tests (12.5%) and a large OI category (5.26%).

The verification results of the composite scheme (Table 6.2) show similar but smaller improvements as compared to the identification results. For example, in the case of the 32/32 IFSC with the 32/32 DFPT, the total error rate is reduced to 2.83% from the individually obtained total error rates of 3.82% and 4.78% respectively. The percentage of conducted tests that had to be repeated before a decision could be reached was 1.61%. A low repetition rate was a salient feature of the results for all the combinations of parameter sets investigated. The best results were obtained with the 12/12 REF and the 16/16 SSP which yielded a total error rate of only 2.04%. It is appropriate to mention here that the distance thresholds used in the composite scheme were the minimum total error thresholds obtained from verification experiments conducted with each parameter set individually (Section 3.6). By appropriate modification of these thresholds, the scheme can be even further weighted against false acceptance or weighted against false rejection depending on the requirements of the application.
Chapter VII
DISCUSSION

7.1 SUMMARY

An investigation has been conducted into the viability of text-independent speaker recognition schemes that make use of acoustic characterizations obtained from short-duration reference and test speech segments of each speaker. An initial study was performed to ascertain the relative potential for speaker recognition of various parametric representations of speech, viz., the linear prediction coefficients, cepstrum coefficients, reflection coefficients, log area ratio coefficients, speech spectrum parameters, inverse filter spectral coefficients and the direct FFT parameters. The Mahalanabis minimum distance metric was used to classify the unknown test speaker. The identification accuracies obtained ranged from 81.5% to 88.5% for the various parameter sets for a 5 second test data utterance. The highest accuracy of 88.54% was obtained using the inverse filter spectral coefficients. In general, the accuracies obtained with all the parameter sets tended to increase when the test utterance lengths were increased. Verification accuracies varied between 94.4% and 96.2% with the inverse filter spectral coefficients again coming out on top.

- 100 -
Next, the potential for feature selection among the elements of each parameter set was investigated. A dynamic programming technique was used for this purpose. The feature selection was performed in the orthogonal feature space with the orthogonalization being carried out through principal component analysis. A feature effectiveness criterion based on minimization of the difference between the sum of all inter-speaker distances and the sum of all the intra-speaker distances was shown to yield promising results. Almost all the parameter sets registered improved accuracies in identification tests conducted with the chosen feature subsets. The increase in identification accuracy was significant and was between 2% to 4% for the various parameter sets. The verification results showed similar trends but with smaller percentage improvements in accuracy.

The performance of a classifier making use of a multi-reference characterization of each reference speaker in the population was investigated. This is the well-known nearest neighbour classification rule. The identification results obtained with a 2 second test utterance length were comparable to those obtained earlier with the Mahalanobis distance metric. A decision strategy that pooled decisions from a block of several successive test utterances, to make an overall decision for the block as a whole was formulated. This resulted in significant improvements in identification accuracy.
Finally, a composite scheme for speaker recognition was proposed. This scheme involved the logical combination of the separate decisions from two independent tests performed with two different parameter sets. The use of this decision strategy led to significant improvements in recognition accuracy. For instance, combining the IFSC and the DFFT parameter sets led to an identification success rate of 95.11% as compared to individual accuracies of 88.54% and 88.02% respectively. The failure rate was 3.06% with a new "UNIDENTIFIED" (don't know) category accounting for the remaining 1.83%.

It must be noted here that no mention has been made in this work of computation time considerations, as this was not the major thrust of this work. In general, the computation time involved in speaker recognition tests can be divided into contributions from on-line and off-line computations. Typically the former involves those computations required to obtain a reference characterization for the speakers in the population, while the latter includes those necessary to obtain the test speaker's characterization and the calculation and comparison of distances. For certain applications the latter could also be done off-line. It is felt that, in general, the techniques proposed here do not exhibit any serious computational limitations, keeping in mind the state of the art.
7.2 CONCLUSIONS

The broad conclusions of this work are as follows:

1. The LPC model of speech does appear to offer a very reliable set of parameters for speaker recognition.

2. Spectral information appears to be, in general, more favourable for speaker recognition than time domain information.

3. The orthogonal speech parameter sets obtained through principal component analysis appear to have a potential for feature selection resulting in improved recognition accuracies.

4. The nearest neighbour classification rule, while not encouraging in itself, does seem to show a potential for yielding high accuracy identification with the incorporation of blocking strategies into the decision logic.

5. The proposed composite speaker recognition scheme results in significant improvements in recognition accuracy and seems to be extremely promising as the basis of a high accuracy speaker recognition system.

7.3 CONTRIBUTIONS OF THIS WORK

The contributions of this work are the following:

1. An exhaustive evaluation of the speaker recognition potential of various speech parameter sets with the use of short-duration reference and test speech segments.
2. The formulation of the DFFT parameter set and the investigation of its performance as speaker-characterizing features for text-independent speaker recognition.

3. The investigation of the performance of the IFSC parameter set as speaker-characterizing features for text-independent speaker recognition.

4. The investigation of techniques for improving recognition accuracies through feature selection and the incorporation of decision strategies into the recognition logic.

7.4 SUGGESTIONS FOR FUTURE WORK

1. An interesting observation resulted from the implementation of the proposed composite speaker recognition scheme. In the case of identification tests that resulted in errors, the same incorrect identification was repeated for several consecutive test utterance segments that were used. This is possibly due to the strong correlation between adjacent frames of speech. It is recommended that the recognition tests be repeated with the incorporation of the following provision. The repeat test utterance must be obtained after a period of time of about 2 to 3 minutes has elapsed after the first test utterance. This would not be too
impractical a constraint in many practical implementations of recognition schemes. The results obtained from such recognition tests would be revealing.

2. The underlying assumption of stationarity of the speech process is somewhat debatable. In this connection it would be interesting to investigate the use of multiple reference covariance matrix characterizations, particularly for the non-steady (problematic) speakers in the population.

3. The possible inclusion of other speech characteristics such as zero crossing rate, pitch, accent aspects and nasal features to supplement the parameter sets used in this study should also be considered.

4. The incorporation of heuristic features into the decision strategy for speaker recognition is also worth considering. This could, for instance, take the form of the presence of peaks in certain well-defined regions of the spectrum of the speech of certain speakers.

5. Variations of the spectrally based schemes investigated in this work can be considered. For instance, the speech spectrum parameters can be reformulated with the availability of a more flexible choice of the number of filters in the filter bank.
and their centre frequencies. This could also be
done for the other spectral schemes involving the use
of the FET algorithm, by redefining the mode of
averaging of the spectral coefficients. The above
could also be done in a speaker-dependent manner, if,
in doing so, inter-speaker differences are
highlighted.
APPENDIX I

The modelling of the vocal tract as a nonuniform acoustic tube formed by concatenating $p$ uniform cylindrical sections of equal length[7][10], leads to the set of parameters called the reflection coefficients. The reflection coefficients $\{k_i, i=1, 2, \ldots, p\}$ are related to the linear prediction coefficients $\{a_i, i=1, 2, \ldots, p\}$ by the following relations[7]:

\[
\begin{align*}
  X_i &= a_i^{(i)} \\
  a_j^{(i-1)} &= \frac{a_j^{(i)} + a_i^{(i)} \cdot a_{j-1}^{(i)}}{1 - k_i^2} & (A1.1) \\
  1 \leq j \leq i-1, \ i = p, p-1, \ldots, 1
\end{align*}
\]

The recursion is initiated by,

\[
\begin{align*}
  a_j^{(p)} &= a_j & (A1.2) \\
  1 \leq j \leq p
\end{align*}
\]

The acoustic tube model yields another set of parameters that are related to the ratios of the areas of the uniform cylindrical sections. These are called the log area ratio coefficients. The log area ratio coefficients $\{G_i, i=1, 2, \ldots, p\}$ can be obtained from the reflection coefficients as follows[7]:

- 107 -
\[ G_i = \log \left( \frac{A_i+1}{A_i} \right) = \log \left( \frac{1-K_i}{1+K_i} \right) \quad (A1.3) \]

\[ 1 \leq i \leq P \]

where \( \{A_i, i=1,2,\ldots,P\} \) are the areas of the \( P \) cylindrical sections of the model.

The cepstrum coefficients \( \{c_i, i=1,2,\ldots,P\} \) can be obtained by computing the cepstrum of the inverse filter \( H(z) \) and hence are related to the linear prediction coefficients through the following relations[7]:

\[ c_i = a_i + \sum_{j=1}^{i-1} \frac{i}{j} c_j a_{i-j} \quad (A1.4) \]

\[ i=1,2,\ldots,P \]
APPENDIX II

The m-feature subsets selected through the dynamic programming algorithm using FFC I are presented below for the LPC, REP, LAR and DPFT parameter sets.
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT</td>
<td>x</td>
<td>x</td>
<td>*</td>
<td>*</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BN</td>
<td>*</td>
<td>*</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>*</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS</td>
<td>x</td>
<td>x</td>
<td>*</td>
<td>x</td>
<td>x</td>
<td>*</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DV</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>*</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JD</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>*</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JZ</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>*</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KD</td>
<td>*</td>
<td>x</td>
<td>x</td>
<td>*</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KT</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>*</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MF</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>*</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RW</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>*</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TO</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>*</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A.2.1. Feature subsets selected through dynamic programming using 
FEC I - Area coefficients (6 out of 12). The asterisks (*) 
indicate parameters to be added for an 8-parameter subset.
<table>
<thead>
<tr>
<th>Parameter No.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>*</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BN</td>
<td></td>
<td>*</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DV</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>*</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JD</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>JZ</td>
<td>X</td>
<td>*</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KD</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KT</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MF</td>
<td>X</td>
<td>X</td>
<td>*</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RW</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>*</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TO</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>*</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A.2.2. Feature subsets selected through dynamic programming using FEC I - Reflection coefficients (6 out of 12). The asterisks (*) indicate parameters to be added for an 8-parameter subset.
<table>
<thead>
<tr>
<th>SPEAKER INITIALS</th>
<th>PARAMETER NO.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BN</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>*</td>
<td>X</td>
<td>*</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>*</td>
<td>X</td>
<td>*</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DV</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JD</td>
<td>X</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JZ</td>
<td>X</td>
<td></td>
<td>*</td>
<td>X</td>
<td></td>
<td>X</td>
<td>*</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KD</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>*</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KT</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>*</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MF</td>
<td>*</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RW</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>*</td>
<td>X</td>
<td>*</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TO</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>*</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A.2.3. Feature subsets selected through dynamic programming using FIC - LPC parameters (6 out of 12). The asterisks (*) indicate parameters to be added for an 8-parameter subset.
<table>
<thead>
<tr>
<th>SPEAKER INITIALS</th>
<th>PARAMETER NO.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32</td>
</tr>
<tr>
<td>AT</td>
<td>X X X X X X X X X</td>
</tr>
<tr>
<td>BN</td>
<td>X X X X X X X X X X X X X X X X</td>
</tr>
<tr>
<td>CC</td>
<td>X X X X X X X X X X X X X X X X</td>
</tr>
<tr>
<td>CS</td>
<td>X X X X X X X X X X X X X X X X</td>
</tr>
<tr>
<td>DV</td>
<td>X X X X X X X X X X X X X X X X</td>
</tr>
<tr>
<td>JD</td>
<td>X X X X X X X X X X X X X X X X</td>
</tr>
<tr>
<td>JZ</td>
<td>X X X X X X X X X X X X X X X X</td>
</tr>
<tr>
<td>KD</td>
<td>X X X X X X X X X X X X X X X X</td>
</tr>
<tr>
<td>KT</td>
<td>X X X X X X X X X X X X X X X X</td>
</tr>
<tr>
<td>MF</td>
<td>X X X X X X X X X X X X X X X X</td>
</tr>
<tr>
<td>RW</td>
<td>X X X X X X X X X X X X X X X X</td>
</tr>
<tr>
<td>TO</td>
<td>X X X X X X X X X X X X X X X X</td>
</tr>
</tbody>
</table>

Table A.2.4. Feature subsets selected through dynamic programming using FEC I - DFPT parameters (16 out of 32).
REFERENCES


32. M. Senidher, A. Badreldin, N. Mohankrishnan "Cluster analysis through functional mapping for text-independent speaker recognition" to be published.


VITA AUCTORIS

1953 Born on January 18th in Aurangabad, Maharashtra, India.

1968 Obtained the Indian School Certificate from Doveton Corrie Boys High School, Madras, India.

1974 Obtained the Bachelor of Technology Degree in Electrical Engineering from the Indian Institute of Technology, Madras, India.

1977 Obtained the Master of Technology Degree in Electrical Engineering from the Indian Institute of Technology, Kanpur, India.

1984 Candidate for the Ph.D. Degree in Electrical Engineering at the University of Windsor, Windsor, Ontario, Canada.