Adaptive Speech Analysis Techniques for Biometrics Applications

by

Tanweer Mozaffar,

A Project

Presented to Ryerson University

In partial fulfilment of the

requirement for the degree of

Master of Engineering

in the Department of

Electrical and Computer Engineering

Toronto, Ontario, Canada, 2003

Author's Declaration

I hereby declare that I am the sole author of this project report.

I authorize Ryerson University to lend this project report to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this project report by photocopying or by other means, in total or in part, at the request of the other institutions or individuals for the purpose of scholarly research.

Abstract

Utilizing biometrics for personal authentication is becoming convenient and considerably reliable and more accurate than current methods (such as: the utilization of passwords or personal identification number (PIN)). This is because; biometrics links the event to a particular individual (e.g. a password may be used by someone other than the authorized user), is convenient (nothing to carry or remember), accurate (provides positive authentication) and is becoming socially acceptable.

Adaptive frequency sub-band or time frame recombination approaches have been introduced in this project. As soon as any sub-band or time frame combination yields sufficiently confident and reliable information adaptively, feature extraction could be obtained from there and the sub-band or time frame that does not provide any significant information could be merged. In this project, we have implemented and compared different speech analysis techniques for biometrics applications. A feature extraction technique of speaker recognition investigated in this project is based on text-dependent scenario. The definition of speaker recognition used in this project is defined as: "a sample of speech from an unknown speaker is analyzed and compared with the speech samples from a set of known speakers stored in database. A choice is made, based on which speaker from the set of known speakers corresponds best to the unknown speaker."

Recognition accuracy rate in the range of 78%-86% for males and 60%-70% for females have been obtained with better computational speed for 8 band adaptive frequency sub-band filtering method for linear prediction coefficients (LPC), LPC derived cepstral coefficients (LPCC) and Mel frequency cepstral coefficients (MFCC) feature extraction techniques.

Keywords: Biometrics, sub-band adaptive filtering, adaptive segmentation, linear prediction coefficients (LPC), Cepstral coefficients (CC), Mel-frequency cepstral coefficients (MFCC).

Borrow List

Ryerson University requires the signatures of all persons using or photocopying this project report. Please sign, and give address and date.

Acknowledgments

The completion of this project would not have been possible without the valuable advice, continual guidance and technical expertise of my supervisor Prof. Dr. Sridhar Krishnan.

My gratitude goes to my wife Nausheen and my little daughter Anushah for their tolerance, patience, support and encouragement throughout my studies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The rapid advancement of telecommunications and computer technology in recent years is revolutionizing various sectors of industry. Although telecommunications and computer technology have been shared many technologies for sometime, and up to quite recently, they have been employed in distinctly different ways. Telecommunications concentrated more closely on immediate applications using continuous media such as sound and video, at the same time computer technology focused on the applications of storage, retrieval and manipulation of information. However, as the two technologies began to share in an integrated infrastructure, these distinctions between the two technologies started to disappear. A profound consequence of this has been the development of the new technological realities, which form the bases of our rapidly progressing information society. One such technological reality that warrants particular attention, for example is the Internet. This is because it is growing very rapidly and is used for a growing range of services. A large category of the services over the Internet includes those requiring a means of identification of users in order to restrict access to sensitive or personal data.

Examples of these services such as: fraudulent use of ATMs, cellular phones, smart cards, desktop PCs, workstations, and computer networks. In order to work for such type of services effectively, there is a growing need for methods to verify the identity of users reliably. The conventional means of identification such as passwords, secret codes and personal identification numbers (PIN) can easily be compromised, shared, observed or forgotten. In view of this, it appears that the required optimal reliability in determining the identities of users may only be achieved through the use of **biometrics.**

Biometrics is the oldest form of identification. Humans recognize each other's faces. On the telephone, your voice identifies you as who is the person on the line. On a paper contract, your signature identifies you as the person who signed it. Your photograph identifies you as the person who owns a particular passport. What makes biometrics useful from many of these applications is that it (biometrics) can be stored in a database, and retrieved at ease.

1.1 Biometrics

A biometrics is a unique, measurable, characteristic of a human being for automatically recognizing or verifying identity. This method of identification is preferred over conventional methods involving passwords and personal identification numbers (PIN) for various reasons such as:

- The person to be identified does not have to present anything but him-self or her-self.
- The critical variable for identification cannot be lost or forged.
- It (biometrics) uses unique and non-transferable physical or behavioral characteristics.
- It (biometrics) complement existing security systems combined with conventional authentication methods.

1.2 Biometrics Types

Biometrics techniques identify or verify the identity of an individual based on measurable physiological or behavioral characteristics. Therefore biometrics could be divided into two types: behavioral or physiological.

- *Behavioral:* Behavioral types are unique but variable. Behavioral characteristics are acquired over a period of time through the learning process and are more reflection of an individual's psychological make-up. Although in general, physical characters such as size and sex have a major influence on behavioral characteristics. Examples of behavioral type biometrics techniques are such as: i-Voice patterns (Speaker Recognition), ii-Signature verification and iii-Key Stroke patterns.

- *Physiological:* Physiological types are unique and permanent. Physiological characteristics are relatively stable physical characteristics that any individual is born with and does not change with time without significant duress. Examples of physiological type biometrics technique are such as: i-Finger prints, ii-Hand geometry, iii-Retinal & Iris scanning and iv-Facial recognition.

Table 1.1 presents the grading of different technologies; a grade of '4' denotes '*Very High* reliability or acceptance or stability', a grade of '3' denotes '*High* reliability or acceptance or stability', a grade of '2' denotes '*Medium* reliability or acceptance or stability' and a grade of '1' denotes '*Low* reliability or acceptance or stability'.

1.3 Biometrics: Truth and Fiction

The message is that, biometrics work great only if the verifier can verify two things: first, that the biometrics came from the person present at the time of verification, and second that the biometrics matches the master biometrics or the template biometrics on file. Biometrics is powerful and useful, but they are not

keys. They are useful in situations where there is a trusted path from the reader to the verifier; in those cases all you need is a unique identifier. Biometrics is unique identifier, but they are not secrets.

Technology	*Ease of Use*	*User Accept.*	*Accuracy*	*Long Term Stability*	*Error Incidence*
Finger Prints	**3**	**2**	**3**	**3**	**Dryness, Dirt, Age**
Hand geometry	**3**	**2**	**3**	**2**	**Hand Injury, Age**
Retina	**1**	**2**	**4**	**3**	**Glasses**
Iris	**2**	**2**	**4**	**3**	**Lighting**
Face	**2**	**2**	**3**	**2**	**Lighting, Age, Glasses, Hair**
Voice	**3**	**3**	**3**	**2**	**Noise, Colds**
Signature	**3**	**3**	**3**	**2**	**Changing Signatures**

Table1.1: Comparison & Grading of Different Biometrics Technologies [31]

1.4 Biometrics: The 'People' Element

Physical access biometrics deployments have major impact in any organization, and in deployment of biometrics technology the "people" element should be taken into consideration. This "people" element includes cultural background, hygiene issues and privacy concerns. For example, an employee might feel reluctant to press their fingers or palm on the biometrics device to gain access simply because of the hygienic concern or an employee who is exposed for the first time to an iris-scan or retina-scan device might afraid and concerned about the physical damage of his/her eye.

While now days there are many biometrics security devices are in use it is important to remember the human factor when deploying a physical access biometrics security solution. Organizations should consider the attitudes and perceptions of employees and personnel when asking them to volunteer their personal biometrics identities in exchange for access to company resources and information.

1.5 Speaker Recognition

Biometrics systems automatically recognize a person by using distinguishing traits. Speaker recognition is a performance biometrics i.e. you perform a task to be recognized. Your voice, like other biometrics, cannot be forgotten or misplaced, unlike knowledge-based (e.g. password) or professional-based (e.g. key) access control methods.

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information in speech waves. This technique makes it possible to use the speaker's voice to verify their identity and control access to services such as: voice dialing, telephone banking, database access services, security control for confidential information areas and remote access to computers.

Similar to humans, in order recognize voices, the voices must be familiar to machines. The process of 'getting to know' speakers is referred to as training and consists of collecting data from speech samples of people to be recognized. The second component of speaker recognition is testing namely the task of comparing unidentified speech samples to the training data and making the recognition. The speaker of a test speech sample is referred to as the target speaker. General overviews of speaker recognition have been given in [1] [2].

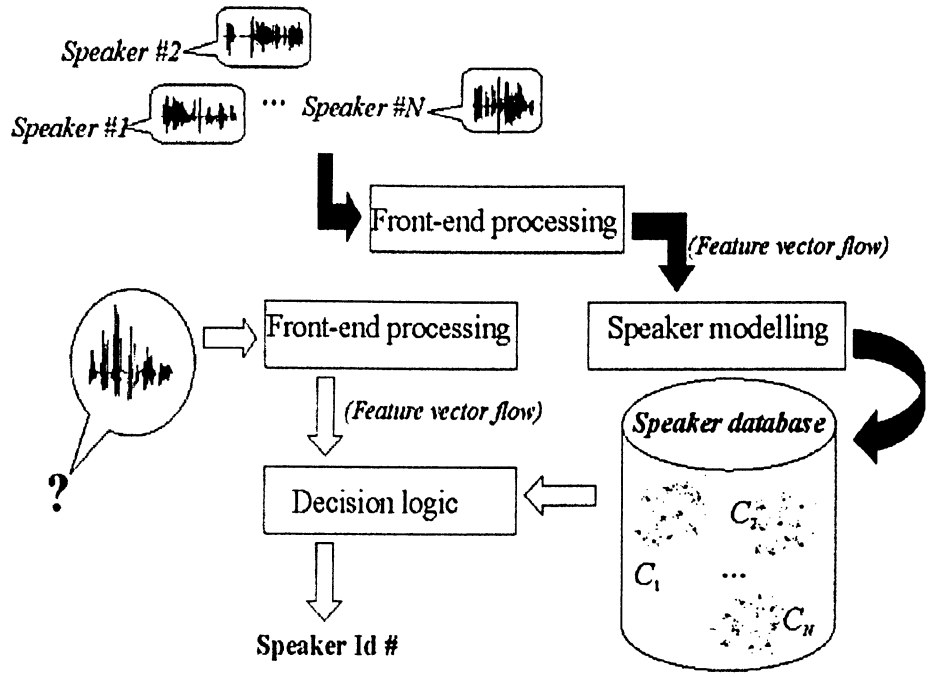


Figure 1.1: Typical Speaker Recognition System Block Diagram

1.6 Classification of Speaker Recognition

Speaker recognition can be divided into two classes:

Identification: Speaker identification is the process of determining which registered speaker provides a given utterance. Speaker identification exists in the area of speaker recognition, which includes both identification and verification of speakers.

Verification: Speaker verification, on the other hand is the process of accepting or rejecting the identity claim of a speaker. Speaker verification is further divided into:

- Text dependent: In text dependent, verification of the speaker's identity is based on his/her speaking of one or more specific phrases such as: password, card number and PIN number etc. [15].

- Text independent: In text independent, speaker models capture characteristics of somebody's speech which show up irrespective of what one is saying [23].

- Text prompted: In text prompted, speaker verification system will select a random word, read it to the called and ask the caller to repeat it exactly [22].

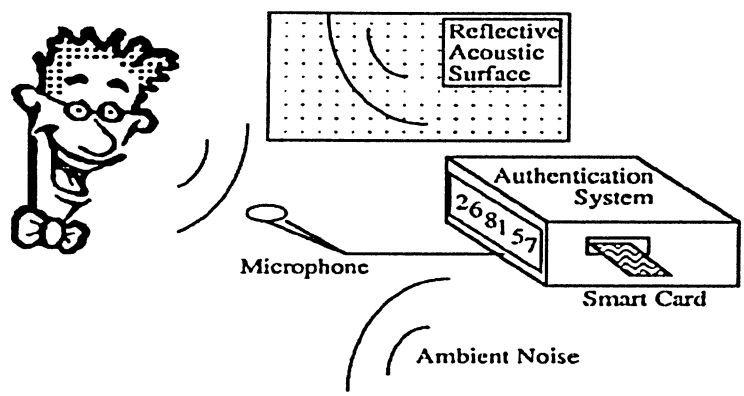


Figure 1.2: Typical Verification/ Identification setup. [8]

All technologies of speaker recognition have its own advantages and disadvantages and may require different treatment and techniques. The choice of which technology to use, is application specific. Many factors can contribute to identification and verification errors.

1. Misspeak or misread prompted phrases.
2. Extreme emotional states (e.g. stress or duress).
3. Time varying (intra-or intersession) microphone placement.
4. Poor or inconsistent room acoustics (e.g., noise)
5. Channel mismatch (e.g., using different microphones for enrollment and verification)
6. Sickness (e.g., head colds can alter the vocal tract)
7. Aging (the vocal tract can drift away from models with age).

Table 1.2: Human and environment error factors.

Table-1.2 lists some of the human and environmental factors that contribute to the errors. These factors generally are outside the scope of algorithms or better corrected by means of other than algorithms (i.e. better microphones). These factors are important, however because no matter how good a speaker recognition algorithm is, human error (e.g. misreading or misspeaking) ultimately limits its performance.

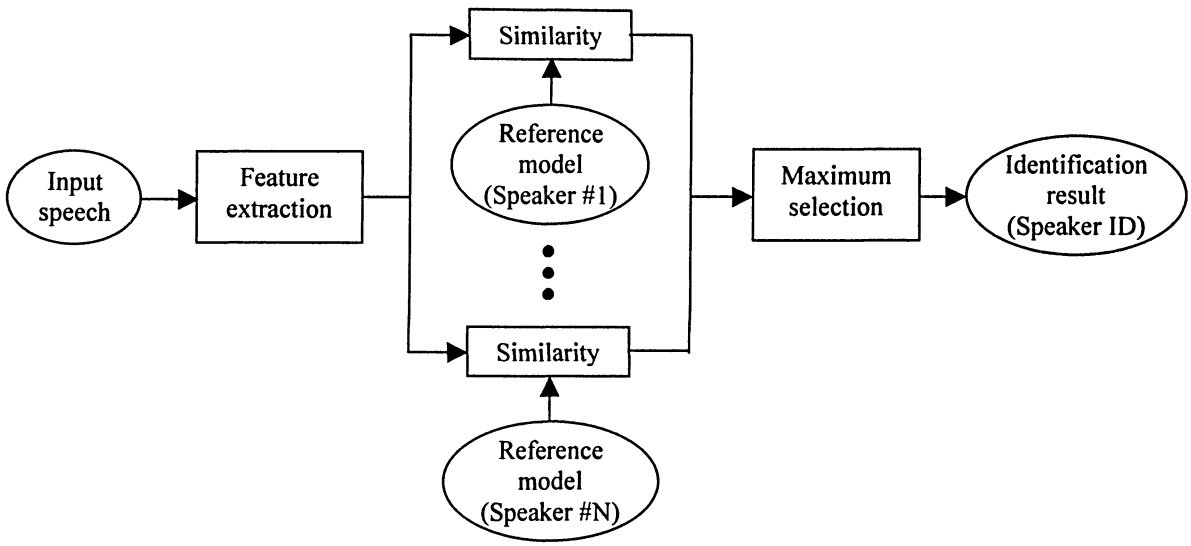


Figure 1.3: Typical Block Diagram of Speaker Identification

Typical block diagram of speaker identification is shown in Figure 1.3. After speech input of unknown speaker, a feature extraction technique is applied to extract feature model of the input speech sample. The extracted speech model is then compared with the stored database. Identification of the unknown speaker is made as to which speaker in the database best corresponds to the unknown speaker.

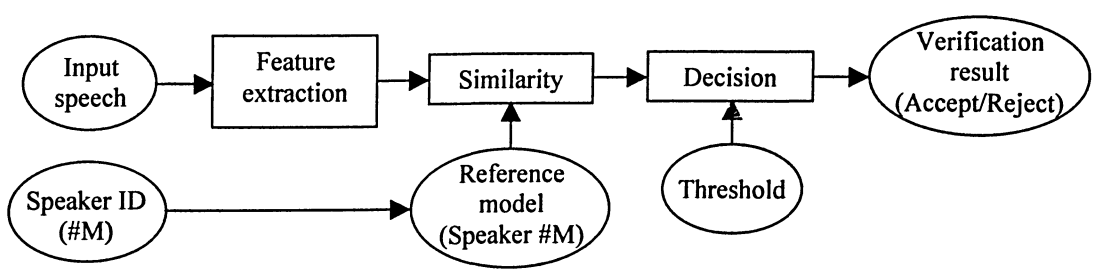


Figure 1.4: Typical Block Diagram of Speaker Verification

Typical block diagram of speaker verification is shown in Figure 1.4. After speech input of claimed speaker, a feature extraction technique is applied to extract feature model of the input speech sample. The extracted speech model and the ID of the claimed speaker is then compared with the stored database speech sample and database ID. With the help of suitable threshold value verification of the claimed speaker is made in result of acceptance or rejection as to which speaker and ID in the database best corresponds to the claimed speaker speech and ID.

1.7 Objective of the Project

Speech signals can be very different even for the same word, depending on speaker's stress and noise. Furthermore, the length of speech signal consists of thousands of samples even for a simple word, shows a large variation among individuals.

Therefore, in most of the speaker recognition algorithms, feature extraction is performed to reduce the dimensionality. Generally, in these cases speech signal is divided into either equal number of small fixed time frames such as: 25ms or 30ms frames, or divided into fixed number of frequency sub-bands and extract feature vectors for every time frame or frequency sub-band [13],[14].

Information in speech signal may not be severely loss as long as some of the sub-bands or time frames supply sufficiently reliable information and provide better feature vectors as compared to other sub-bands or time frames [15]. It is perhaps obvious that a core issue in the design of any sub-band or time frame system is the choice of the number and spanning of frequency sub-bands or time frames.

Feature vector information of a speech signal could be quite different of any individual for any particular frequency sub-band or time frame as compare to any other individual. Therefore in this scenario, information in one sub-band or time frame may provide significant information for one speaker while there will

not be any information of feature extraction for other speaker in the same span of frequency sub-band or time frame.

An adaptive recombination of sub-bands or time frames could provide better feature vectors (models) to discriminate them from other speech sample feature vectors. As soon as any sub-band combination yields sufficiently confident and reliable information adaptively, feature extraction could be obtained from there and the sub-band or time frame that does not provide any significant information could be merged.

Therefore, the objective of this project is to implement adaptive time segmentation and adaptive frequency sub-band filtering techniques in order to obtain optimal number of feature extraction vectors (model) of a given speech sample. In this project, we have introduced four different novel approaches of speech analysis to evaluate the performance of conventional feature extraction techniques such as: LPC, LPCC and MFCC.

Method-1 is based on non-zero time samples of a given speech signal. In this method we have used average feature vectors of every speaker out of their 10 speech samples to establish the database during training session.

Method-2 is based on adaptive frequency sub-band filtering approach. In this method, first the speech sample is divided into fixed number of frequency sub-bands. In order to get optimal number of frequency sub-bands, we have combined the selected sub-bands into a bigger sub-band in an adaptive way by applying weight against minimum vector distance between the sub-bands.

Method-3 is based on adaptive time segmentation approach. In this method, first the speech sample is divided into 25ms fixed time frames. In order to get optimal number of time frames we have combined the smaller time frames into bigger time frames in an adaptive way by applying weight against minimum vector distance between the time frames.

Method-4 has utilized the combined techniques of frequency sub-band and adaptive time segmentation together. In this method, first the speech sample is divided into fixed number of frequency sub-bands and then adaptive time segmentation technique has been applied for each sub-band respectively.

The results of all four methods in the context of LPC, LPCC and MFCC feature techniques are discussed in detail in Chapter 4.

1.8 Report Organization

This report is organized as follows: Speaker recognition techniques are reviewed in Chapter 2. Conventional feature extraction techniques of speech analysis including: linear prediction coefficients (LPC), LPC derived cepstral coefficients (LPCC) and Mel frequency cepstral coefficients are presented. Also overviews of feature matching (pattern) techniques are given such as: minimum distance measure, vector quantization (VQ) and hidden markov model (HMM). Chapter 3 presents the implementation of the four novel techniques introduced in this project for adaptive frequency sub-band and time segmentation. Chapter 4 covers the contribution of the project and provides the discussion and conclusion of the obtained results.

Chapter 2

Speaker Recognition Techniques

In speech communication systems, the speech signal is transmitted, stored and processed in many ways. Technical concerns lead to a wide variety of representations of the speech signal. In general, there are two major concerns in any system:

- Preservation of the message context in the speech signal.
- Representation of the speech signal in a form that is convenient for transmission or storage, or in a form that is flexible, so that modifications can be made to the speech signal without seriously degrading the message context.

The representation of the speech signal must be such that the information context can easily be extracted by human listeners or automatically by machine.

2.1 Speech Signal

Speech is a complicated signal, produced as a result of several transformations occurring at several different levels: linguistic, articulatory and acoustic. Differences in these transformations present a difference in the acoustic properties of the speech signal. Speaker related differences are a result of a combination of anatomical differences inherent in the vocal tract and the learned speaking habits of different individuals. In speaker recognition, all these differences can be used to discriminate speakers.

2.2 Signal Model

Real world processes generally produce observable outputs, which can be characterized as signals. The signals can be discrete or continuous, stationary or non-stationary and pure or corrupted in nature. A problem of fundamental interest is to characterize such real world signals in terms of signal models. There are several reasons why one is interested in applying signal models.

First of all, a signal model can provide the basis for a theoretical description of a signal processing system that can be used to process the signal so as to provide a desired output. For example, if we are interested in enhancing a speech signal corrupted by noise and transmission distortion, we can use the signal model to design a system, which will optimally remove the noise and undo the transmission distortion.

A second reason why signal models are important is that, they are potentially capable of letting us learn a great deal of about the signal source without having to have the source variable. Finally, the most important reason why signal models are important is that they often work extremely well in practice and enable us to realize important practical systems such as prediction systems and recognition systems in a very efficient manner.

The goal of speaker recognition is to obtain models of a speaker's patterns in feature space, which can be used to identify or verify the speaker of a test

utterance. Therefore an important step in the speaker recognition process is to extract sufficient information for good discrimination and at the same time to have captured the information in a form and size that is suitable to effective modeling.

Several popular signal analysis techniques have been emerged as standards in the literature. These algorithms are intended to produce a 'perceptually meaningful' parametric representation of the speech signals that imitate some of the behavior observed in the human auditory system. Of course, and perhaps more importantly these algorithms are also designed to maximize recognition performance.

2.3 Feature Extraction

The purpose of feature extraction is to convert the speech waveform of a speaker to some type of parametric representation for further analysis and processing. The speech signal is a slowly timed varying signal when examined over a sufficiently short period of time, i.e. its characteristics are quasi-stationary. However, long periods of time the signal characteristic change to reflect the different speech sounds being spoken. Therefore, short time spectral analysis is the most common way to characterize the speech signal. A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Coefficients (LPC), Cepstral Coefficients (CC) and derived versions of LPC and CC.

2.3.1 Linear Prediction Coefficients (LPC)

The techniques and methods of linear prediction have been available in the engineering literature for a long time. The basic idea behind linear predictive analysis is that a speech sample can be approximated as a linear combination of past speech samples. The linear prediction method provides a robust, reliable and accurate method for estimating time invariant system. The idea of LPC is

based on the speech production model which is the characteristic of the vocal tract, and can be modeled by an all pole filter.

The speech production process could be generally assumed to be the convolution of the excitation *E(z)* from the glottis and the all pole transfer function (vocal/tract/chords) *H(z)* to result in speech *S(z)* as shown in Figure 2.1. The excitation is periodic train pulses for the voiced speech and it is random for unvoiced speech. It is possible to represent the vocal tract in a parametric form as the transfer function *H(z)*. In order to estimate the parameters of *H(z)* from the observed speech waveform, it is necessary to assume some form for *H(z)*. Ideally, the transfer function should contain poles as well as zeros.

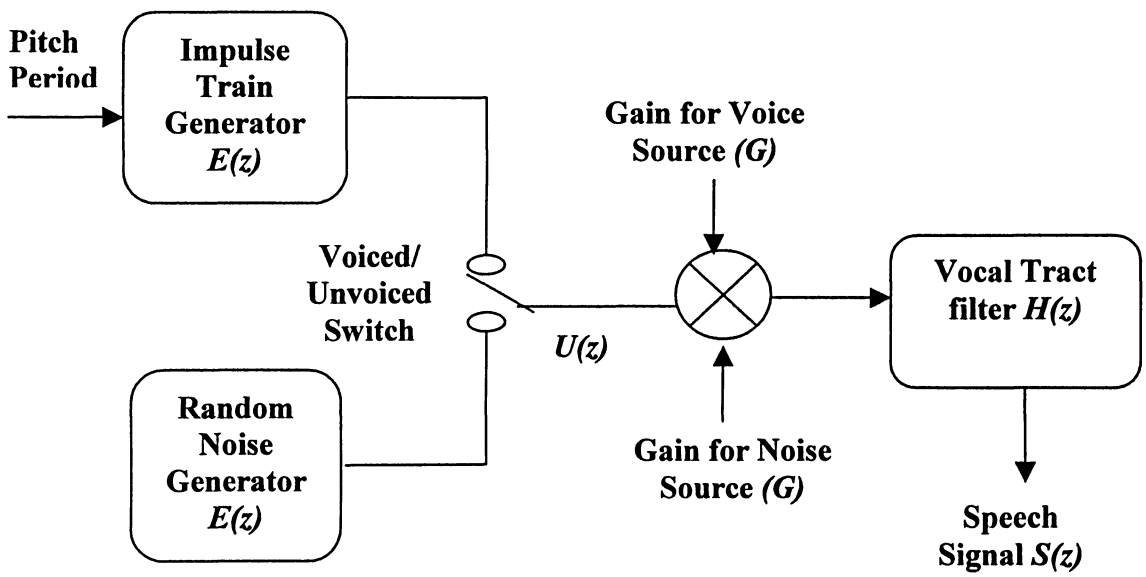


Figure 2.1: Block Diagram of Speech Production Model [9]

However, if only the voiced regions of speech are used then an all-pole model for *H(z)* is sufficient [4].

$$\hat{H}(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \qquad (2.1)$$

In equation 2.1 ' a_k 'denotes the LPC and ' p ' is the model order.

2.3.1.1 Formulations of LPC

As applied to speech processing, the term linear prediction refers to a variety of essentially equivalent formulations of the problem of modeling of the speech waveform. The differences among these formulations are often those of philosophy or way of viewing the problem.

The common technique of LPC formulation is the autocorrelation method, which multiplies the signal by a time window *w(n)* so that *x(n)=w(n)*s(n)* has finite number of durations. Thus *x(n)=0* outside the range of $0 \leq n \leq N-1$. LPC models, all *x(n)* samples within each frame; thus when the signal is non-stationary, the LPC coefficients describe a smoothed average.

Let *E* be the energy in the error:

$$E = \sum_{-\infty}^{+\infty} e^2(n) = \sum_{-\infty}^{+\infty} \left[x(n) - \sum_{k=1}^{p} a_k x(n-k) \right]^2 \qquad (2.2)$$

Where *e(n)* is the residual corresponding to the windowed signal *x(n)*. The value of a_k that minimize *E* are found by setting $\partial E / \partial a_k = 0$ for $k=1,2,3,\ldots P$. This yields *'P'* linear equations ($i=1,2,3,\ldots$P) in *'P'* unknowns a_k.

2.3.1.2 Solutions of the LPC Equations

In order to effectively implement a linear predictive analysis system, it is necessary to solve the linear equations in an efficient manner. Although a variety of techniques can be applied to solve a system of *'P'* linear equations in *'P'* unknowns, but these techniques are not equally efficient. Because of the special properties of the coefficient matrices it is possible to solve the equations much more efficiently than is possible in general.

In autocorrelation method, the P linear equations to be solved can be viewed in matrix form as $\boldsymbol{Ra} = \boldsymbol{r}$, where '$R$' is a '$P$ x P' matrix of elements; '$\boldsymbol{r}$' is a column vector $[R(1),R(2),...,R(p)]^T$, and '$\boldsymbol{a}$' is a column vector of LPC coefficients $[(a_1,a_2,...a_p)]^T$. Computing the LPC vector directly requires inversion of the '$\boldsymbol{R}$' matrix and multiplication of the resultant 'P x P' matrix with the 'r' vector. But due to the Toeplitz nature of the matrix of coefficients 'R' allows the efficient Levinson-Durbin recursive procedure to compute a_k coefficients [4].

2.3.2 Cepstral Coefficients (CC)

Linear predictive analysis is one of the powerful techniques used in speech analysis. A by-product of the LPC analysis is the generation of prediction residues, or prediction errors $e(n)$. Theoretically, if the all-pole model was perfect, $e(n)$ would be very small. Unfortunately, this simplified model is not suitable for nasal and fricative sounds, the detailed acoustic theory calls for both poles and zeros in the vocal tract transfer function. $e(n)$ essentially carry all information not captured by the LPC coefficients. In speaker recognition $e(n)$ are usually ignored, only the LPC coefficients or some transformations of the LPC parameters (e.g., LPCC are used to compose feature vectors [11]).

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \qquad (2.3)$$

Where $s(n)$ are speech samples, a_k are LPC coefficients, 'P' is the LPC analysis order. Therefore cepstral coefficients were another transform developed as a primary improvement over the direct usage of the linear prediction coefficients as feature vectors. As shown in Figure 2.1 it de-convolved $S(z)$ to yield $E(z)$ and $H(z)$. If $s(n)$ is the input speech signal, the real cepstrum is given by [5] and is computed as; Real *Cepstrum* = $ifft(\log|fft(s(n))|)$.

$E(z)$ is an intra-speaker varying feature and varies depending on the emotional status and age of an individual. $H(z)$; in general is assumed to be unique for each

speaker. The lower frequency contents of the cepstrum represents *H(z)* and the higher frequencies represents *E(z)*.

2.3.2.1 Linear Prediction Cepstral Coefficients (LPCC)

The majority of speaker recognition systems use some type of short-time spectral analysis. The methods used usually assume that speech is a short-time stationary process. The LPC cepstral coefficients *c(i)* for *p*-th order linear prediction coefficients *a(i)* is given by [12]:

$$c(1) = -a(1)$$

$$c(i) = -a(i) - \sum_{k=1}^{i-1} (1 - {}^{k}\!/_{i})\, a(k)\, c(i-k); 1 \le i \le p \quad (2.4)$$

LPCC have been widely used for a few decades and it has been proven that it is more robust and reliable than LPC.

2.3.3 Mel Frequency Cepstral Coefficients (MFCC)

While most feature extraction techniques attempt to capture information on the vocal tract transfer function from the gross spectral shape of the input speech, the accuracy and robustness of the speech representation may deteriorate dramatically due to the spectral distortion caused by the additive background noise. The well-known MFCC [3], though adopted by most automatic speaker recognition systems for its superiority in clean speech recognition, do not cope well with noisy speech. This is the most popular form of parameterization for speech recognition. This is a feature, derived based on the psychoacoustics modeling which studies human auditory perception.

One main difficulty in conventional feature extraction algorithms is concerned with the vocal tract transfer function whose accurate and robust description is crucial to effective speaker recognition.

The human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency 'f'

measured in Hz, a subjective pitch is measured on a scale called the 'Mel' scale. The Mel-frequency scale is linear frequency spacing below 1kHz and a logarithmic spacing above 1kHz. Therefore for 'f_{hertz}' frequency and 'f_{mel}' Mel-scale, the approximation is given by [3]:

$$f_{mel} = 2595 * \log_{10}(1 + f_{hertz}/700) \qquad (2.5)$$

2.4 Feature (Pattern) Matching

The feature (pattern) matching involves the computation of a matching score between the input feature vector (to be authenticated) and the stored speaker feature models, which are constructed from the speech signals.

The speaker models are either template models or stochastic models. The stochastic modeling assumes the speech production process to be a non-parametric random process and the template models assumes speech production to be a parametric random process. The vector quantization (VQ) source modeling and distance measure are some of the important template modeling techniques; while hidden markov model (HMM) has been widely used as a stochastic model for modeling the speech production.

2.4.1 Template Models

The process of pattern matching is deterministic for template models. The pattern matching process involves the comparison of a given set of input feature vectors against the speaker model for the claimed identity and computing a matching score. The two types of template models, distance measure and vector quantization are discussed below.

2.4.1.1 Distance Measure

A Euclidean distance measure is perhaps the most famous technique used for template matching. The distance between the stored speaker model $\hat{x}$ and the input speech feature vector x_i is given by [5]:

$$d = \left[\sum (\hat{x} - x_i)^2\right]^{1/2} \qquad (2.6)$$

Where 'd' is the minimum distance between the two vectors for Euclidean distance measure method.

2.4.1.2 Vector Quantization

Vector quantization (VQ) is another template modeling. In the VQ source modeling, a codebook is designed to represent the frames of speech by clustering them by standard clustering procedures. The two important factors for the VQ-method of modeling is the size of codebook designed and the method to generate the codebook [5].

The best way is to form a codebook for each speaker where the codebooks are represented by their centroids. These codebooks are then stored as database of the enrolled speakers. If an input speaker has to be authenticated, the distortion is calculated between the feature vector and each of the codebooks. The unknown speaker is identified as the one, which has the lowest distortion.

2.4.2 Stochastic Modeling

A stochastic model that is very popular for modeling sequences is the Hidden Markov Model (HMM). A probabilistic function of a (Hidden) Markov chain is a stochastic process generated by two interrelated mechanisms, an underlying Markov chain having a finite number of states, and a set of random functions, one of which is associated with each state. At discrete instants of time, the process is assumed to be in a unique state and an observation is generated by the

random function corresponding to the current state. The underlying Markov chain then changes states according to its transition probabilistic matrix [6].

The observer sees only the output of the random functions associated with each state and can not directly observe the state of the underlying Markov chain, hence the term Hidden Markov model

2.5 Summary

Due to the simplicity with distance measure model, it is adapted as the method of choice in this project. In summary, a key issue for implementing an accurate speaker recognition system is the set of acoustic features extracted from the speech signal. The set is required to convey as much speaker dependent information as possible. The standard methodology to extract these features from the signal follows the use of LPC feature extraction technique.

LPC based on the underlying assumption that acoustic characteristics of human speech are mainly due to the vocal tract resonance, which form the basic spectral structure of the speech signal. However, human speech is a nonlinear phenomenon, which involves nonlinear biomechanical, and physiological factors and therefore, LPC derived parameters such as LPCC and MFCC can offer sub-optimal description of the speech dynamics.

Speaker recognition uses the acoustic features of speech that have been found to differ between individuals. These acoustic patterns reflect both anatomy (e.g., size and shape of the throat and mouth) and learned behavioral patterns (e.g., voice pitch, speaking style). This incorporation of learned patterns into the voice templates has earned speaker recognition its classification as a "behavioral biometrics."

The various technologies used to process and store voice template include LPC, LPCC and MFCC. Performance degradation can result from changes in behavioral attributes of the voice and from enrollment using one microphone and verification on another microphone.

Chapter 3

Implementation and Results

To apply mathematical tools without loss of generality, the speech signal can be represented by a sequence of feature vectors. Traditionally pattern recognition models are divided into two components: feature extraction and pattern matching. Although this division is convenient from the perspective of designing system components, these components are not independent.

A key issue for implementing an accurate speaker recognition system is the set of acoustic features extracted from the speech signal. This set is required to convey as much speaker dependent information as possible. Therefore, the goal is to design a system that minimizes the probability of verification or identification errors. Thus, the underlying objective is to discriminate between the given speaker and all others.

In this project we have evaluated the performance of speech analysis techniques for biometrics applications based on LPC, LPC derived CC (LPCC) and MFCC. For fairness of comparison we kept all different feature vectors of the same dimensionality (10^{th} order LPC). It is known in literature that a 10^{th} order LPC

accurately represents/models voiced and unvoiced speech and is used as the model order in this project.

3.1 Speech Database

We know that there is a great variability among speakers producing the same vowel. Therefore in order to achieve significant discrimination among speech models of different speakers we have decide to use the word 'HELLO'. The word 'HELLO' consists of 'H (whisper)', 'E (vowel)', 'L (semi-vowel)' and 'O (vowel)'. Semi-vowels are quite difficult to characterize. They are best described as vowel-like sounds and hence are similar in nature to the vowels. While the characteristics of 'H', are invariably those vowels, which follows 'H' therefore vocal tract assumes the position for the following vowel during the production of 'H'.

One hundred and twenty (120) speech samples from two males and two females in age from 25 to 40 years, with the same regional accent and without noticeable defects were chosen for the recordings. They were aware of the nature of the experiment. In order to take into account the important effects of changes in speaker's voice over time, the recordings were made on two different days with an interval of about 3-months. In each recording session, the speakers said the same word 'HELLO'.

The words were spoken in a quiet environment into a computer microphone with audio format of PCM, 8kHz, 8 bit, Mono. Recording has been made in two sets of speaker pairs (1-male and 1-female) on two different computers with different microphones. Speaker#1 and speaker#2 were recorded on one set of computer and speaker#3 and speaker#4 were recorded on another set of computer.

3.2 Experimental Setup

Four different methods of speech signal analysis in MATLAB environment have been implemented to evaluate the performance of LPC, LPCC and MFCC feature extraction techniques.

3.2.1 Method-1

This method is based on non-zero time samples of a given speech signal. All three features extraction techniques (LPC, LPCC and MFCC) have been implemented and tested in this method. In order to establish the speaker model ten speech samples of two males and two females each, have been recorded during training session. The reason of recoding ten speech samples of each speaker is to get the average 10^{th} order LPC model of each speaker.

Training Phase:

As shown in Figure 3.1, ten speech samples of every speaker have been taken from stored database. Only non-zero values have been selected of each speech sample and calculated the 10^{th} order LPC/LPCC/MFCC model. The average 10^{th} order LPC/LPCC/MFCC model has been calculated from all ten samples and saved as template (feature vector) of known speakers. Same procedure has been repeated for rest of speakers' speech samples.

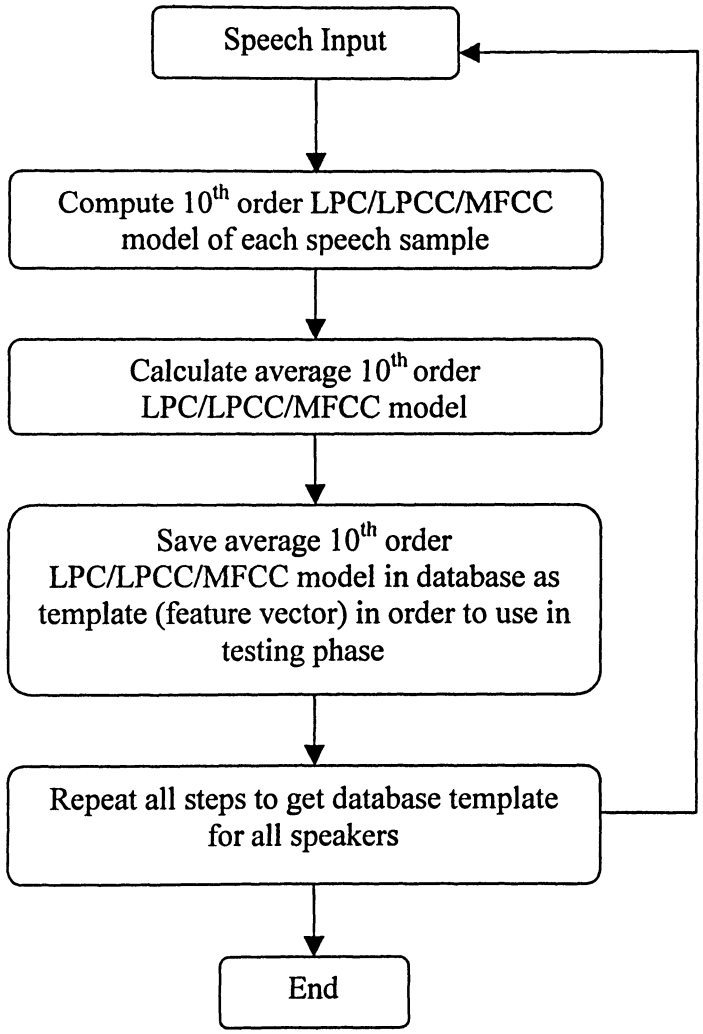


Figure 3.1: Method-1, Typical Training Phase Flow Diagram

Testing Phase: Since we have implemented our experiment in the context of text-dependent speaker recognition. Therefore, the unknown speaker to be recognized has to record the same word 'HELLO' which has been used during training session. As explained in Figure 3.2, 10^{th} order LPC/LPCC/MFCC model of unknown speaker has been calculated by following the same procedure as in training session of this method. Minimum vector distance has been computed between unknown speaker's feature vector and all known templates (feature vectors) from database with the help of Euclidean distance measure method. Recognizing the speaker whose vector distance is minimum among all speakers.

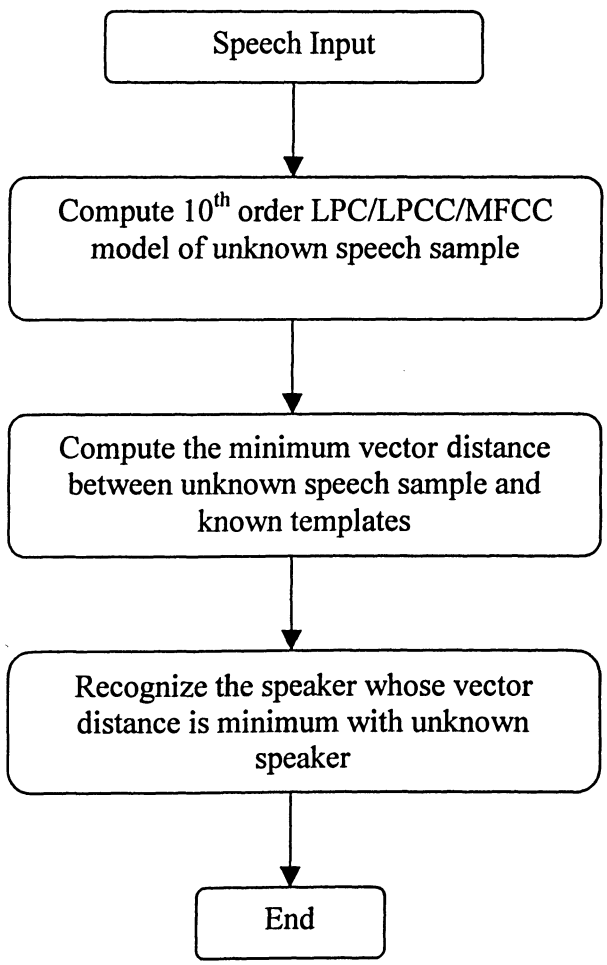


Figure 3.2: Method-1, Typical Testing Phase Flow Diagram

All the above test experiments were carried out with 120 samples of four speakers. The results of this method are as shown in Table 3.1

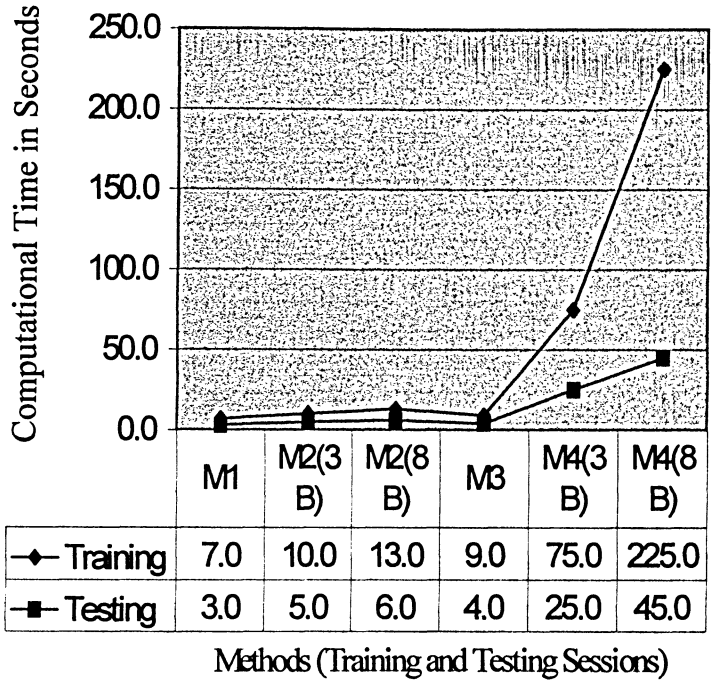


Table 3.1: Method-1 Results

3.2.2 Method-2

In the conventional speech feature extraction process, each feature vector is generated by utilizing the entire frequency spectrum of a given speech frame. Therefore, when the speech signal is partially degraded by an anomaly, which is localized in time and frequency, the feature vectors that are generated within the time span of that anomaly are completely contaminated. A logical way to tackle this problem is to split the entire frequency domain into a number of sub-bands and to use the spectral information contained in each of these sub-bands to extract independent feature vectors.

A critical issue of using sub-band is that the sub-band vectors consists of less spectral information than that of the full band. A possible method to handle this

problem is to recombine the intermediate sub-band outcomes at certain stage. Ideally, for this purpose, the outcomes of different sub-bands are to be combined in a constructive way so that sub-bands that are specific to the target speaker are emphasized while others are de-emphasized (combined). In this method of experiment we have used three sub-band (spanning [0-0.4kHz, 0.4-4kHz and 4-8kHz]) and eight sub-band (spanning [0-1kHz, 1-2kHz, 2-3kHz, 3-4kHz, 4-5kHz, 5-6kHz, 6-7kHz and 7-8kHz]) approach.

Training Phase: The training phase is shown in Figure 3.3. Speech samples of every speaker have been taken from stored database. Then filter the speech sample into sub-bands with the help of a digital filter with ten filter coefficients. After division of speech sample into sub-bands the 10^{th} order LPC/LPCC/MFCC model have been calculated within each sub-band. Since this method is an adaptive sub-band filtering, therefore in order to get optimal number of sub-bands, Euclidean distance measure method has been applied between the LPC/LPCC/MFCC models of every sub-band. Decision of combining the sub-bands into one band has been taken against the weight applied on minimum distance measure between sub-bands.

After getting the optimal number of sub-bands, the 10^{th} order LPC/LPCC/MFCC model has been calculated again for every sub-band and saved in database as a template.

Testing Phase: The testing phase is explained in Figure 3.4. The unknown speaker has recorded the same word 'HELLO', which has been used during training session. The 10^{th} order LPC/LPCC/MFCC model (feature vector) of unknown speaker have been calculated for all optimal number of sub-bands by following the same procedure as in training session of this method. Euclidean distance measure has been applied between unknown speaker template and all known template in database to recognize the unknown speaker. Recognizing the speaker whose vector distance is minimum among all speakers.

All the above experiments were carried out with 120 speech samples of four speakers. The results are as shown in Table 3.2 and Table 3.3.

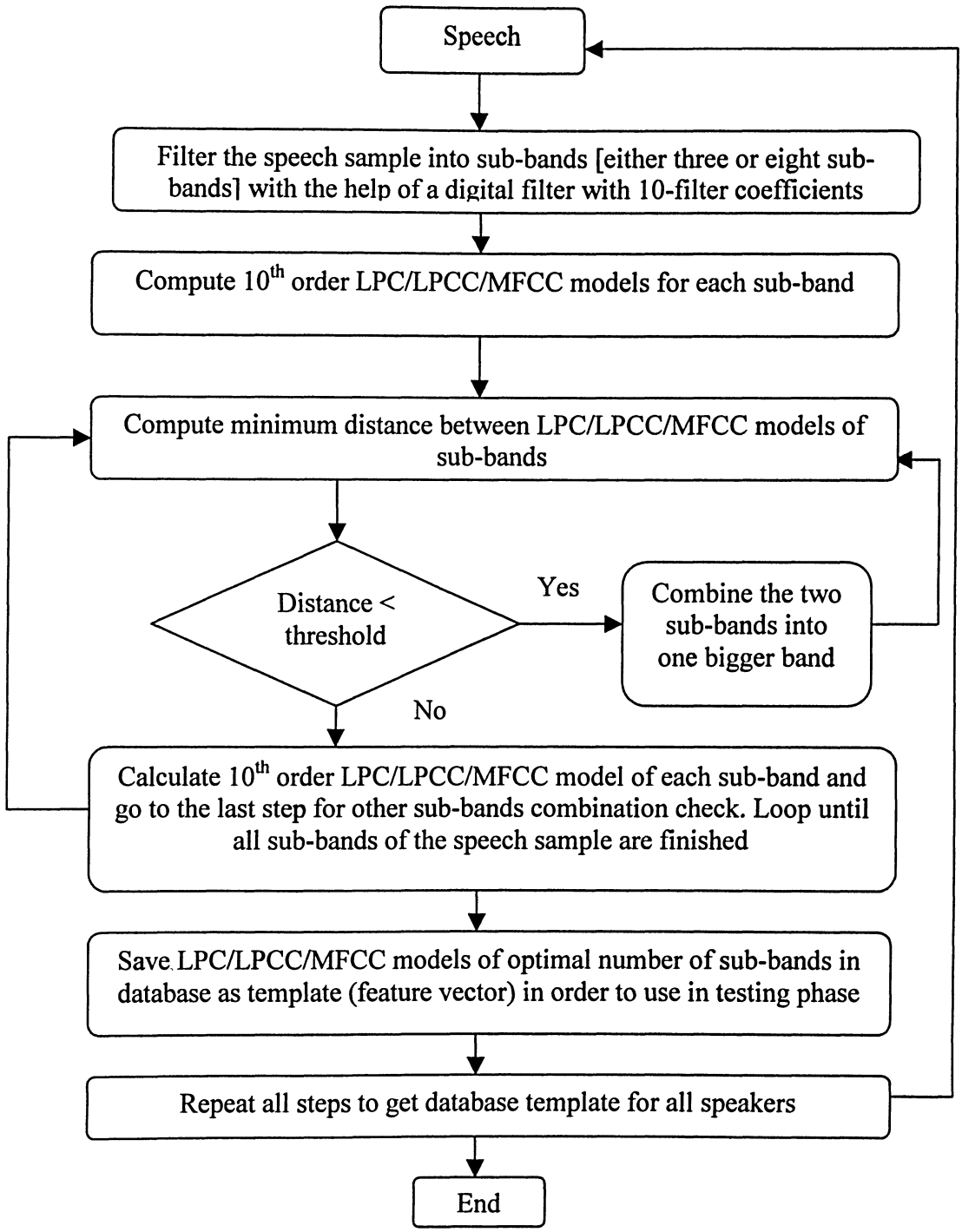


Figure 3.3: Method-2, Typical Training Phase Flow Diagram

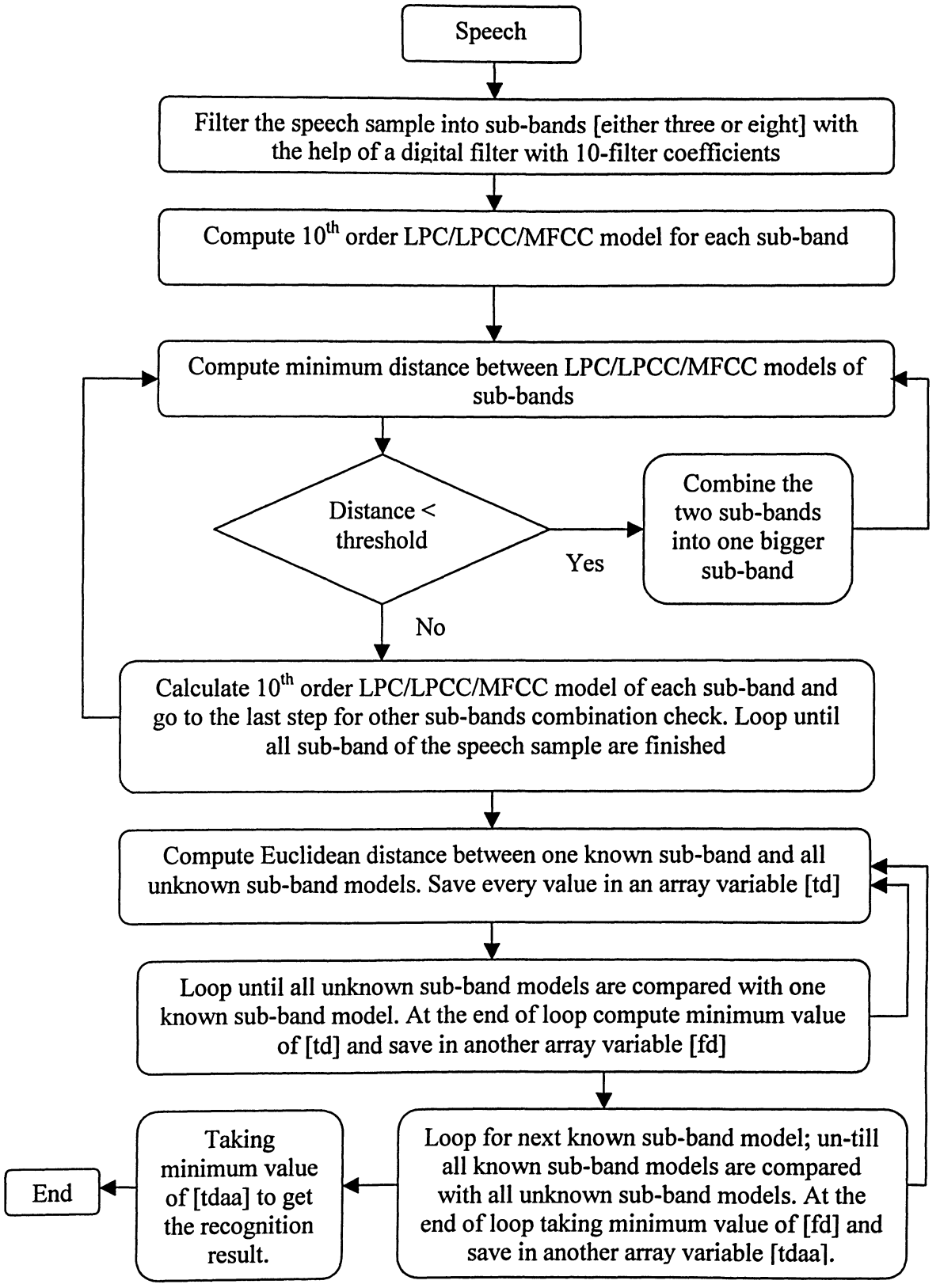


Figure 3.4: Method-2, Typical Testing Phase Flow Diagram

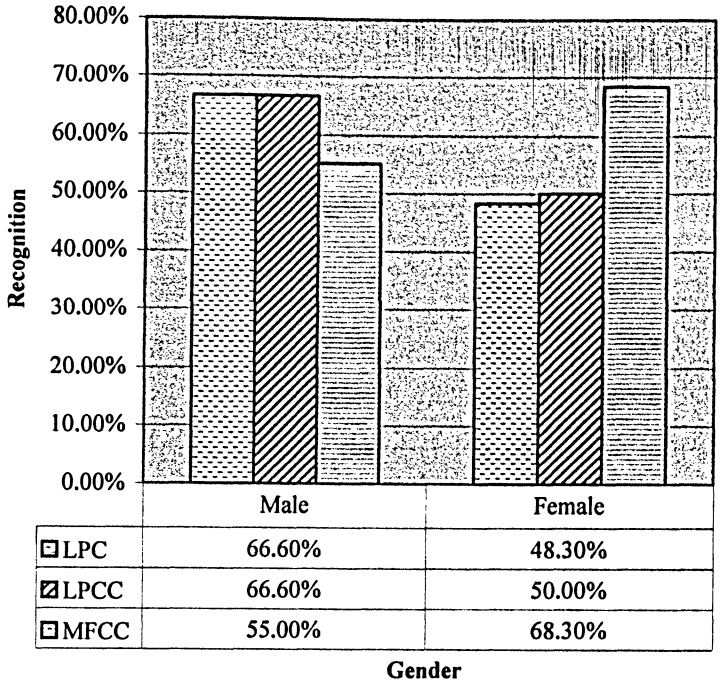


Table 3.2: Method-2 Results (3 Sub-bands Approach)

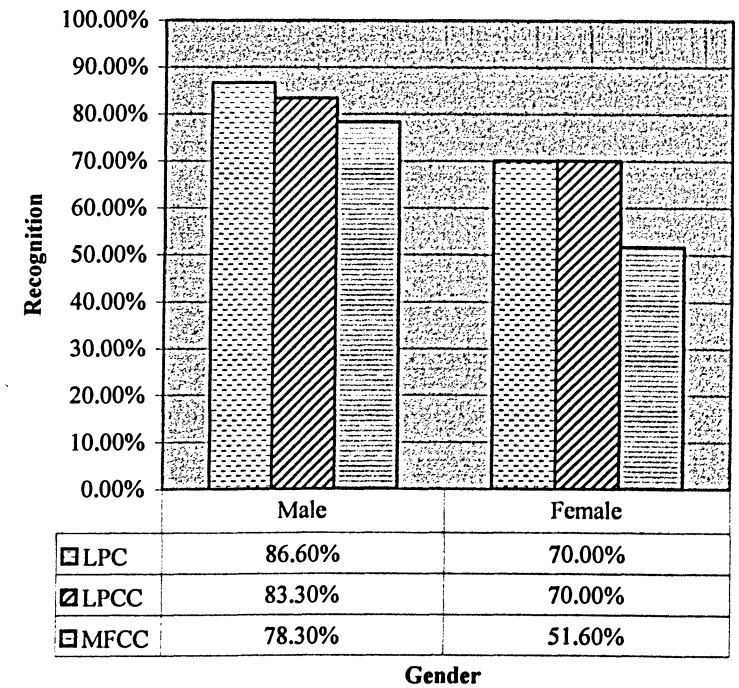


Table 3.3: Method-2 Results (8 Sub-bands Approach)

3.2.3 Method-3

In this method of experiment first the speech sample is divided into 25ms time frame interval and then adaptive time segmentation is applied. We have implemented and tested all three features extraction techniques LPC, LPCC and MFCC in this method.

Training Phase: The training phase is shown in Figure 3.5. Speech samples of every speaker have been taken from stored database. The speech signal is then divided into 25ms time frames. After the division of speech signal into small time frame, the 10^{th} order LPC/LPCC/MFCC has been calculated within every time frame. In order to get optimal number of time frames, Euclidean distance measure method has been applied between the LPC/LPCC/MFCC models of each time frame. Decision of combining the time frames into bigger time frames has been taken against the weight applied on minimum distance measure between time frames.

After getting the optimal number of time frames, the 10^{th} order LPC/LPCC/MFCC has been calculated again of each final time frame and saved in database as template (feature vector) of known speakers. Same procedure has been repeated for the rest of speakers' speech samples.

Testing Phase: The testing phase is explained in Figure 3.6. The unknown speaker has recorded the same word 'HELLO', which has been used during training session. The 10^{th} order LPC/LPCC/MFCC model (feature vector) of unknown speaker for all optimal number of time frames have been calculated by following the same procedure as in training session of this method. Euclidean distance measure has been applied to recognize the unknown speaker. Recognizing the speaker whose vector distance is minimum among all speakers.

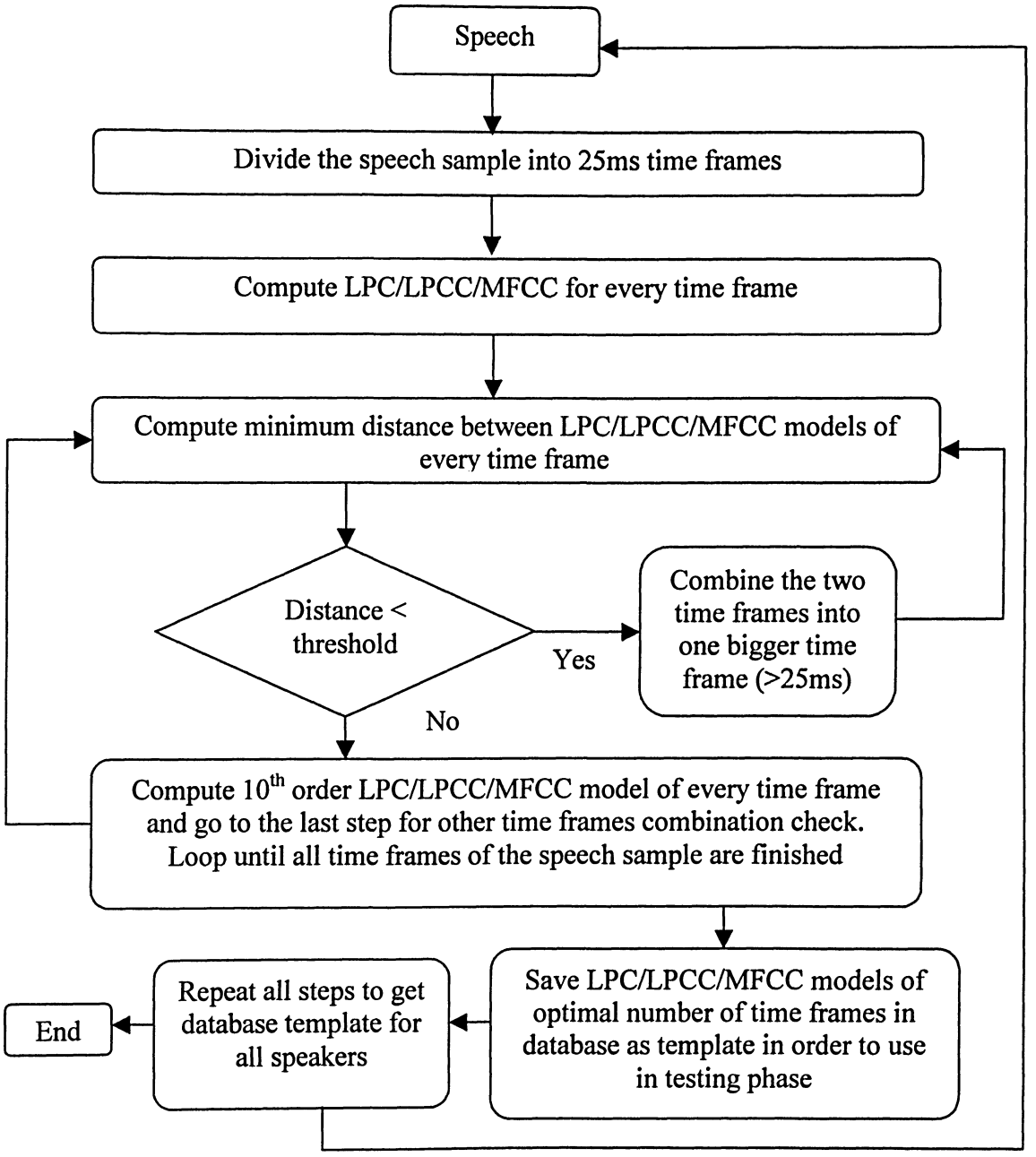


Figure 3.5: Method-3, Typical Training Phase Flow Diagram

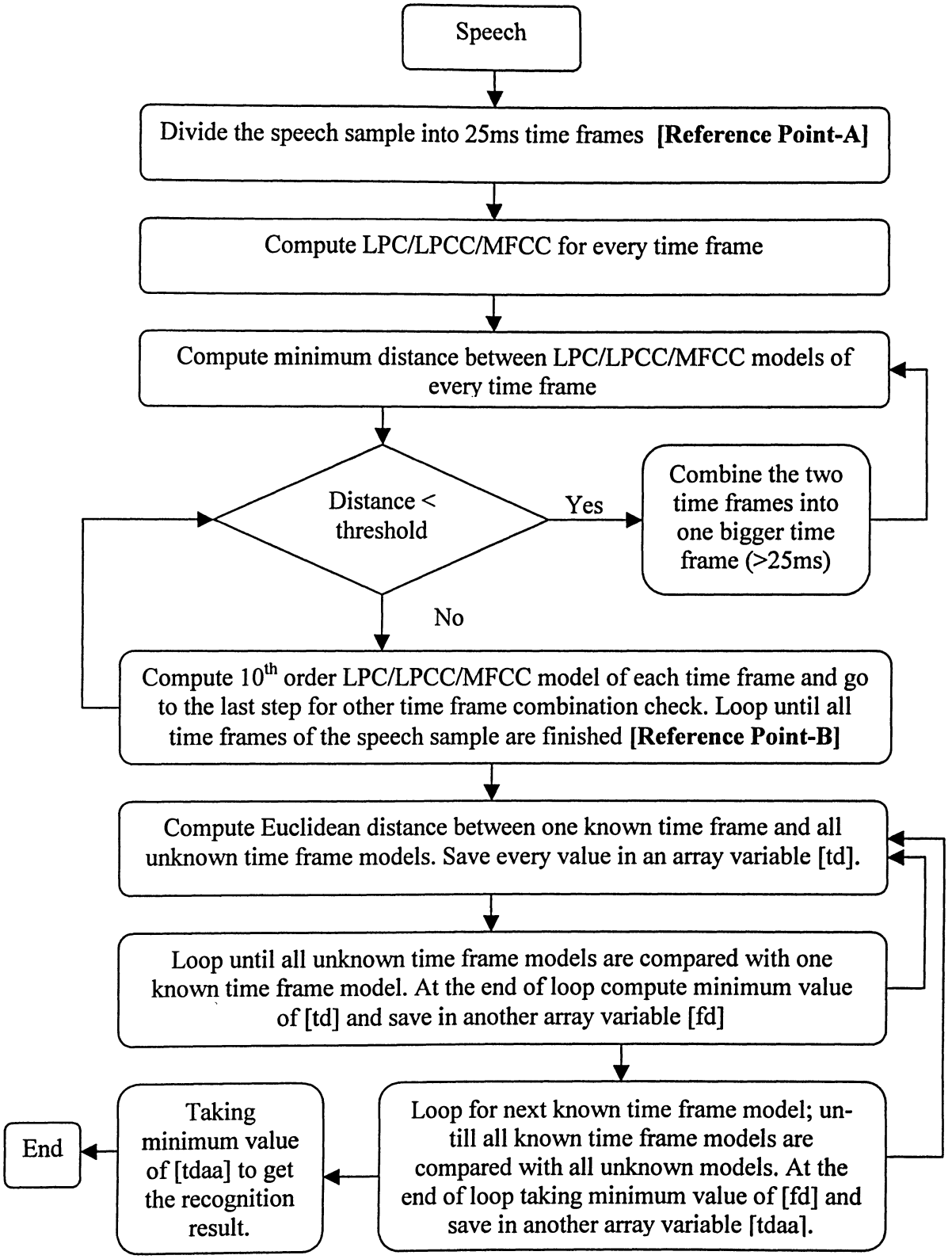


Figure 3.6: Method-3, Typical Testing Phase Flow Diagram

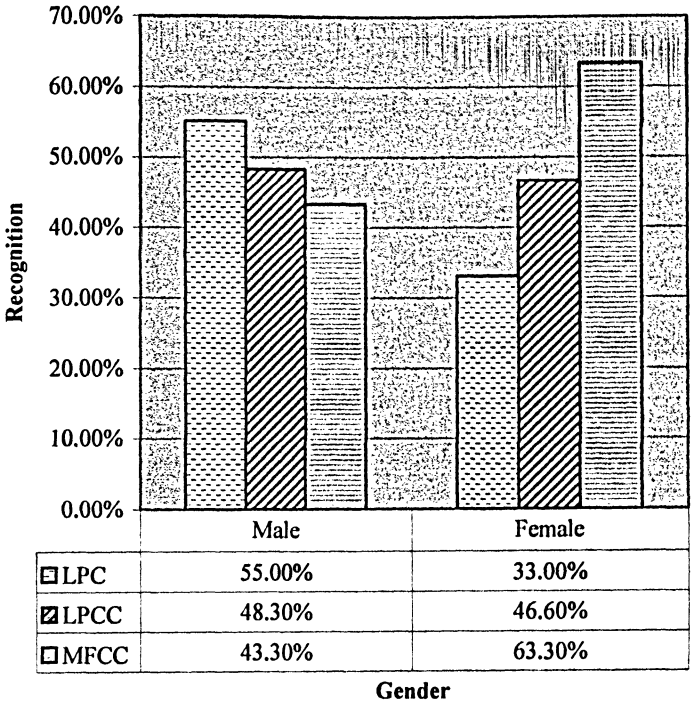


Table 3.4: Method-3 Results

3.2.4 Method-4

In this method frequency sub-band and adaptive time segmentation techniques are combined together. In this method first the speech sample is divided into fixed number of sub-bands and then adaptive time segmentation is applied in order to compute the optimal number of time frames in every sub-bands respectively.

Training Phase: The training phase is shown in Figure 3.7. Speech samples of every speaker have been taken from stored database. Then the speech sample is filtered into sub-bands (either 3 or 8 sub-bands) with the help of a digital filter with ten filter coefficients. After the sub-band filtering the speech signal is divided into 25ms time frame with in each sub-band. The 10^{th} order LPC/LPCC/MFCC model has been calculated within each time frame. Euclidean distance measure method has been applied for LPC/LPCC/MFCC models of each time frame. Decision of combining the time frames into bigger time frames has been taken against the weight applied on minimum distance measure between the time frames.

After getting the optimal number of time frames, the 10^{th} order LPC/LPCC/MFCC have been calculated again of each time frame within every sub-band and saved as template in database. Number of feature vectors in every sub-band has been calculated by following the same steps. Same procedure has been repeated for the rest of speakers' speech samples.

Testing Phase: The testing phase is explained in Figure 3.8. The unknown speaker has recorded the same word 'HELLO', which has been used during training session. The 10^{th} order LPC/LPCC/MFCC model (feature vectors) of unknown speaker for optimal number of time frames in every sub-band have been calculated by following the same procedure as in training session of this method. Euclidean distance measure has been applied to recognize the unknown speaker. Recognizing the speaker whose vector distance is minimum among all speakers.

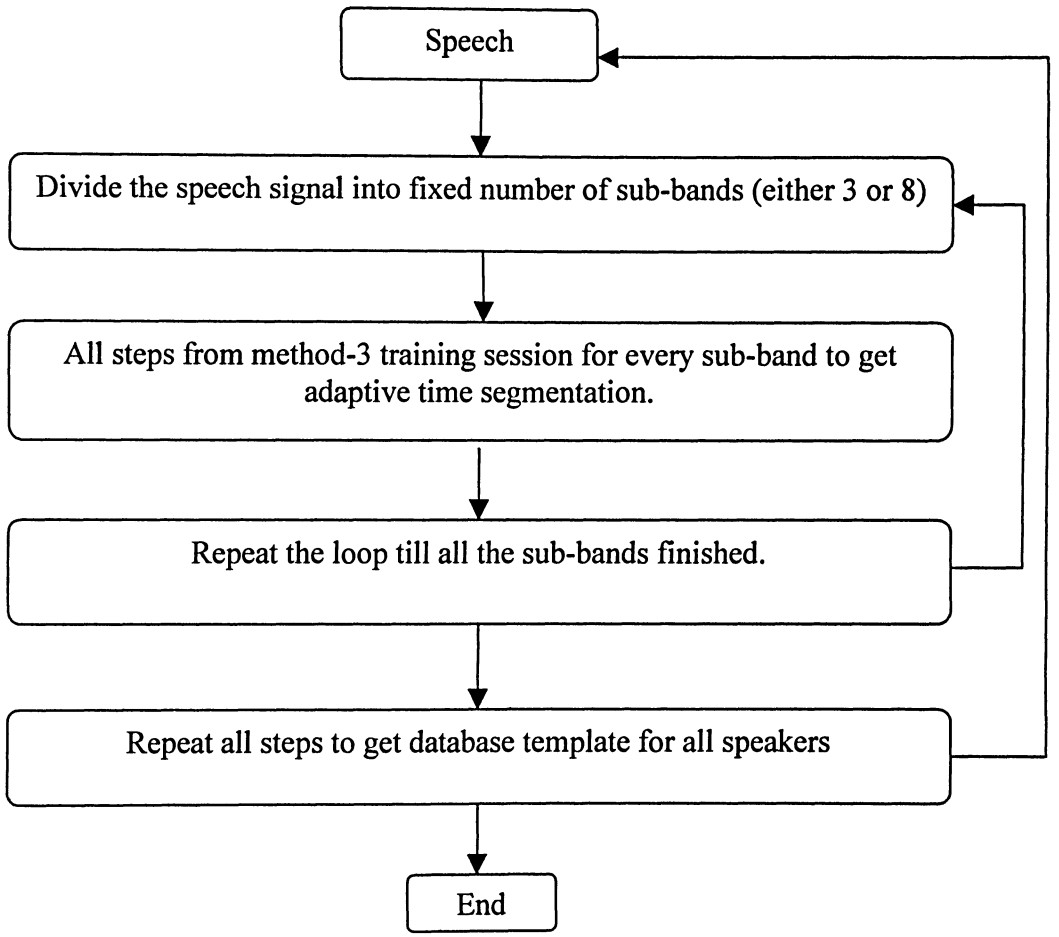


Figure 3.7: Method-4, Typical Training Phase Flow Diagram

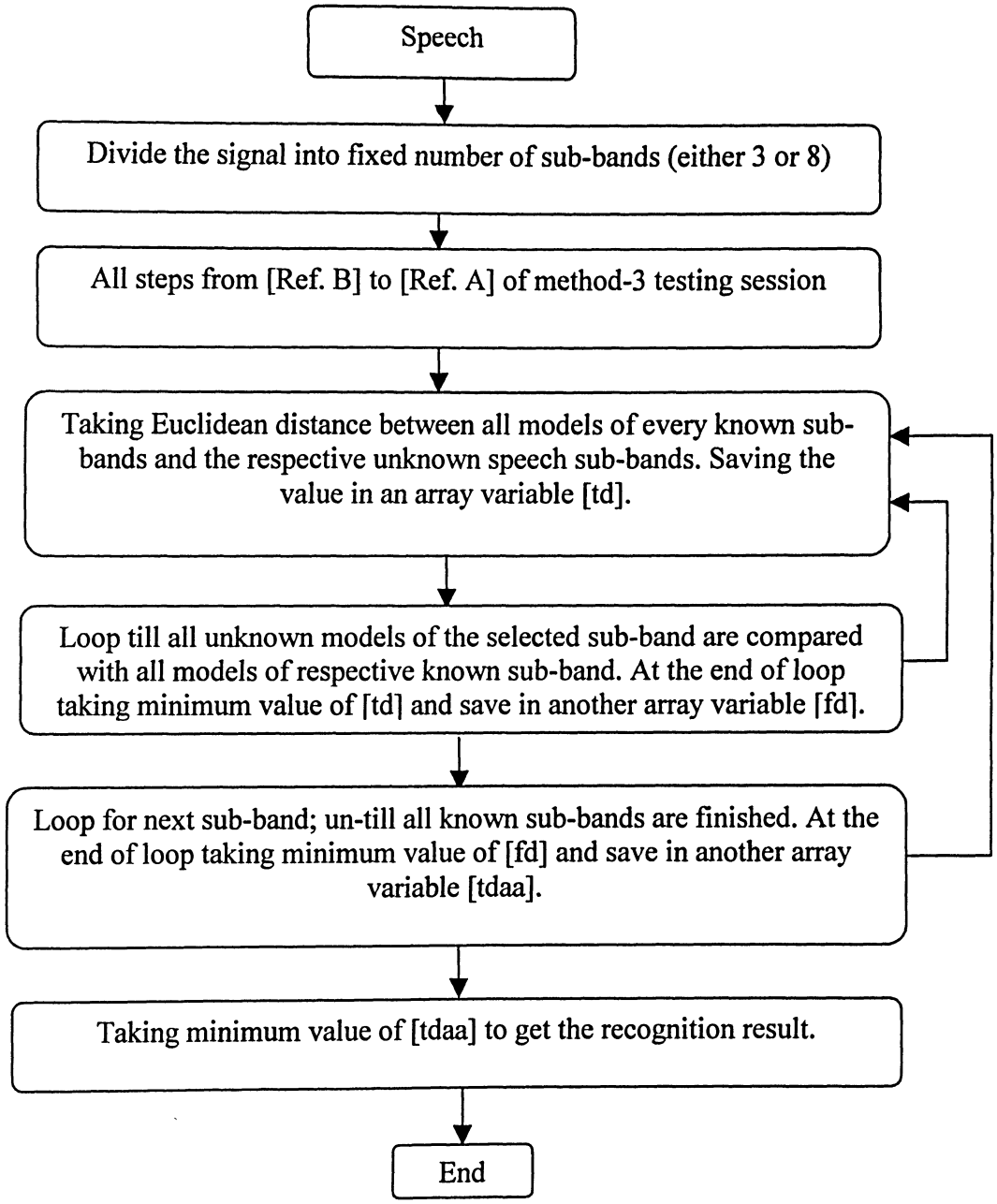


Figure 3.8: Method-4, Typical Testing Phase Flow Diagram

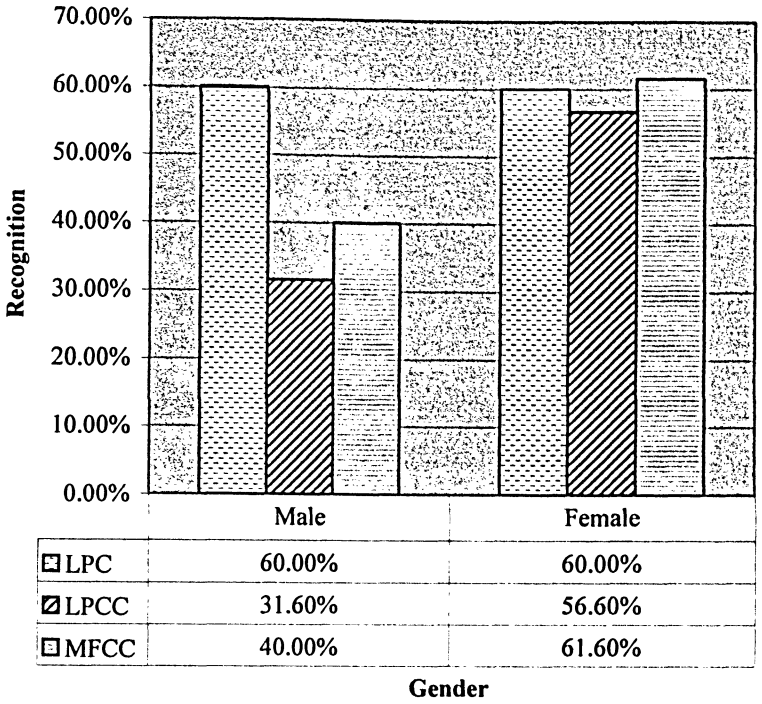


Table 3.5: Method-4 Results (3 Sub-bands Approach)

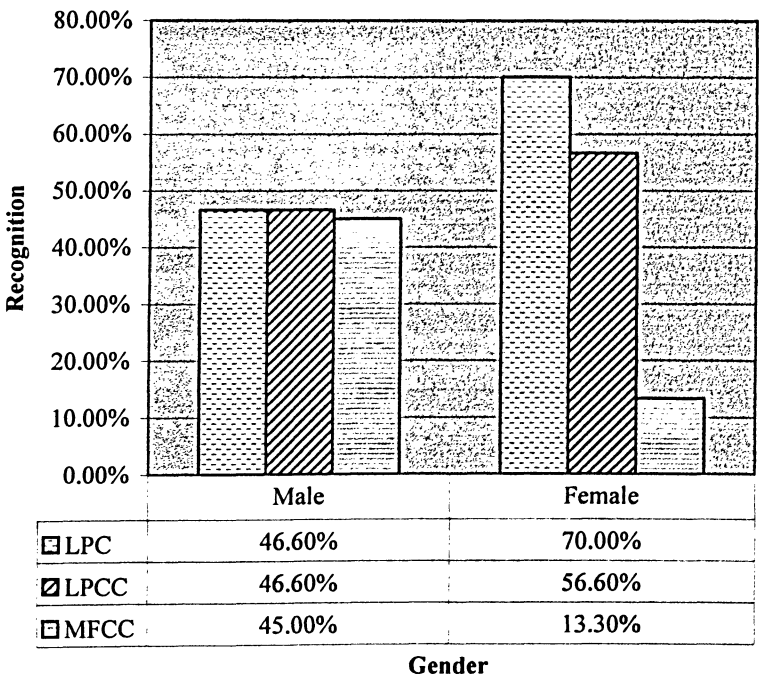


Table 3.6: Method-4 Results (8 Sub-bands Approach)

3.3 Summary

In this chapter we have presented our experiment of different methods and accuracy of recognition, which we have achieved with different feature extraction techniques in different methods. The total number of four speakers, two male and two female were involved in the training and testing session. One pair of male and female has recorded their samples with different microphone and other pair of speakers with different microphone. All samples have been collected in different period of time. We have recorded 30 speech samples of each speaker and the results in percentages have been indicated are out of 60 sample of male and 60 sample of female speakers.

We have implemented four different methods to extract feature vectors. In Method-1 we have used only non-zero values of speech sample to compute average feature vectors from 10 models.

In Method-2 we have implemented adaptive sub-band approach. In this method first we have divided the speech sample into fixed number of sub-band (either 3 or 8 sub-bands) and then computed optimal number of sub-band by applying weight factor.

In Method-3 we have implemented adaptive time segmentation approach in order to get optimal number of models. We have started with small interval of 25ms frame and combine them by applying weight factor.

While in last Method-4 we have joined frequency sub-band approach and Method-3 (adaptive time segmentation) together. We have divided first the speech sample into fixed number of sub-bands (either 3 or 8 sub-bands) and then applied adaptive time segmentation for every sub-band.

Chapter 4

Discussion and Conclusion

In this project we have implemented four different methods of signal processing for speech signal feature extraction. The tables/graphs presented in Chapter 3 and in this chapter are organized in two major categories in order to realize and compare the classification accuracy in speaker recognition in context of biometrics application. First category represents recognition accuracy results of every feature extraction technique with respect to the methods implemented and the second category provides the recognition accuracy of the four methods. The experimental results are summarized as follows:

- Table 3.1 represents results of our Method-1 applied for LPC, LPCC and MFCC feature extraction techniques. LPC feature extraction technique gave better results of 65% for male speakers in comparison to LPCC and MFCC. While LPC-derived cepstral coefficients feature extraction technique gave high accuracy of 91.6% for female speakers in comparison to LPC and MFCC feature extraction techniques. For female speakers, as expected LPCC results showed an improvement in results from LPC, but did not perform well with MFCC features. While for male

speakers showed unexpectedly poor results with LPCC and MFCC feature vectors.

- Table 3.2 represents results of Method-2 based on 3 sub-band adaptive band filtering approach applied for LPC, LPCC and MFCC feature extraction techniques and it can be realized that in this method LPC and LPCC both gave the same accuracy results of 66% for male speakers while MFCC gave better performance of 68.3% for female speakers as expected from the use of MFCC feature extraction technique. For male speakers MFCC did not provide better results as compared to LPC or LPCC. It is clear that in case of female speakers MFCC gave a significant improvement over LPC and LPCC accuracy results.

- Table 3.3 represents results of Method-2 with 8 sub-band adaptive band filtering approach applied for LPC, LPCC and MFCC feature extraction technique. It can be realized that this method performed well overall, and gave good results as compared to 3 sub-band approach. LPC provide 86.6% recognition accuracy for male speakers and 70% accuracy with LPC and LPCC for female speakers. LPCC provided a recognition accuracy of 83.3% and MFCC provided a recognition accuracy of 78.3% accuracy results for male speakers, which are not good improvement over LPC but still can be considered good result as compared to 3 sub-band approach for male speakers. While in the case of female speakers MFCC gave poor result of 51.6%.

- Table 3.4 represents results of Method-3 based on adaptive time segmentation applied to LPC, LPCC and MFCC feature extraction technique. LPC provided 55% accuracy result for male speakers while MFCC gave 63.3% recognition accuracy for female speaker. This indicates a significant improvement over LPC (33% recognition rate) and LPCC (46.6% recognition rate). In case of male speakers this method showed unexpected degradation for LPC from 55% recognition rate to LPCC with 48.3% recognition rate and 43.3% with MFCC.

- Table 3.5 represents results of Method-4 based on 3 fixed sub-bands with adaptive time segmentation approach applied for LPC, LPCC and MFCC feature extraction techniques. According to our results in this method, LPC performs well and gave 60% for male speakers while MFCC gave 61.6% for female speaker over 60% of LPC and 56.6% of LPCC. But LPCC for male speakers did not show any improvement over LPC.

- Table 3.6 represents results of Method-4 based on 8 fixed sub-bands in context of adaptive time segmentation approach applied for LPC, LPCC and MFCC feature extraction techniques. It can be realized that LPC, LPCC and MFCC provides almost the same result of 46.6% for male speakers while LPC of 70% results have been achieved for female speakers. It can be noted that the performance of LPCC of 56.6% and MFCC of 13.3% degraded from LPC of 70% results significantly for female speakers.

We would also like to present a quick overview of the performance of feature extraction techniques in our different methods. We have summarized our results in the Tables 4.1, 4.2 and 4.3.

- Table 4.1 represents the recognition accuracy result with respect to LPC feature extraction technique. It can be realized that Method-2 with 8 sub-bands adaptive band filtering approach has performed very well with 86.6% for male speaker recognition, while Method-1 has provided 88.3% accuracy for females speaker recognition. Where as Method-3 with 8 sub-band in context of adaptive time segmentation gave poor result of 55% for male speakers and 33.3% accuracy result for female speakers.

- Table 4.2 represents the recognition accuracy result with respect to the LPCC feature extraction technique. It can be realize that Method-2 with 8 sub-bands adaptive band filtering provided 83.3% result for male speakers while Method-1 provided 91.6% recognition accuracy for

female speakers. While Method-2 with 8 sub-bands adaptive band filtering gave higher results, but it is not giving any improvement for LPC results for male speakers. Method-4 with 3 sub-bands in context of adaptive time segmentation gave poor result for male speakers and it reduced to 50%.

- Table 4.3 represents the recognition accuracy result with respect to the MFCC feature extraction technique. From the Table it is clear that our Method-2 with 8 sub-bands adaptive band filtering gave very good results of 78.3% for male speakers while 88.3% accuracy results from Method-1 for female speakers. Method-3 gave the poorest results of 43.3%, which is also a significant degradation of Method-3 with respect to LPC and LPCC. Whereas Method-4 with 8 sub-bands in context of adaptive time segmentation showed unexpected poor results of 13.3% recognition rate as compared to LPC (70% recognition rate) and LPCC (56.6% recognition rate) for female speakers.

4.1 Conclusion

From the above discussion, we can see that we have achieved some results, which have provided very good performance results as per our expectation and as reported by different peoples in their research work. While at the same time we have seen some unexpected results, which we have got from different methods. In this situation it is hard to draw a solid line to conclude the success of different adaptive speech analysis methods, which we have implemented for feature extraction techniques. From above we have seen that all methods have some pros and cons of their applications in context with male and female speakers. The conclusions we can summarize are as follows:

- By using LPC as feature vectors for speaker recognition; our method-2 with 8 sub-bands adaptive band filtering approach gave good results for both male and female speakers, and as shown in Table 4.1, 86.6% for male speaker and 70% for female speaker recognition accuracy.

- In context of using LPC-derived cepstral coefficients (LPCC) as feature vectors for speaker recognition; again Method-2 with 8 sub-bands adaptive band filtering approach gave good results of 83.3% for male speakers and 70% of female speakers recognition accuracy.

- In last for MFCC as feature vectors for recognition again method-2 with 8 sub-bands adaptive band filtering gave 78.3% for male speakers and 60% for female speakers.

- From Table 4.1, 4.2 and 4.3 it can also be concluded that our method-2 with 3 sub-bands adaptive filtering and method-3 for MFCC results have shown improvement over LPC and LPCC for female speakers but non of them have shown better results for LPCC and MFCC over LPC for male speakers.

In this report several adaptive time segmentation, and sub-band filtering approaches, and their combinations were demonstrated for biometrics application. From the results obtained with 8 sub-bands adaptive band filtering it can be concluded that Method-2 has performed well and provided good recognition accuracy results for male speakers as well as female speakers in all feature extraction techniques. 8-band adaptive recombination of Method-2 also showed that there is a need to properly choose the number of frequency sub-band. While choosing the number of sub-bands to implement adaptive filtering, it is better to divide the speech signal into more number of sub-bands so that sufficiently reliable spectral information can be obtained in certain number of sub-bands. 3-band adaptive recombination approach has provided poor result as expected (due to not having much choice to obtain optimal number of sub-bands) as compared to 8-band approach and confirmed our objective to obtain optimal number of frequency sub-band of a speech signal. Method-2 results also conformed that adaptive recombination of frequency bands provides a better way to handle the non-stationary characteristics of speech signal.

From Table 4.1, 4.2, 4.3, it can be realized that the recognition accuracy for male and female speakers are not same. To achieve, even better recognition accuracy for female speakers with Method-2, following possible scenarios could be tested:

- Higher or lower order models other than 10^{th} order models can be used during training and testing sessions to test the recognition. The model order could be selected in such a way that it completely characterizes the pitch information. This area needs some investigation, and is a worthwhile step given the differences in recognition rates achieved between male and female speakers throughout the experimental studies conducted in this project.

- Different words other than word 'HELLO' such as: 'ZERO', 'ONE', 'TWO', 'THREE', or different small sentences can be used to evaluate the performance of the methods.

- The speech signal may be divided into more sub-bands in the span of 0-4kHz frequency bands and in less number of sub-bands in span of 4kHz-8kHz frequency bands. This allows for a finer analysis in the perceptual range, and may bring additional information.

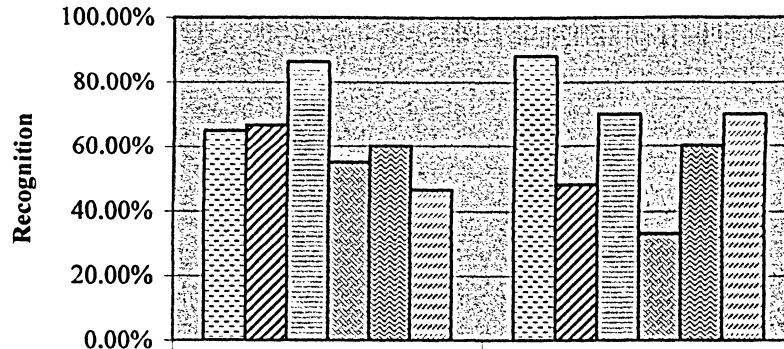
- Other feature matching technique such as: Vector Quantization or Hidden Markov Model can be applied to get better recognition accuracy.

There is no direct and fair comparison between our recognition accuracy results and other accuracy results as reported in different research literatures, because of different way to obtain feature extraction and feature matching techniques. Such as:

- We have used only Euclidean distance measure for feature matching in all our experiments, while most researchers have reported their accuracy success with vector quantization and Hidden Markov Model.

- We have used LPC derived cepstral coefficients to obtain MFCC, while in other reported work they have utilized filter banks technique to obtain MFCC feature vectors.

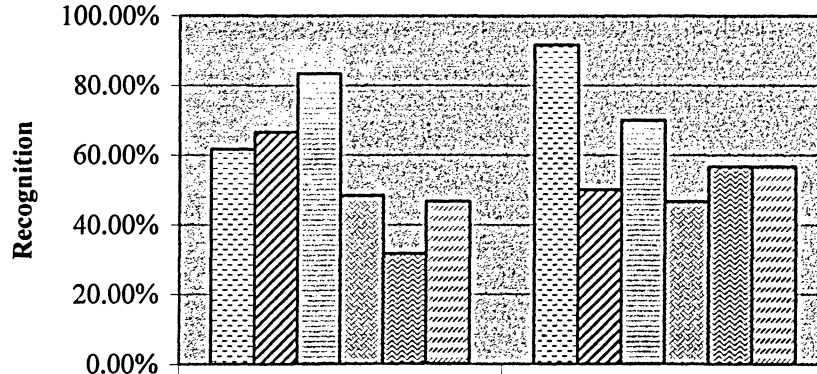
- We have trained and tested all methods with computer microphone as communication channel while most of the reported work has been done with telephone channel.



	Male	Female
M1	65.00%	88.30%
M2(3B)	66.60%	48.30%
M2(8B)	86.60%	70.00%
M3	55.00%	33.30%
M4(3B)	60.00%	60.00%
M4(8B)	46.60%	70.00%

Gender

Table 4.1: LPC Based Results



	Male	Female
M1	61.60%	91.60%
M2(3B)	66.60%	50.00%
M2(8B)	83.30%	70.00%
M3	48.30%	46.60%
M4(3B)	31.60%	56.60%
M4(8B)	46.60%	56.60%

Gender

Table 4.2: LPCC Based Results

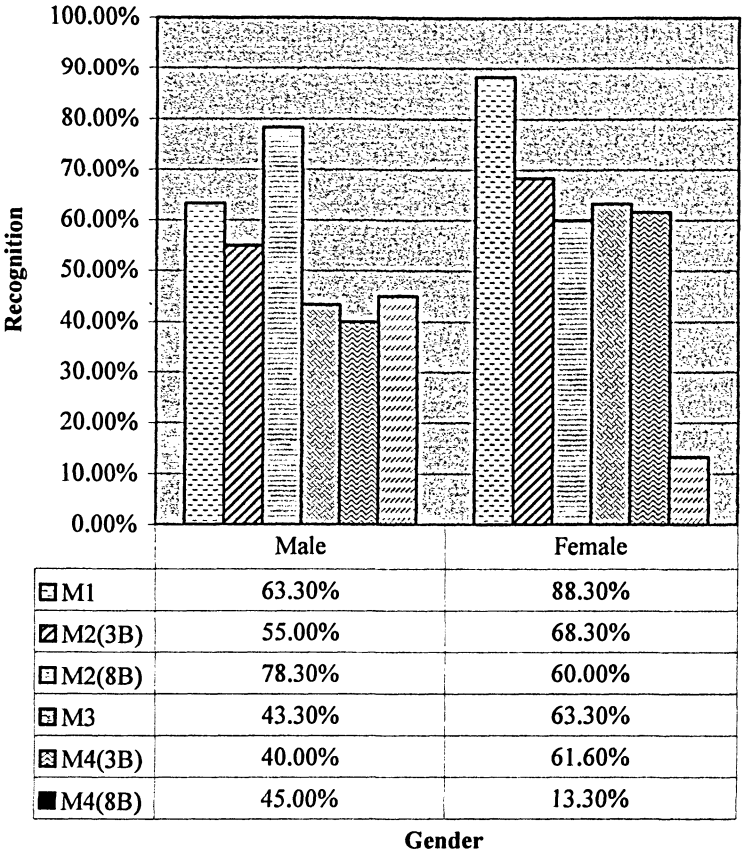


Table 4.3: MFCC Based Results

4.2 Summary

In this chapter we have presented our discussion on our experiments results. We have shown that which of our method gave a reasonable good recognition accuracy results. Though there are some pros and cons of using one of the methods we have implemented in context of LPC, LPCC and MFCC feature extraction techniques but we have seen that Method-2 (optimization of number of sub-bands models with the help of adaptive frequency sub-band filtering) has given good and reasonable speaker recognition accuracy result in comparison of other methods we have implemented and tested in this project.

Bibiography

[1] A. Rosenberg; "Automatic speaker verification: A review"; Proceedings of IEEE, Vol. 64, April 1976, pages: 475-487.

[2] G. R. Doddington; "Speaker recognition-identifying people by their voices"; Proceedings of IEEE, Vol. 73, Nov. 1985, pages: 1651-1664.

[3] Wong E., Sridharan S.; "Comparison of linear prediction cepstrum coefficients and Mel-frequency cepstrum coefficients for language identification";, 2001, Proceedings of 2001 International symposium on Intelligent multimedia video and speech signal processing, 2-4 May 2001, pages: 95-98.

[4] O'Shaughnessy D.; "Linear prediction coding"; Potentials, IEEE, Vol. 7, Issue 1, February 1998, pages: 29-32.

[5] Premakanthan P., Michael W. B.; "Speaker verification/recognition and the importance of selective feature extraction: Review"; Proceedings of the 44th IEEE 2001 Midwest symposium on Circuits and systems, 2001, vol. 1, 14-17 August 2001, pages: 57-61.

[6] Levinson S., Rabiner L., Sondhi M.; "Speaker independent isolated digit recognition using hidden markov models"; IEEE International conference on Acoustics, speech and signal processing, ICASSP' 83, vol. 8, April 1983, pages: 1049-1052.

[7] Kahn M., Grast P.; "The effects of five voice characteristics on LPC quality";, IEEE International conference on Acoustics, speech and signal processing, ICASSP' 83, vol. 8, April 1983, pages: 531-534.

[8] Campbell J. P. Jr.; "Speaker recognition: A tutorial"; Proceedings of IEEE, vol. 85 Issue: 9, Sept. 1997, pages: 1434-1462.

[9] Rabiner L. R., Schafer R. W.; "Digital processing of speech signals"; Prentice Hall signal processing series-1978.

[10] Imai S.; "Cepstral analysis synthesis on the Mel-frequency scale"; IEEE International conference on Acoustics, speech and signal processing, ICASSP' 83, vol. 8, April 1983, pages: 93-96.

[11] Furui S.; "Cepstral analysis technique for automatic speaker verification"; IEEE Transaction on Acoustic, speech and signal processing, vol. ASSP-29, 1981, pages: 254-272.

[12] Tohkura Y.; "A weighted cepstral distance measure for speech recognition"; IEEE Transaction on Acoustics, speech and signal processing, vol. ASSP-35 No. 10, October 1987, pages: 1414-1422.

[13] Boulard H., Dupont S.; "A new ASR approach based on independent processing and recombination of partial frequency bands"; Proceedings, fourth international conference on Spoken language, 1996, ICSLP-96 vol. 1, -3-6 October 1996, pages: 426-429.

[14] Tiberwala S, Hermansky H.; "Sub-band based recognition of noise speech"; IEEE International conference on Acoustics, speech and signal processing, vol. 2, 21-24 April 1997, pages: 1255-1258.

[15] Sivakumaran P., Ariyaeeinia A.M.; "The use of sub-band cepstrum in speaker verification"; Proceedings of IEEE international conference on Acoustics, speech and signal processing 2000, ICASSP'00. vol. 2, 5-9 June 2000, pages: II1073-II1076.

[16] Jialong He, Li Liu, Palm G.; "On the use of residual cepstrum in speech recognition"; Proceedings of IEEE international conference on Acoustics, speech and signal processing 1996, ICASSP'96. vol. 1, 7-10 May 1996, pages: 5-8.

[17] Gu L., Rose K.; "Perceptual harmonic cepstral coefficients for speech recognition in noisy environment"; Proceedings of IEEE international conference on Acoustics, speech and signal processing 2001, ICASSP'01. vol. 1, 7-11 May 2001, pages: 125-128.

[18] Assaleh, K.T., Mammone R.J.; "New LP-derived features for speaker identification"; IEEE Transaction on Speech and audio processing, vol. 2, Issue 4, October 1994, pages: 630-638.

[19] Assaleh, K.T., Mammone R.J.; "Robust cepstral features for speaker identification"; IEEE international conference on Acoustics, speech and signal processing 1994. ICASSP-94, vol. 1, 19-22 April 1994, pages: I/29-I/32.

[20] Mian G.; "Some factors influencing the performances of a speaker recognition system based on LPC"; IEEE international conference on Acoustics, speech and signal processing 1979. ICASSP-79, vol. 4, pages: 781-784.

[21] Tuzun O.B., Demirekler M., Nakiboglu K.B.; "Comparison of parametric and non-parametric representation of speech for recognition"; Proceedings of 7th Mediterranean Electrotechnical conference, April 1994, vol. 1, pages: 65-68.

[22] Chi Wei Che, Qiguang L., Dong-Suk Yuk; "An HMM approach to text-prompted speaker verification"; IEEE international conference on Acoustics, speech and signal processing 1996. ICASSP-96, vol. 2, 7-10 May, pages: 673-676.

[23] Gish H., Schmidt M.; "Text independent speaker identification"; IEEE, Signal processing magazine, vol. 11, Issue: 4, October 1994, pages: 18-32.

[24] Atal B.S.; "A model of LPC excitation in terms of eigenvectors of the autocorrelation matrix of the impulse response of the LPC filter"; IEEE international conference on Acoustics, speech and signal processing 1989. ICASSP-89, vol. 1, 23-26 May, pages: 45-48.

[25] Wohlford R., Wrench E.Jr., Landell B.; "A comparison of four techniques for automatic speaker recognition"; IEEE international conference on Acoustics, speech and signal processing 1980. ICASSP-80, vol. 5, April 1980, pages: 908-911.

[26] Matsumoto H., Moroto M.; "Evaluation of Mel-LPC cepstrum in a large vocabulary continuous speech recognition"; IEEE international conference on Acoustics, speech and signal processing 2001. ICASSP'01, vol. 1, May 2001, pages: 117-120.

[27] Donghoon Hyun, Chulhee Lee; "Optimization of Mel-cepstrum for speech recognition"; Conference proceedings of IEEE International conference on System, man and cybernetics, SMC'99, vol. 1, 12-15 October 1999, pages: 500-503.

[28] Chulhee Lee, Donghoon Hyun, Euisum Choi, Jinwook Go, Chungyong Lee; "Optimizing feature extraction for speech recognition"; IEEE Transaction on Speech and audio processing, vol. 11, Issue 1, January 2003, pages: 80-87.

[29] Hunt M.J., Lefebvre C.; "Speaker dependent and independent speech recognition experiments with an auditory model"; IEEE international conference on Acoustics, speech and signal processing 1998, ICASSP'98, vol. 1, 11-14 April 1998, pages: 215-218.

[30] Sanjit K.M.; "Digital Signal Processing Laboratory Using Matlab"; McGraw Hill –1999.

[31] "A practical guide to biometric security technology"; IT Professional, January/February 2001.

[32] B.S. Atal and S.L.Hanauer; "Speech analysis and synthesis by linear prediction of the speech wave."; J. Acoust. Soc. Vol. 50, 1971, pages: 637-655.

Appendix-A

Computational block diagram for recognition is shown in Figure A.1. The computational speed was computed on 350MHz Pentium-II processor with Matlab v5. The variables from a1 to a6 represent optimal number of models of one of the registered speaker, while variables from b1 to b6 represents optimal number of models for unknown speaker.

In the first step of computation, minimum Euclidean distance has been calculated between one speech model of known registered speaker and unknown speech models (such as: between a1 and b1, a1 and b2, a1 and b3 and so on) and saved in an array variable 'td'. The variables (c1, c2, c3, c4, c5, c6) in array 'td' hold the minimum distance values between a1-b1, a1-b2, a1-b3, a1-b4, a1-b5 and a1-b6 respectively. In the next step of computation, minimum value has been calculated among the variables in array 'td' and saved in another array variable 'fd'. The variables (d1, d2, d3, d4, d5, d6) hold the values for all known models comparison.

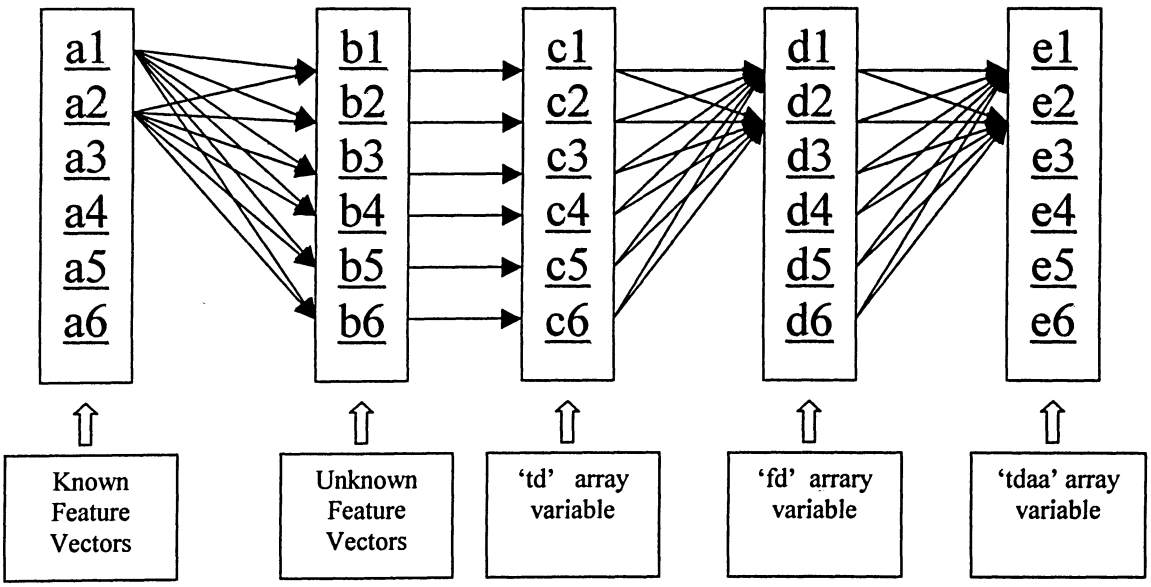


Figure A.1: Computational Algorithm Block Diagram for Recognition

Once all variables in array 'fd' have been calculated, the minimum value based on Euclidean distance has been computed among d1, d2, d3, d4, d5, d6 and saved in another array variable 'tdaa'. All the above steps of computation have been repeated for rest of the registered speakers. At the end of the computation all variables (e1, e2, e3, e4, e5, e6) of array 'tdaa' hold comparative values between all known and unknown speakers. To recognize the unknown speaker a minimum value has been calculated among variables of array 'tdaa'. The speaker having minimum value among all the registered speakers is recognized as the unknown speaker.

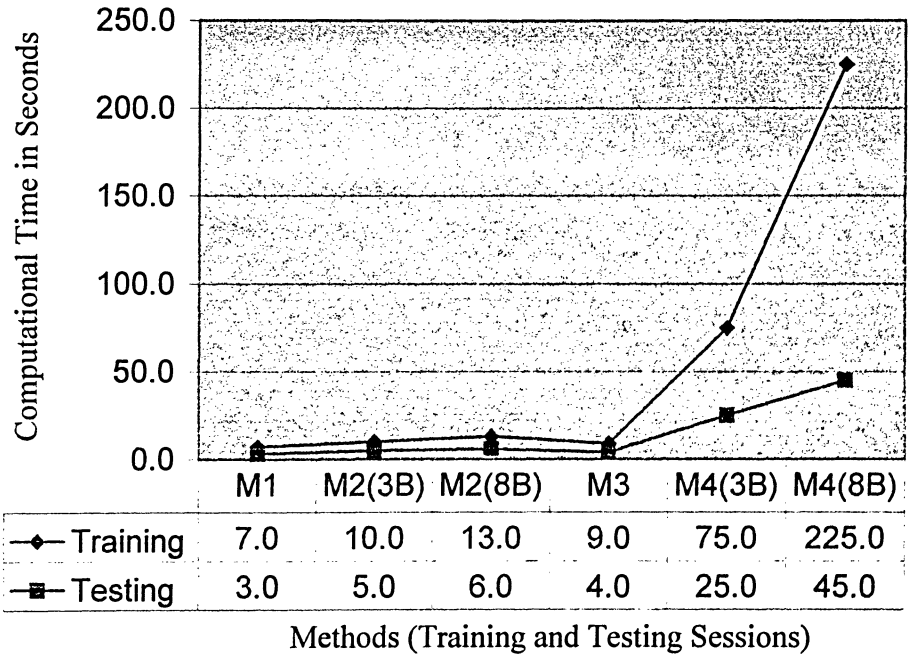


Table A.1: Computational Time in Training and Testing Sessions

From the Table A.1 it could be observed that only Method-4 is computationally expensive as compared to the other Methods. The computational complexity of Method-4 is applicable to both the training and the testing sessions. As seen in this report Method-2 has provided good recognition accuracy, and also consume less computational resources. The computational speed was computed on 350MHz Pentium-II processor with Matlab v5.