# Development Of An Automatic Speaker Recognition System Using Neural Networks

**NC Saminah Hanif,**

**PC Maliha Alam,**

**NC M. Ali**

**NC Zulfiqar Aijaz**

**TCC-11**

**DEVELOPMENT OF AN AUTOMATIC SPEAKER RECOGNITION SYSTEM USING NEURAL NETWORKS**

## DEDICATION

***WE DEDICATE OUR PROJECT TO ALL THE INNOCENT CHILDREN MARTYRED IN IRAQ WAR.***

## TABLE OF CONTENTS

## LIST OF FIGURES

**LIST OF TABLES**

**ACKNOWLEDGEMENTS**

Humblest gratitude to Allah Almighty - the All-knowing and the All-Powerful, the Creator, the Most Beneficent, Most Merciful. He is the Omni-Present and the Omni-Potent. Indeed, the working of the universe is nothing but a manifestation of the Great Powers He possesses. Without His consent not even a single breath could enter or leave our bodies, let alone the undertaking of work on this project. May He bestow us with His Guidance and make things clear for us when they get vague and confusing. Ameen.

Among His creations, we would like to thank our parents, who understood our concerns and spared us household chores, which would have definitely hindered the regular attention required for this work. This adds to the ever-growing list of favors that they have done for us. All this time their prayers have been invaluable, for which we can offer no compensation.

Our DS, Asst Prof (H) Aamir Masood Khan, is one person who requires special mention. The interest and eagerness exhibited by him to assist work in the pattern recognition domain has been phenomenal, which has been complemented by his excellent management and organizational skills. Also his patient , consistent and professional guidance helped us in achieving our project goals.

Finally , thanks to Military College of Signals (NUST) for providing us the opportunity to enhance our technical and practical skills.

# CHAPTER 1

## 1.1 An Overview to Pattern Recognition

***"To understand is to perceive patterns" - Isaiah Berlin***

One of the most impressive capabilities of the human brain is the ability to recognize patterns in nature. The brain is considered to be one of the most complex systems developed throughout the evolution of life in this planet. Complex systems are composed of many different subparts, which reveal emergent collective behavior that is not predicted by the analysis of its individual components. Systems presenting intelligent and creative behavior as well as systems that are able of performing pattern recognition are considered intrinsically complex.

A *pattern* can be stated as a distinct set of characteristics that can be differentiated and classified from a collection of information or data. ***Pattern Recognition*** *is the process whereby classification of objects (patterns) is done into a number of categories (classes)*. The purpose of a *pattern recognition system* is to make out, by looking at the observed information, the pattern class that would produce measurements analogous to the observed data. A block diagram showing the different stages involved in pattern recognition is shown below.

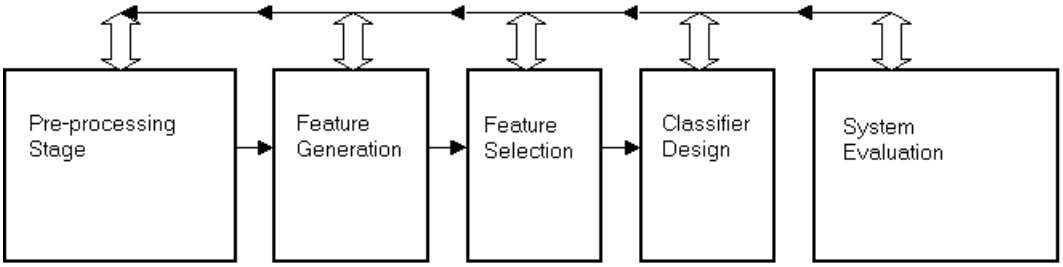



***Figure 1.1: Stages involved in a Pattern Recognition System***

The original signal is subjected to the *Pre-processing stage* to bring it in a form that facilitates the pattern recognition process. This usually involves cutting out certain portion of the signal that is not relevant for pattern recognition. In the speech signal this means cutting out the unvoiced part of the signal. De-noising and normalization are also applied as a part of preprocessing of the speech signal in order to bring the signal in a suitable form for the next stage, i.e. the Feature Generation stage.

***Feature generation*** is an essential stage whose purpose is to transform a given input sample space into a new “feature space” (which is usually of lower dimension) in order to obtain condensed or “abridged” information. In other words the information pertaining to classification is contained within a comparatively small number of features. In the speech signal this may be done by the application of transformations such as the Fourier or Wavelet Transform. This produces the FFT or DWT coefficients, which are then considered as features. As a consequence of this stage, we get a much better form of input data.

The requirement was to select the adequate Fast Fourier transform (FFT) and Discrete wavelet transform (DWT) coefficients that would yield the optimal results when subjected to the classification process. The coefficients to be selected may not be visible at first. So a sensible division of the coefficients can be made in the form of suitable ranges. Each of these ranges may then be separately tested and according to the results, only the selection of coefficients that give the best results are retained as features. Consequently, this is called the *feature selection* stage.

Systems designed to perform classification are termed *classifiers.* The classification can be *supervised* or *unsupervised* depending on whether a set of training data is available or not. Correct classification is dependent on the amount of differentiating info present in the measurements and it being put to efficient use in the preprocessing and feature generation stages.

Once the classifier design has been completed, the performance of the designed classifier needs to be assessed. For this purpose *the classification error rate* is normally determined. This is done at the system evaluation stage.
The various stages in the pattern recognition system are not independent. The feedback arrows in the figure amply suggest this fact. On the other hand, all these stages are interrelated. . One may want to revert to certain modifications in any of the stages involved, if the results are below expectation. This facility is provided so that the overall system performance can be enhanced and better results achieved.

## 1.2 Problem Definition

### 1.2.1 Speaker Recognition Systems

***Speaker Recognition*** is the process of automatically recognizing who is speaking on the basis of information obtained from speech waves. It is a generic term for the classification of a speaker's identity from an acoustic signal. This technique makes it possible to verify the identity of persons accessing systems, which is access control by voice, in various services. These services include voice dialing, banking transactions over a telephone network, telephone shopping, database access services, information and reservation services, voice mail, security control for confidential information areas, and remote access to computers. It is a technology that is moving towards making lives more convenient by introducing new, helpful services.

An ***Automatic Speaker Recognition System*** extracts, characterizes and recognizes the information in the speech signal conveying speaker identity. This system can be further divided into speaker identification and speaker verification.

***Speaker identification*** is the process of determining which of the registered speakers a given utterance comes. ***Speaker verification*** is the process of accepting or rejecting the identity claim of a speaker. Most of the applications in which voice is used as a key to confirm the identity claim of speaker are classified as speaker verification.

### 1.3 Motivation

ASV and ASI are probably the most natural and economical methods for solving the problems of unauthorized use of computer and communications systems and multilevel access control. With the ubiquitous telephone network and microphones bundled with computers, the cost of a speaker recognition system might only be for software.

Biometrics systems automatically recognize a person by using distinguishing trait. Speaker recognition is a performance Biometrics, i.e. a task is performed to be recognized. Speaker's voice, like other biometrics cannot be forgotten or misplaced, unlike knowledge based (e.g. passwords) or possession based (e.g. .key) access control methods. Speaker recognition systems can be made somewhat robust against noise and channel variations [1], [2], ordinary human changes (e.g., time of day voice changes and minor head colds), and mimicry by humans and tape recorders [3].

### 1.4 Project Goals

I. To build a simple, robust, complete and representative automatic Speaker Recognition System

II. To clearly investigate the nature of the speech signals collected from various speakers

III. To apply various DSP techniques and transforms for feature extraction that allows high degree of recognition and differentiation between for speakers.

IV. To apply Pattern Recognition methods for developing a classification system that can classify the correct speaker with minimum misclassification error

### 1.5 System Specifications

In order to measure the performance of the speaker recognition system developed by us, we used a database consisting of four speakers (2 males and 2 females). The speech samples collected were recorded in different environments and a large database of speech samples (total 300 samples for each speaker) was obtained to help the system become more robust. In order to build a complete, representative speaker recognition system, the performance was also checked on recorded telephone speech samples separately by collecting 130 samples from each speaker. These samples were subjected to preprocessing stage. Speech features were extracted from the preprocessed speech signal by using different DSP techniques. Appropriately selected features were classified on the basis of the system designed for classification. The system is evaluated by determining the classification error rate.

### 1.6 Organization of the report

Chapter 2 contains an overview of speech and speaker recognition. Chapter 3 explains the how features are extracted and selected. Chapter 4 deals with the classifier design and includes a comprehensive coverage of MLP neural networks using back-propagation Learning algorithm. Chapter 5 describes the results obtained and includes a discussion of these results. Also it concludes the report by discussing the future trends of the proposed system.

# CHAPTER 2

## 2.1 Speech and Speaker Recognition Systems

### 2.1.1 Speech

Speech can be described as the act of producing sound through the use of the vocal chords to create linguistic acts that communicate information from an initiator to a recipient. It has both expressive and receptive elements. The success of a speech act depends on numerous factors, including the presence or absence of a variety of speech disorders and the ability of the speaker to express the intended message.

### 2.1.2 Speech Processing

Speech processing is the study of speech signals and the processing methods of these signals. The signals are usually processed in a digital representation whereby speech processing can be seen as the intersection of digital signal processing and natural language processing. Speech processing can be divided in the following categories:

I. Speech Recognition, which deals with analysis of the linguistic content of a speech signal.

II. Speaker Recognition, where the aim is to recognize the identity of the speaker.

III. Enhancement of speech signals, e.g. Noise reduction,

IV. Speech Coding for compression and transmission of speech.

V. Voice Analysis for medical purposes, such as analysis of vocal loading and dysfunction of the vocal cords.

VI. The artificial synthesis of speech, which usually means computer generated speech.

Speech contains many characteristics that are specific to each individual. Many of these characteristics are independent of the linguistic message for an utterance. These, in speech recognition are generally considered a source of degradation. For instance, each utterance from an individual is produced by the same vocal tract, tends to have a typical pitch range (particularly for each gender), and has a characteristic articulator movement that is associated with dialect or gender. All these factors have a strong effect on the speech that is highly correlated with the particular individual who is speaking. For this reason, listeners are often able to recognize the speaker identity fairly quickly, even over the telephone. Artificial systems that recognize speakers rather than speech have been the subject of much research over the last 20 years, and commercial systems are already in use [4].

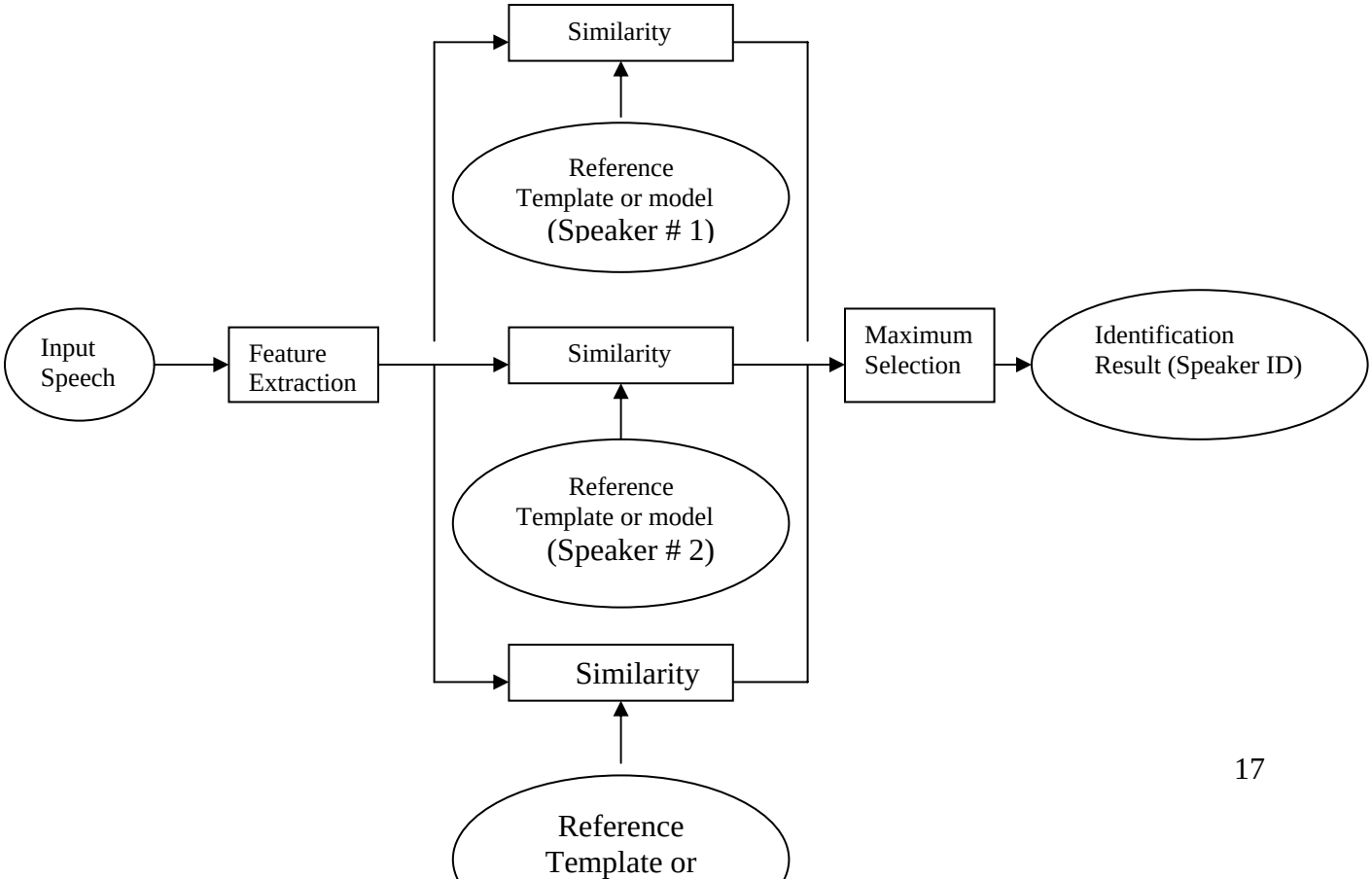
### 2.2 Principles of Speaker Recognition

Speaker recognition is the task of identifying a speaker by his or her voice as explained in the first chapter. Systems performing speaker recognition operate in different modes. A *closed set mode* is the situation of identifying particular speaker as one in a finite set of reference speakers [5]. In an *open set* system, a speaker is either identified as belonging to a finite set or is deemed not to be a member of the set [5].

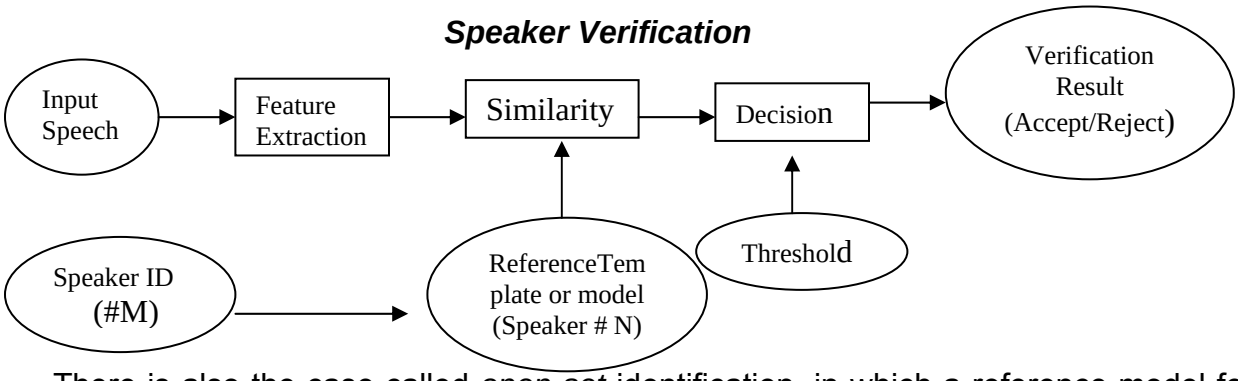
As speaker recognition encompasses speaker *identification* and speaker *verification*, in the case of speaker identification, the speaker is classified as being one of a finite set of speakers. As in the case of speech recognition, this will require the comparison of a speech utterance with a set of references for each potential speaker. For the case of speaker verification, the speaker is classified as having the purported identity or not. That is, the goal is to automatically accept or reject an identity that is claimed by the speaker. In this case the user will first identify himself/herself (e.g. by speaking a PIN code), and the distance between the associated reference and the pronounced utterance will be compared to a threshold that is determined during training.

Speaker identification and verification each require the calculation of a score reflecting the distance between an utterance and a set of references. The fundamental difference between speaker identification and speaker verification is the number of decision alternatives. In identification the number of decision alternatives is equal to the size of the population, while in speaker verification, performance approaches a constant independent of the size of the population unless the distribution of physical characteristics of speakers is extremely biased.

***Figure 2.1 shows the basic structures of speaker identification and verification systems.***



**Speaker Identification**




There is also the case called *open set* identification, in which a reference model for an unknown speaker may not exist. This is usually the case in forensic applications. In this situation, an additional decision alternative, *the unknown does not match any of the models*, is required. In both verification and identification processes, an additional threshold test can be used to determine if the match is close enough to accept the decision or if more speech data are needed.

Speaker recognition methods can also be divided into ***text-dependent*** and ***text-independent*** methods. The former require the speaker to say key words or sentences having the same text for both training and recognition trials, whereas the latter do not rely on a specific text being spoken. In a text dependent application, the recognition system has prior knowledge of the text to be spoken and it is expected that the user will cooperatively speak this text. The prior knowledge and constraint of the text can greatly boost performance of a recognition system. In a text independent application, there is no prior knowledge by the system of the text to be spoken, such as when using extemporaneous speech. Text dependent recognition is more difficult but more flexible.

As speaker and speech recognition system merge and speech recognition accuracy improves, the distinction between text dependent and text independent applications will decrease. Of the two basic tasks, text dependent speaker verification is currently the most commercially viable and useful technology although there has been much research conducted on both tasks [6]

Both text-dependent and independent methods share a problem however. These systems can be easily deceived because someone who plays back the recorded voice of a registered speaker saying the key words or sentences can be accepted as the registered speaker. To cope with this problem, there are methods in which a small set of words, such as digits, are used as key words and each user is prompted to utter a given sequence of key words that is randomly chosen every time the system is used. Yet even this method is not completely reliable, since it can be deceived with advanced electronic recording equipment that can reproduce key words in a requested order. Therefore, a text-prompted speaker recognition method has recently been proposed. [7]

### 2.3 Speaker Verification and identification Errors

#### 2.3.1 Sources of Verification Errors

Many factors can contribute to verification and identification errors. The sources of these errors include

I. Misspoken or misread prompted phrases

II. Extreme emotional states (e.g., stress or duress)

III. Time varying (intra or inter session) microphone placement.

IV. Poor or inconsistent room acoustics (e.g., multipath and noise)

V. Channel mismatch (e.g., using different microphones for enrollment and and verification)

VI. Sickness (e.g., head colds can alter the vocal tract)

VII. Aging (the vocal tract can drift away from models with age)

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.

### 2.3.2 Speaker Verification Errors

As the automatic speaker verification systems gain wide spread use, it is imperative to understand the errors made by these systems. There are two types of errors: the false acceptance of an invalid user (FA or TYPE 1) and the false rejection of a valid user (FR or TYPE 11). False acceptance errors are ultimate concern of high security speaker verification applications; however they can be traded off for false rejection errors.

## 2.4 Feature Parameters

Speaker identity is correlated with the physiological and behavioral characteristics of the speaker. These characteristics exist both in the spectral envelope (vocal tract characteristics) and in the supra-segmental features (voice source characteristics and dynamic features spanning several segments).

*Desirable* attributes of features for an automatic system are as under:

• Occur naturally and frequently in speech
• Easily measurable
• Not change over time or be affected by speaker's health
• Not be affected by reasonable background noise nor depend on specific transmission characteristics
• Not be subject to mimicry

Or in other words these are the requirements for a *practical*, *robust* and *secure* Automatic Speaker Recognition System. However it is also sensible to state that no

single feature has all these attributes. Features derived from spectrum of speech have proven to be the most effective in automatic systems.

Speaker Recognition requires the extraction of *speaker characteristic* features, which may or may not be independent of the particular words spoken. Since speech signal is a slowly time varying signal, when examined over a sufficiently short period of time. Over long periods of time, the signal characteristics 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 speaker recognition task, such as Linear Prediction Coding (LPC), Mel Frequency Cepstrum Coefficients (MFCC), and others.

### 2.5 Normalization

The most significant factor affecting automatic speaker recognition performance is variation in the signal characteristics from trial to trial (intersession variability and variability over time). Variations arise from the speaker themselves, from differences in recording and transmission conditions, and from background noise.

The basic reason for doing normalization is to ensure that every thing comes in the same range of values. This is related to the relative signal amplitude. Higher signal amplitude does not necessarily mean a high priority/value or high discriminatory information is present in that part of the signal.

Another important reason for normalization is the fact that speakers cannot repeat an utterance precisely the same way from trial to trial. It is well known that samples of the same utterance recorded in one session are much more highly correlated than samples recorded in separate sessions. There are also long-term changes in voices. It is important for speaker recognition systems to accommodate to these variations. Normalization is done to achieve exactly this.

### 2.6 Pattern Matching

The pattern-matching task of speaker verification system involves computing a match score, which is a measure of similarity of the input feature vector to some model.

Speaker models are constructed from the features extracted from the speech signal. To enroll users into the system a model of voice, based on extracted features, is generated and stored. Then to authenticate a user, the matching algorithm compares/scores the incoming speech signal with the model of claimed user.

There are two types of pattern matching models:

In *stochastic models*, the pattern matching is probabilistic and results in a measure of likelihood, or conditional probability, of observation given the model. For *template models* the pattern matching is deterministic. The observation is assumed to be an imperfect replica of the template, and the alignment of observed frames to template frames is selected to minimize a distance measured.

### 2.7 Text-Dependent Speaker Recognition Methods

Text-dependent methods are normally based on template-matching techniques. In this approach, the input utterance is represented by a sequence of feature vectors, generally short-term spectral feature vectors. The time axes of the input utterance and each reference template or reference model of the registered speakers are aligned using a dynamic time warping (DTW) algorithm and the degree of similarity between them, accumulated from the beginning to the end of the utterance, is calculated. This model compensates for speaking-rate variability

The hidden Markov model (HMM) can efficiently model statistical variation in spectral features. Therefore, HMM-based methods were introduced as extensions of the DTW-based methods, and have achieved significantly better recognition accuracies.[8]

### 2.8 Text-Independent Speaker Recognition Methods

One of the most successful text-independent recognition methods is based on vector quantization (VQ). In this method, VQ codebooks consisting of a small number of representative feature vectors are used as an efficient means of characterizing speaker-

specific features. A speaker-specific codebook is generated by clustering the training feature vectors of each speaker. In the recognition stage, an input utterance is vector-quantized using the codebook of each reference speaker and the VQ distortion accumulated over the entire input utterance is used to make the recognition decision.

Temporal variation in speech signal parameters over the long term can be represented by stochastic Markovian transitions between states. Speech segments are classified into one of the broad phonetic categories corresponding to the HMM states. After the classification, appropriate features are selected. In the training phase, reference templates are generated and verification thresholds are computed for each phonetic category. In the verification phase, after the phonetic categorization, a comparison with the reference template for each particular category provides a verification score for that category. The final verification score is a weighted linear combination of the scores from each category.

### 2.9 Text-Prompted Speaker Recognition Method

In the text-prompted speaker recognition method, the recognition system prompts each user with a new key sentence every time the system is used and accepts the input utterance only when it decides that it was the registered speaker who repeated the prompted sentence. The sentence can be displayed as characters or spoken by a synthesized voice. Because the vocabulary is unlimited, prospective impostors cannot know in advance what sentence will be requested. Not only can this method accurately recognize speakers, but it can also reject utterances, whose text differs from the prompted text, even if it is spoken by the registered speaker. A recorded voice can thus be correctly rejected.

### 2.10 Future Directions

Although many recent advances and successes in speaker recognition have been achieved, there are still many problems for which good solutions remain to be found. Most of these problems arise from variability, including speaker-generated variability and variability in channel and recording conditions. It is very important to investigate feature parameters that are stable over time, insensitive to the variation of speaking

manner, including the speaking rate and level, and robust against variations in voice quality due to causes such as voice disguise or colds. It is also important to develop a method to cope with the problem of distortion due to telephone sets and channels, and background and channel noises.

From the human-interface point of view, it is important to consider how the users should be prompted, and how recognition errors should be handled. Studies on ways to automatically extract the speech periods of each person separately from a dialogue involving more than two people have appeared as an extension of speaker recognition technology.

# CHAPTER 3

# Feature Extraction and Selection

## 3.1 Overview

The purpose of feature generation is to extract “relevant” information into a relatively very small number of features as compared to the length of original input sample vector To apply mathematical tools without loss of generality, the speech signal can be represented by a sequence of feature vectors. In this chapter the selection of appropriate features is discussed, along with methods to estimate them. This is known as feature selection and feature extraction.

Traditionally, pattern recognition paradigms are divided into three components: feature extraction and selection, pattern matching, and classification. Although this division is convenient from the perspective of designing system components, these components are not independent. The false demarcation among these components can lead to suboptimal designs because they all interact in real world systems. In speaker verification systems the goal is to design a system that minimizes the probability of

verification errors. Thus, the underlying objective is to discriminate between the given speaker and all others. A comprehensive review of state of art is given in [9]. Different transforms are usually used to extract features from a speech signal. The basic theme behind transform-based features is that an appropriately chosen transform can exploit and remove information redundancies, which usually exist in speech signals. Various examples of these transforms are Discrete Fourier Transform (DFT), Discrete Sine Transform (DST), Discrete Cosine Transform (DCT), Hadamard Transform, Haar Transform, Karhunen-Loeve (K-L) Transform and Discrete Wavelet transform (DWT). A comprehensive detail of each transform can be found in[10].

## 3.2 Fast Fourier Transform (FFT) Coefficients as Features

### 3.2.1 Introduction to Fourier Transform (FT) & Discrete Fourier Transform (DFT*)*

In the domain of signal processing, the Fourier Transform stands out as the most important signal-processing tool that has been in continuous and extensive use for a multitude of decades. It dates back to the nineteenth century when the French mathematician J.Fourier, showed that any signal can be expressed as an infinite sum of periodic complex exponential functions:

$$X(f) = \int_{-\infty}^{\infty} x(t).e^{-j2\pi ft}dt \qquad (3.1)$$

Equation (3.1) is called the Fourier transform of x [t].

This property of periodic functions was then generalized for non-periodic functions and then extended to periodic and non-periodic discrete time signals, thus allowing for the definition of the Discrete Fourier Transform (DFT):

$$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{(-j2\pi kn/N)}, \quad k=0,1,2.....N-1 \qquad (3.2)$$

Where x[n] is a discrete time signal and X[k] is the DFT of x[n].

It is evident that the DFT estimates the Fourier transform of a function using a finite number of its sampled points. This is important as it enables computers to be used for making calculations.

### 3.2.2 The Fast Fourier Transform Algorithms

Equation 3.2 gives the DFT of a discrete signal with N samples. Using this to compute the DFT requires $N^2$ complex additions and multiplication. A simple computation shows that the even frequency coefficients are the coefficients of the Fourier transform of the N/2 periodic signal:

$$f_p[n] = f[n] + f[n+N/2] \qquad (3.3)$$

And that the odd frequency coefficients are the coefficients of the Fourier transform of:

$$fi[n] = (f[n]-f[n+N/2])\ e\text{-}2i\pi n/N \qquad (3.4)$$

Induction verifies that the number of operations required by this method to compute the Fourier transform is of the order of KN $\log_2$ (N), where K is a constant which does not depend on N. Thus the number of computations have been significantly reduced. This is the basic principle of the *Fast Fourier Transform*.

## 3.3 Procedure and Results

### 3.3.1 Preprocessing

The normalized speech signal (sample) obtained for male and female speakers are shown in Figure3.1 and Figure 3.2.

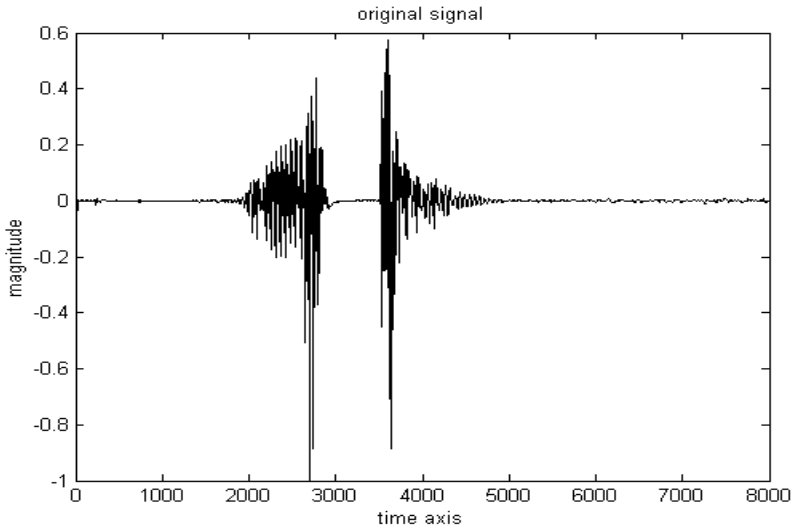



***Figure 3.1 Original Speech Sample (Male Speaker)***

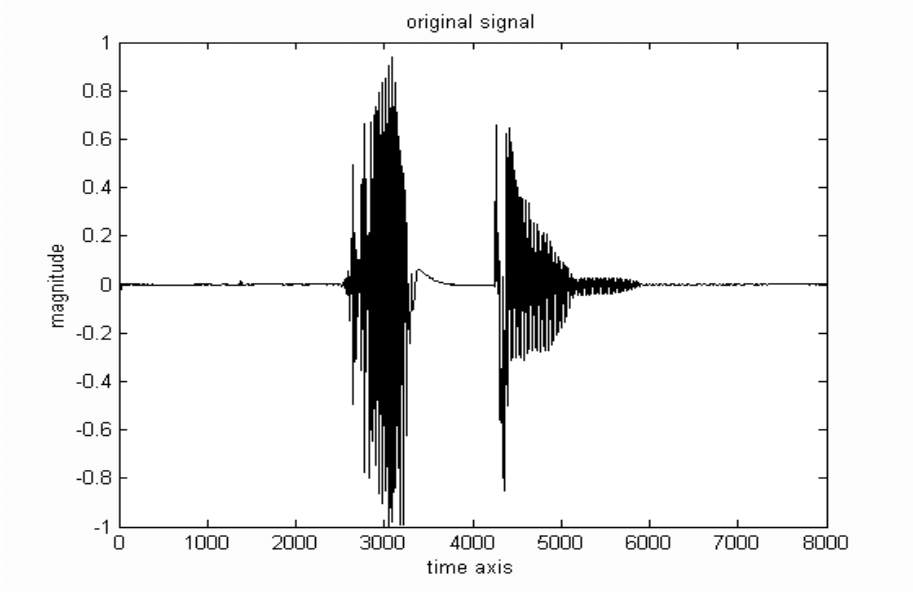



***Figure 3.2 Original Speech Sample (Female Speaker)***

It can be seen from these samples that almost all the speech samples of one person are almost identical but it is also observed that they are recorded differently on time axis. So to solve this problem all the samples are keenly observed and certain threshold

is applied to pick only the voiced part of the signal. The sample is initially 8000 points in case of isolated word and 16000 points if a sentence is spoken. The sampling frequency is 8kHz. After energy thresholding, the un-voiced part is removed from the speech sample, which results in the reduction of sample points to approximately 2500. This extracted voiced part is used for further processing and feature selection. The normalized speech samples of male and female speakers after the removal of un-voiced part by energy thresholding are shown in Figure 3.3 and Figure 3.4 respectively.

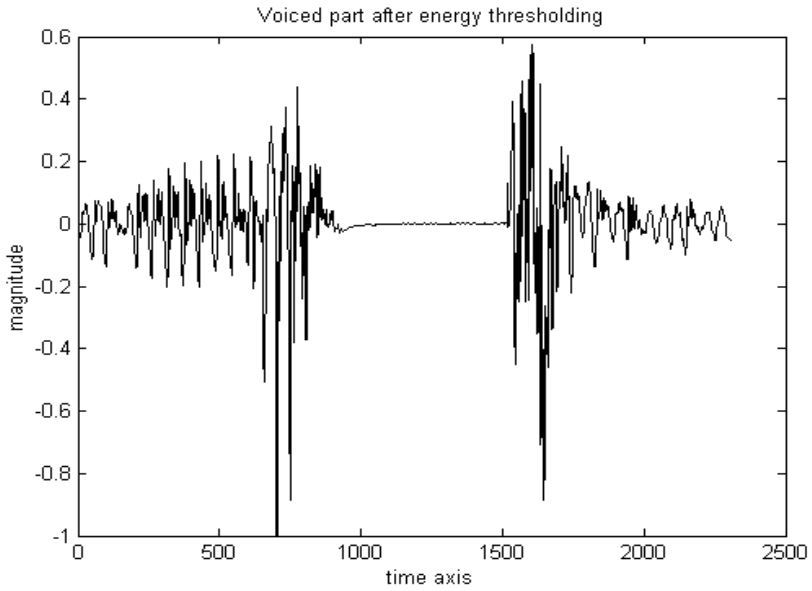



**Figure 3.3 Speech Sample after energy thresholding (Male Speaker)**

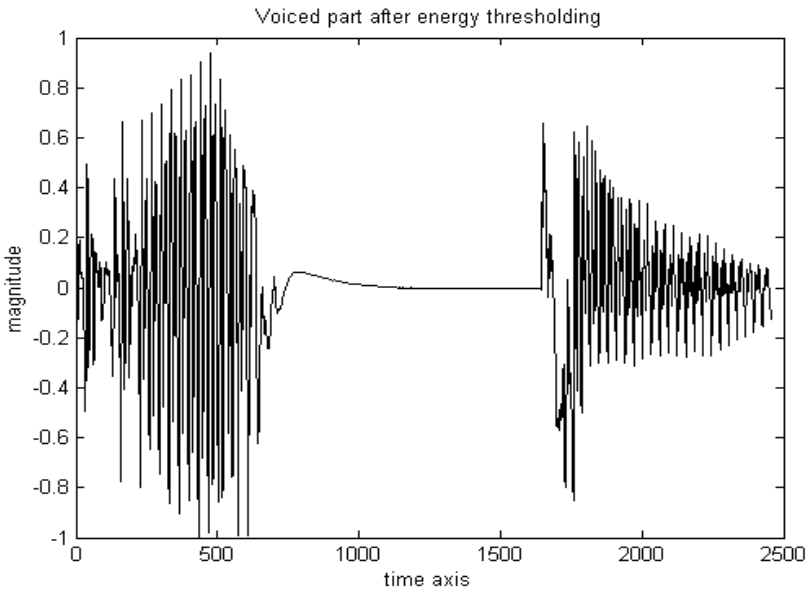



***Figure 3.4 Speech Sample after energy thresholding (Female Speaker)***

### 3.3.2 Feature Extraction

After obtaining the normalized voiced part of the speech sample, 256 point FFT was taken. Due to symmetry, a vector of 128 FFT coefficients is used as a feature vector. first 128 FFT coefficients of 256 point Fast Fourier Transform for a male and female signal after normalization are shown in Figure 3.5 and Figure 3.6 respectively

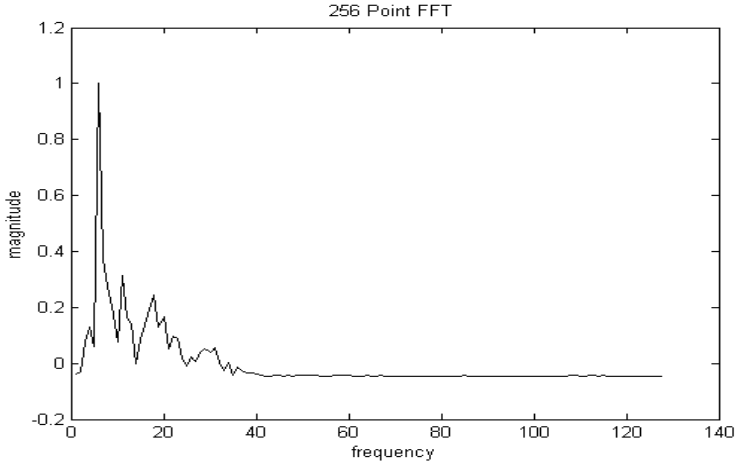



***Figure 3.5 first 128 symmetric points of 256 point FFT (Male Speaker)***

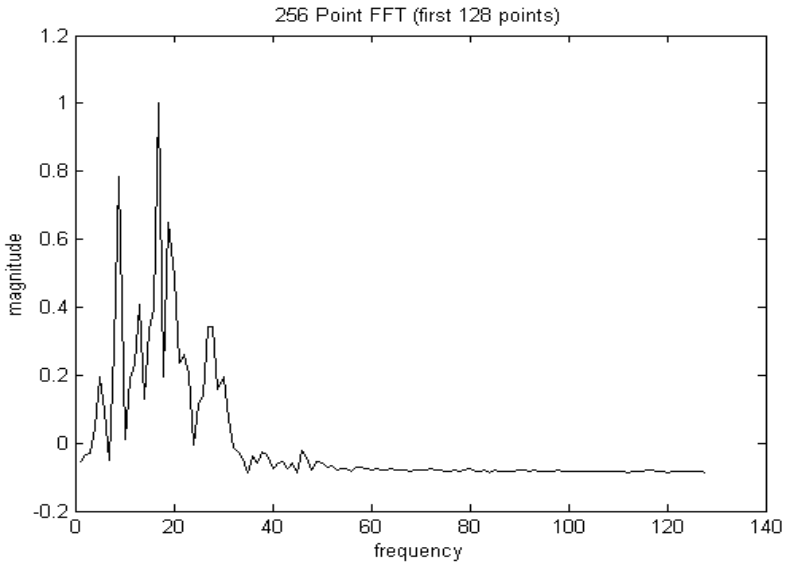



**Figure 3.6 first 128 symmetric points of 256 point FFT (Female Speaker)**

## 3.4 Discrete Wavelet Transform (DWT) Coefficients as Features

### 3.4.1 Introduction to Wavelet Theory

Wavelet theory provides a unified framework for a number of techniques, which had been developed independently for various signal-processing applications. For example multiresolution signal processing, used in computer vision, subband coding developed for speech and image compression and wavelet series expansion. The wavelet transform is of the interest for the analysis of *non-stationary* signals because it provides an alternative to Short Time Fourier Transform (STFT) or Gabor transform. The basic difference is that STFT uses single window analysis while WT uses short windows at high frequencies and long windows at low frequencies, instead of using a single analysis window as in the case of STFT.

The wavelet transform can be seen as signal decomposition onto a set of basis functions called *wavelets.* The wavelets are obtained from a single prototype mother wavelet by dilations and contractions as well as shifts *[11].* Wavelets may be viewed from two aspects, that is, “mathematical” and “algorithmic”. The mathematical aspect is covered first followed by the algorithmic implementation that describes an efficient coding scheme for obtaining the discrete wavelet transform.

### 3.4.2 Mathematical Background

An explanation of the Wavelet transform requires a description of the following terms.

1. Any function f(t) can be written as a linear combination of basis functions and the corresponding coefficients as follows:

$$f(t) = \sum_k \mu_k \phi_k(t) \tag{3.5}$$

Basis functions for the Fourier transform are the complex exponential (sines and cosines) functions.

2. The inner product of two functions is defined as follows:

$$< f(t), g(t) >= \int_a^b f(t).g^*(t)dt, \tag{3.6}$$

where f(t) and g(t) are two functions in $L^2$ [a,b] (that is, they are square integrable in the the interval [a,b]).

3. Two functions f and g are said to be orthogonal to each other if their inner product is zero.
That is:

$$< f(t), g(t) >= \int_a^b f(t).g^*(t)dt = 0 \tag{3.7}$$

4. A set of functions $\{\varphi_k(t)\}$, k = 1,2,3,.... is said to be orthonormal if

$$\int_a^b \phi_k(t)\phi_l^*(t)dt = 0, k \neq l \quad \text{(orthogonality condition)} \tag{3.8}$$

And

$$\int_a^b \{|\phi_k(t)|\}^2 dt = 0, \text{ or equivalently} \tag{3.9}$$

$$\int_a^b \phi_k(t)\phi_l^*(l)dt = \delta_{kl}, \tag{3.10}$$

where $\delta_{k,l}$ is the Kronecker delta function, defined as :

$$\delta_{k,l} = \begin{Bmatrix} 1, \text{ if } k = l \\ 0, \text{ if } k \neq l \end{Bmatrix} \tag{3.11}$$

The orthonormal basis functions are of special importance because the orthonormal bases allow computation of the analysis coefficients in a relatively simple way using the property of orthonormality.

For orthonormal bases, the coefficients $\mu_k$ can be calculated as

$$\mu_k =< f, \phi_k >= \int f(t).\phi_k^*(t)dt, \tag{3.12}$$

And the function f(t) can then be reconstructed by using (3.3) by substituting the $\mu_k$ coefficients. This yields

$$f(t) = \sum_k \mu_k \phi_k(t) = \sum_k < f, \phi_k > \phi_k(t) \tag{3.13}$$

### 3.4.3 The Continuous Wavelet Transform (CWT)

The continuous wavelet transform is defined as follows:

$$CWT \quad {}^{\psi}_{x}(\tau, s) = \Psi^{\psi}_{x}(\tau, s) = \frac{1}{\sqrt{|s|}} \int x(t) \psi^{*} \left( \frac{t-\tau}{s} \right) dt \tag{3.14}$$

Using the definition of the inner product stated earlier, the CWT can be envisaged as the inner product of the test signal with the basis functions $\psi_{\tau,s}(t)$:

$$CWT \quad {}^{\psi}_{x}(\tau, s) = \Psi^{\psi}_{x}(\tau, s) = \int x(t) \psi^{*}_{\tau,s}(t) dt, \tag{3.15}$$

Where

$$\psi_{\tau,s} = \frac{1}{\sqrt{|s|}} \psi \left( \frac{t-\tau}{s} \right) \tag{3.16}$$

The transformed signal is a function of two variables, τ and s, the translation and scale parameters, respectively. ψ is the transforming function (the mother wavelet). The term translation is related to the location of the window whereas the scale (defined as the reciprocal of frequency) corresponds to the compression or dilation of the signal. Low frequencies (high scales) indicate global information of a signal, while high frequencies (low scales) indicate detailed information of a hidden pattern in the signal. Also, larger scales correspond to dilated signals and small scales correspond to compressed signals. The mother wavelet is a prototype that is used for generating all the other windows. Depending upon the function being used, these windows can be shifted, compressed (or dilated) versions of the mother wavelet. The Morlet wavelet and Mexican hat function are two common examples of such functions, though various functions have been developed

### 3.4.4 The Wavelet Series

Wavelet theory provides a unified framework for a number of techniques, which had been developed independently for various signal-processing applications. For example multiresolution signal processing, used in computer vision, subband coding developed for speech and image compression and wavelet series expansion. The wavelet transform is of the interest for the analysis of *non-stationary* signals because it provides an alternative to Short Time Fourier Transform (STFT) or Gabor transform. The basic difference is that STFT uses single window analysis while WT uses short windows at high frequencies and long windows at low frequencies, instead of using a single analysis window as in the case of STFT.

The wavelet transform can be seen as signal decomposition onto a set of basis functions called *wavelets.* The wavelets are obtained from a single prototype mother wavelet by dilations and contractions as well as shifts *[7 sir].* Wavelets may be viewed from two aspects, that is, “mathematical” and “algorithmic”. The mathematical aspect is covered first followed by the algorithmic implementation that describes an efficient coding scheme for obtaining the discrete wavelet transform.

### 3.4.2 Mathematical Background

An explanation of the Wavelet transform requires a description of the following terms.

2. Any function f(t) can be written as a linear combination of basis functions and the corresponding coefficients as follows:

$$f(t) = \sum_{k} \mu_k \phi_k(t) \tag{3.5}$$

Basis functions for the Fourier transform are the complex exponential (sines and cosines) functions.

2. The inner product of two functions is defined as follows:

$$< f(t), g(t) >= \int_{a}^{b} f(t).g*(t)dt, \tag{3.6}$$

where f(t) and g(t) are two functions in $L^2$ [a,b] (that is, they are square integrable in the the interval [a,b]).

3. Two functions f and g are said to be orthogonal to each other if their inner product is zero.

That is:

$$< f(t), g(t) >= \int_a^b f(t).g^*(t)dt = 0 \quad (3.7)$$

4. A set of functions $\{\varphi_k(t)\}$, k = 1,2,3,.... is said to be orthonormal if

$$\int_a^b \phi_k(t)\phi_l^*(t)dt = 0, k \neq l \text{ (orthogonality condition)} \quad (3.8)$$

And

$$\int_a^b \{|\phi_k(t)|\}^2 dt = 0, \text{ or equivalently} \quad (3.9)$$

$$\int_a^b \phi_k(t)\phi_l^*(l)dt = \delta_{kl}, \quad (3.10)$$

where $\delta_{k,l}$ is the Kronecker delta function, defined as :

$$\delta_{k,l} = \begin{Bmatrix} 1, & \text{if } k = l \\ 0, & \text{if } k \neq l \end{Bmatrix} \quad (3.11)$$

The orthonormal basis functions are of special importance because the orthonormal bases allow computation of the analysis coefficients in a relatively simple way using the property of orthonormality.

For orthonormal bases, the coefficients $\mu_k$ can be calculated as

$$\mu_k =< f, \phi_k >= \int f(t).\phi_k^*(t)dt, \tag{3.12}$$

And the function f(t) can then be reconstructed by using (3.3) by substituting the $\mu_k$ coefficients. This yields

$$f(t) = \sum_k \mu_k \phi_k(t) = \sum_k < f, \phi_k > \phi_k(t) \tag{3.13}$$

### 3.4.3 The Continuous Wavelet Transform (CWT)

The continuous wavelet transform is defined as follows:

$$CWT \quad {}^{\psi}_{x}(\tau, s) = \Psi^{\psi}_{x}(\tau, s) = \frac{1}{\sqrt{|s|}} \int x(t) \psi^* \left( \frac{t-\tau}{s} \right) .dt \tag{3.14}$$

Using the definition of the inner product stated earlier, the CWT can be envisaged as the inner product of the test signal with the basis functions $\psi_{\tau,s}(t)$:

$$CWT \quad {}^{\psi}_{x}(\tau, s) = \Psi^{\psi}_{x}(\tau, s) = \int x(t) \psi^*_{\tau,s}(t)dt, \tag{3.15}$$

Where

$$\psi_{\tau,s} = \frac{1}{\sqrt{|s|}} \psi \left( \frac{t-\tau}{s} \right) \tag{3.16}$$

The transformed signal is a function of two variables, τ and s, the translation and scale parameters, respectively. ψ is the transforming function (the mother wavelet). The term translation is related to the location of the window whereas the scale (defined as the reciprocal of frequency) corresponds to the compression or dilation of the signal. Low frequencies (high scales) indicate global information of a signal, while high frequencies (low scales) indicate detailed information of a hidden pattern in the signal. Also, larger

scales correspond to dilated signals and small scales correspond to compressed signals. The mother wavelet is a prototype that is used for generating all the other windows. Depending upon the function being used, these windows can be shifted, compressed (or dilated) versions of the mother wavelet. The Morlet wavelet and Mexican hat function are two common examples of such functions, though various functions have been developed.

### 3.4.4 The Wavelet Series

It is evident that neither the FT, nor the STFT, nor the CWT can be practically computed on computers by using analytical equations, integrals, etc. It is, therefore, necessary to discretize the transforms by sampling the time-frequency (scale) plane. In the case of WT, the scale change can be used to reduce the sampling rate (according to Nyquist's rule), which will save a considerable amount of computation time.Therefore if the time-

scale plane needs to be sampled with a sampling rate of $N_1$ at scale $s_1$, the same plane can be sampled with a sampling rate of $N_2$, at scale $s_2$, where $s_1 < s_2$ (corresponding to frequencies $f_1 > f_2$) and $N_2 < N_1$. The actual relationship between $N_1$ and $N_2$ is

$$N_2 = \frac{s_1}{s_2} N_1, \text{ or} \tag{3.17}$$

$$N_2 = \frac{f_2}{f_1} N_1 \tag{3.18}$$

The discretization can be done without any restriction as far as the analysis of the signal is concerned. If synthesis is not required, even the Nyquist criteria do not need to be satisfied. The restrictions on the discretization and the sampling rate become important if the signal reconstruction is required. The signal can be reconstructed if the time and scale parameters are discretized under certain conditions. The scale parameter s is discretized first on a logarithmic grid. The time parameter is then discretized with respect to the scale parameter, that is, a different sampling rate is used for every scale.

The discretization procedure is mathematically expressed as follows:

The scale discretization is $s = s_o^j$, and translation discretization is $\tau = ks_o^j\tau_o$, where $s_o > 1$ and $\tau_o > 0$. It is important to note that the translation is dependent on scale discretization with $s_o$.

The continuous wavelet transform was defined as:

$$\psi_{\tau,s} = \frac{1}{\sqrt{s}}\psi\left(\frac{t-\tau}{s}\right) \tag{3.19}$$

By inserting $s = s_o^j$, and $\tau = ks_o^j\tau_o$, we get

$$\psi_{j,k}(t) = s_o^{-j/2}\psi(s_o^{-j}t - k\tau_o) \tag{3.20}$$

If $\psi_{j,k}$ constitutes an orthonormal basis, the wavelet series transform becomes

$$\Psi_k^{\psi\, j,k} = \int x(t)\psi^*_{j,k}(t)dt, \text{ or} \tag{3.21}$$

$$x(t) = c_\psi \sum_j \sum_k \Psi_x^{\psi\, j,k}\psi_{j,k}(t) \tag{3.22}$$

A wavelet series requires that $\psi_{j,k}$ are either orthonormal, biorthogonal, or frame.

If $\psi_{j,k}$ are not orthonormal, equation 3.21 becomes

$$\Psi_k^{\psi\, j,k} = \int x(t)\psi^{*\wedge}_{j,k}(t)dt \tag{3.23}$$

where $\psi^{*\wedge}_{j,k}(t)$, is either the dual biorthogonal basis or dual frame.

If $\psi_{j,k}$ are orthonormal or biorthogonal, the transform will be non - redundant, whereas if they form a frame, the transform will be redundant. On the other hand, it is much easier to find frames than it is to find orthonormal or biorthogonal bases.[12]

.

### 3.4.5 The Discrete Wavelet Transform

The wavelet series is basically a sampled version of the CWT, and the information it provides is highly redundant as far as the reconstruction of the signal is concerned. This redundancy requires a significant amount of computation time and resources. The discrete wavelet transform (DWT), on the other hand, provides sufficient information both for analysis and synthesis of the original signal, with a significant reduction in computation time. The DWT is considerably easier to implement when compared to the CWT and various algorithms exist that are used for its computation. The Subband Coding scheme for implementation of DWT is explained below.

### 3.4.6 The Subband Coding and Multiresolution Analysis

In case of DWT, a time-scale representation of a digital signal is obtained using digital filtering techniques. The DWT analyzes the signal at different frequency bands with different resolutions by decomposing the signal into a coarse approximation and detail information. DWT employs two sets of functions, called scaling functions and wavelet functions, which are associated with low pass and high pass filters, respectively. The decomposition of the signal into different frequency bands is simply obtained by successive high pass and low pass filtering of the time domain signal. The original signal (say x [n] ) is first passed through a half-band high pass filter g[n] and a low pass filter h[n]. After the filtering, half of the samples can be eliminated according to Nyquist's rule, since the signal now has a highest frequency of $\pi/2$ radians instead of $\pi$. The signal can therefore be subsampled by two, simply by discarding every other sample. This constitutes one level of decomposition and can mathematically be expressed as:

$$y_{low}[k] = \sum_{n} x[n].h[2k - n] \tag{3.25}$$

where $y_{high}[k]$ and $y_{low}[k]$ are the outputs of the high pass and low pass filters, respectively, after subsampling by two.Although it is not the only choice, DWT coefficients are usually sampled from the CWT on a dyadic grid, that is, $s_o = 2$ and $\tau_o = 1$, yielding $s = 2^j$ and $\tau = k.2^j$, as explained earlier.

Subband coding starts with passing this signal through a half band digital lowpass filter with impulse response h[n]. This filtering is mathematically expressed as a convolution sum as follows:

$$x[n] * h[n] = \sum_{k=-\infty}^{\infty} x[k].h[n - k] \tag{3.26}$$

This decomposition halves the time resolution and doubles the frequency resolution. The above procedure can be repeated for further decomposition. At every level, the filtering and subsampling will result in half the time resolution and double the frequency resolution. Figure 3.7 illustrates this procedure, where x [n] is the original signal to be decomposed, and h [n] and g [n] are the low pass and high pass filters, respectively.

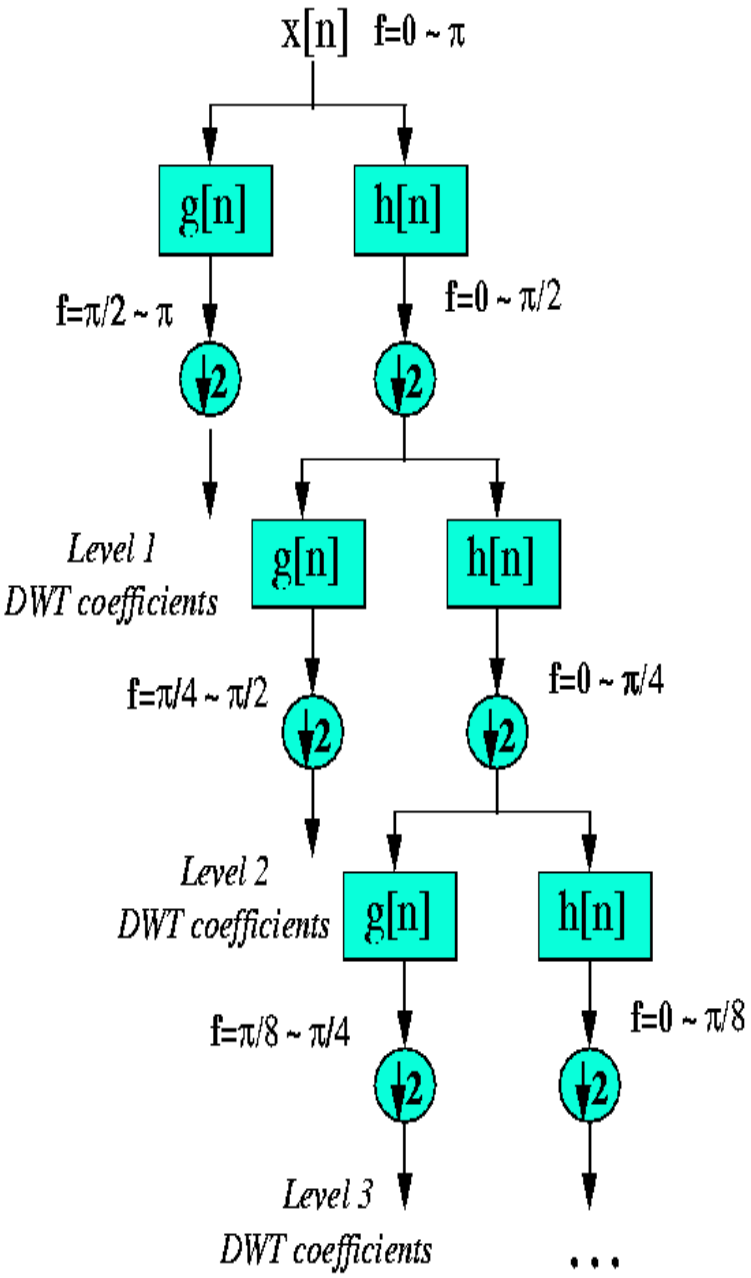



**Figure 3.7 Subband Coding and The Multiresolution Analysis**

One important property of the discrete wavelet transform is the relationship between the impulse responses of the high pass and low pass filters. The high pass and low pass filters are related by:

$$g[L-1-n] = (-1)^n .h[n] \,, \tag{3.27}$$

Where g[n] is the high pass filter, h[n] is the low pass filter, and L is the filter length. The two filters are odd index alternated reversed versions of each other. Low pass to high pass conversion provided by the $(-1)^n$ term.Filters satisfying this condition are known as the Quadrature Mirror Filters (QMF). The two filtering and subsampling operations can be expressed by the following two equations:

$$y_{high}[k] = \sum_n x[n].g[-n+2k] \tag{3.28}$$

$$y_{low}[k] = \sum_n x[n].h[-n+2k] \tag{3.29}$$

The reconstruction in this case is very easy, since halfband filters form orthonormal bases. The above procedure is followed in reverse order for signal reconstruction. The signals at every level are upsampled by two, passed through the synthesis filters g'[n], and h'[n] ( highpass and lowpass, respectively ), and then added. The analysis and synthesis filters are identical to each other, except for a time reversal. Therefore, the reconstruction formula becomes for (each layer)

$$x[n] = \sum_{k=-\infty}^{\infty} (y_{high}[k].g[-n+2k]) + (y_{low}[k].h[-n+2k]) \tag{3.30}$$

However if the filters are not ideal halfband, then perfect reconstruction cannot be achieved. Although it is not possible to realize ideal filters, under certain conditions it is possible to find filters that provide perfect reconstruction. The most well known wavelets are the ones developed by Ingrid Daubechies, and they are known as Daubechies' wavelets[12]

## 3.5 Procedure and Results

### 3.5.1 Preprocessing

The selected feature vectors after removal of unvoiced part through energy thresholding is approximately of 2500 points in length. This preprocessed speech sample was time

aligned since we know that DWT coefficients depend on timing information of the input signal as compared to the Fast Fourier Transform. The original speech samples of male and female speakers and those obtained after removal of unvoiced part are shown in Figure 3.8, 3.9, 3.10 and 3.11.

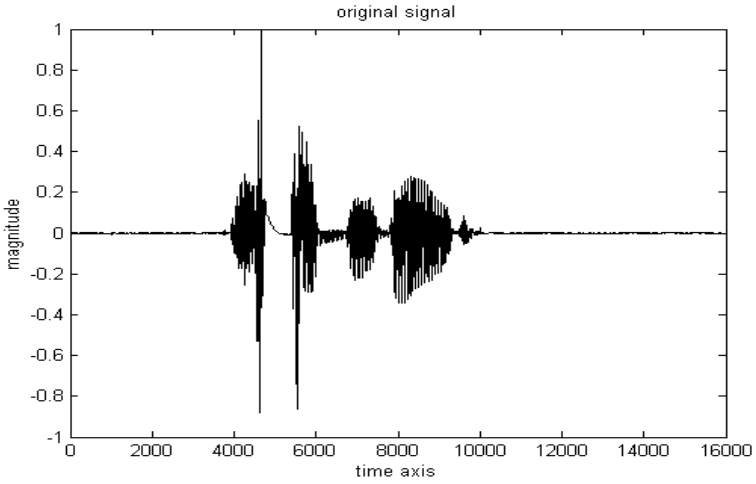



**Figure 3.8 Original Speech sample (Male Speaker)**

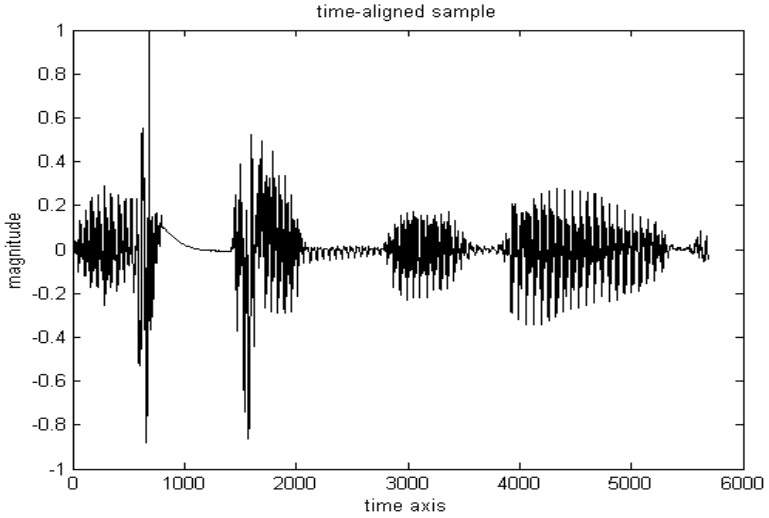



**Figure 3.9 Time-Aligned Speech Sample (Male Speaker)**

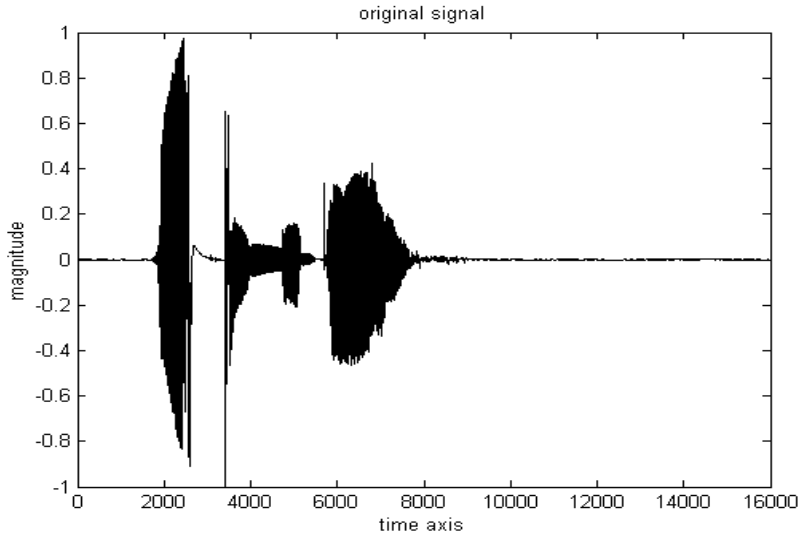



**Figure 3.10 Original Speech Sample (Female Speaker)**

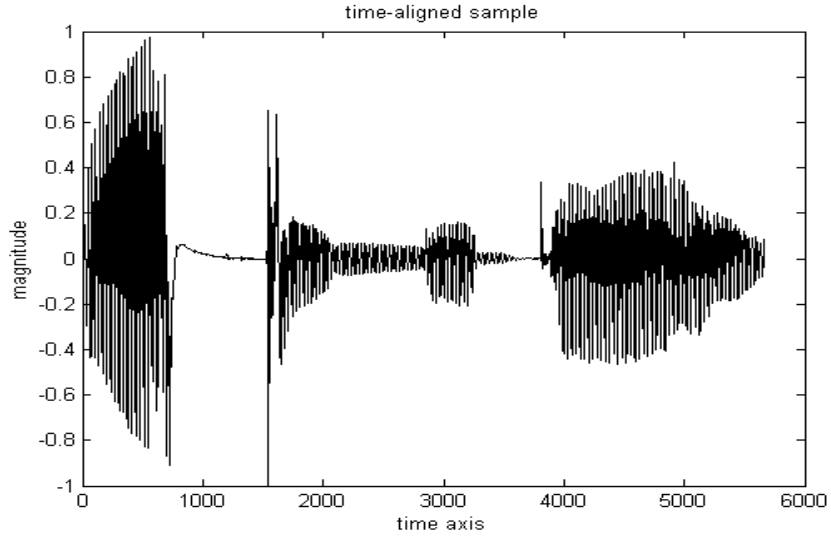


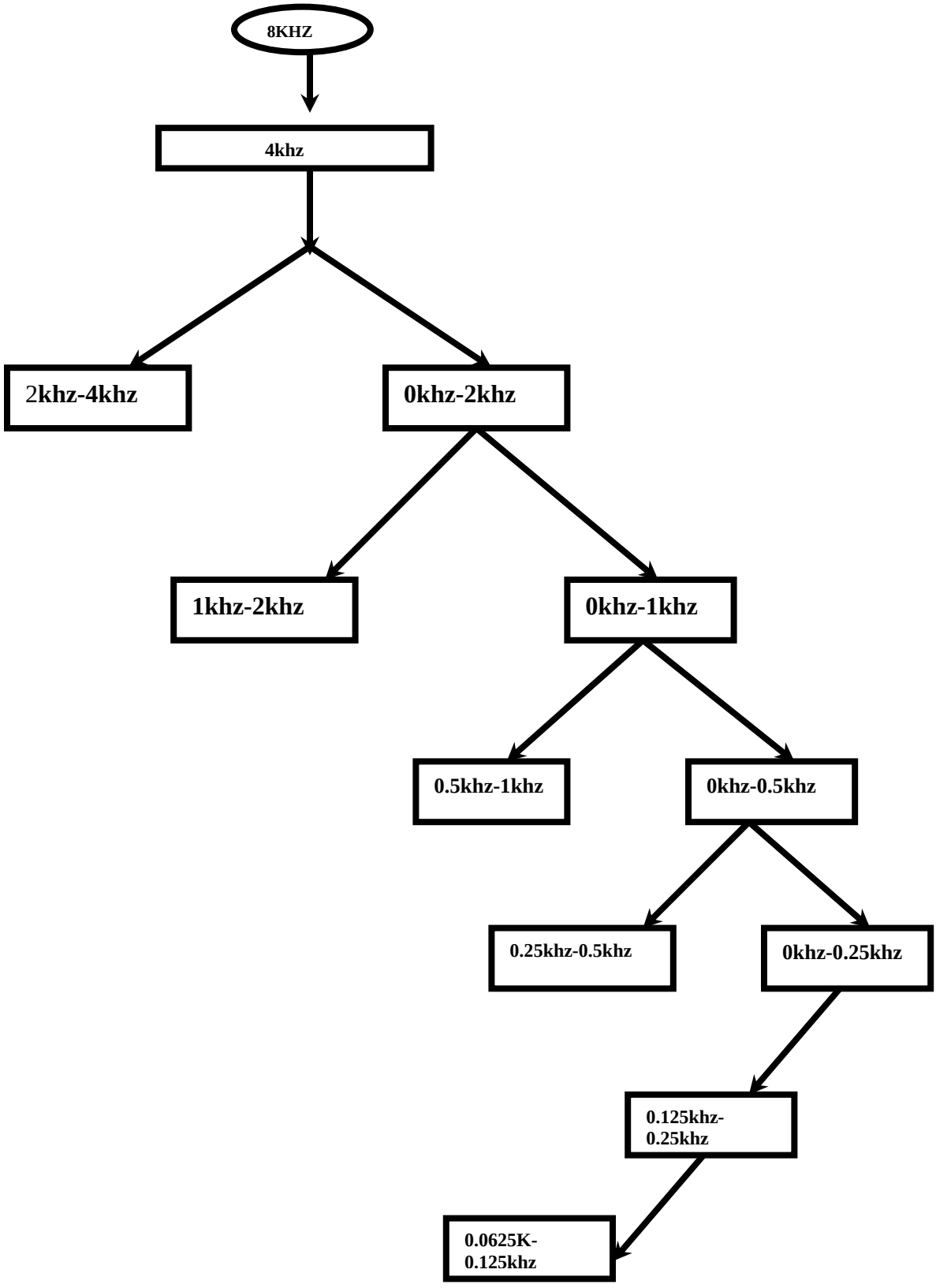

**Figure 3.11 Time Aligned Speech Sample (Female Speaker)**

After timealignment, normalization and zeropadding, the highest energy region of 512 points is selected for feature extraction.

### 3.5.2 Feature Extraction

After the preprocessing, the time aligned speech signal was applied to the subband-coding algorithm for obtaining the discrete wavelet coefficients as feature vectors for the classifier stage. Since the data length is 512, eight levels of decomposition were taken of each of the 256-point data length separately giving detailed and approximate

coefficients. The subband decomposition of voice frequency band is shown in Figure 3.8.



The length of the DWT coefficient for each 256-length region was 306 points as shown in Figure 3.9 and Figure 3.10 for male and female speaker respectively. Daubechies wavelet was used as the mother wavelet for the decomposition .As the frequency range of the speech signal is 0-4 kHz, the frequency bands in the range 0.125 –1 kHz were used which could prove to be most convenient for recognition. After careful consideration of the output signal the most relevant band chosen was in the frequency range of 0.125-0.25 kHz and 0.5 – 1 kHz. The coefficients in this frequency range were finally selected as features. The 54 point DWT coefficients for each of 256 length were combined to give the complete feature vector of 108 points. The selected features are an input to the Multiplayer Perceptron (MLP)

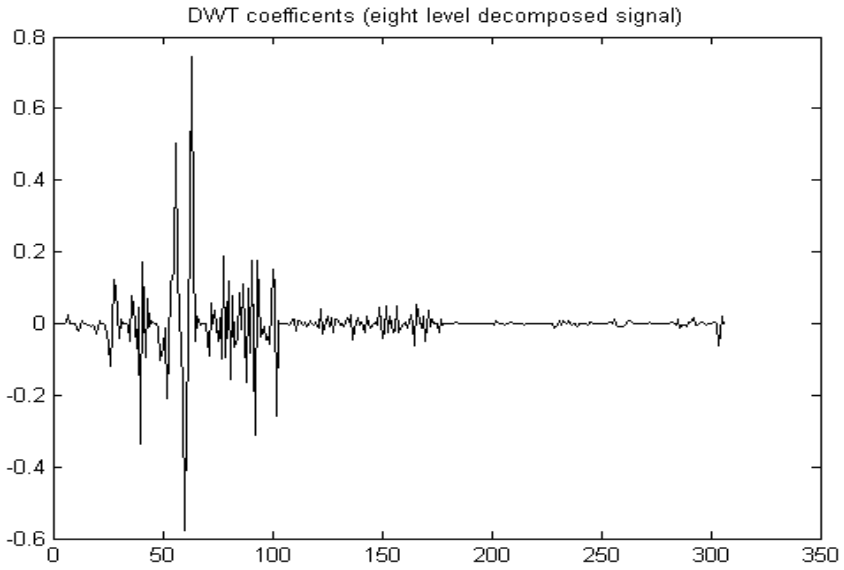



**Figure 3.9 8 level DWT Coefficients (Male Speaker)**

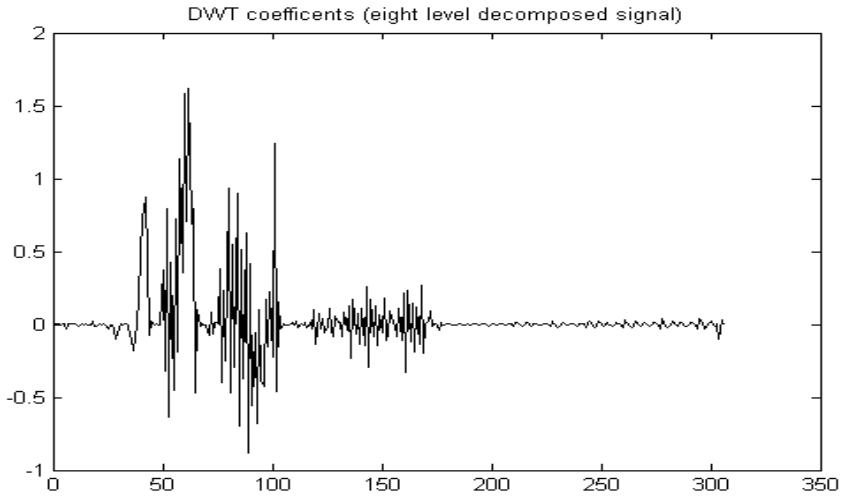



**Figure 3.10 8 level DWT Coefficients (Female Speaker)**

## CHAPTER 4

### 4. Classification

Systems designed to perform classification are termed *classifiers*. There are several approaches to the design of a classifier and they can be grouped into three classes: classifiers based on bayes decision theory, linear and non-linear classifiers. The first approach builds upon probabilistic arguments stemming from the statistical nature of the generated features. Linear classifiers examples such are the perceptron algorithm and the least squares method. For problems that are not linearly separable, non-linear classifiers are used. The classification can be *supervised* or *unsupervised* depending on whether a set of training data is available or not. Correct classification is dependent on the amount of differentiating info present in the measurements and it being put to efficient use in the preprocessing and feature generation stages.

### 4.1 An Introduction to Neural Networks

The human nervous system is comprised of nervous tissue, which is made up of nerve cells called neurons. Though the size and shape of neurons differ in different parts of the nervous system, they are basically similar. It consists of a cell body and the axon- one long fiber that transmits impulses away from the cell body. Terminal branches are called dendrites. Impulses are transmitted between neurons with the help of a junction called as synapse. *When a neuron receives excitatory input that is sufficiently large compared with its inhibitory input, it sends a spike of electrical activity down its axon*. *Learning occurs by changing the effectiveness of the synapses so that the influence of one neuron on another changes*.

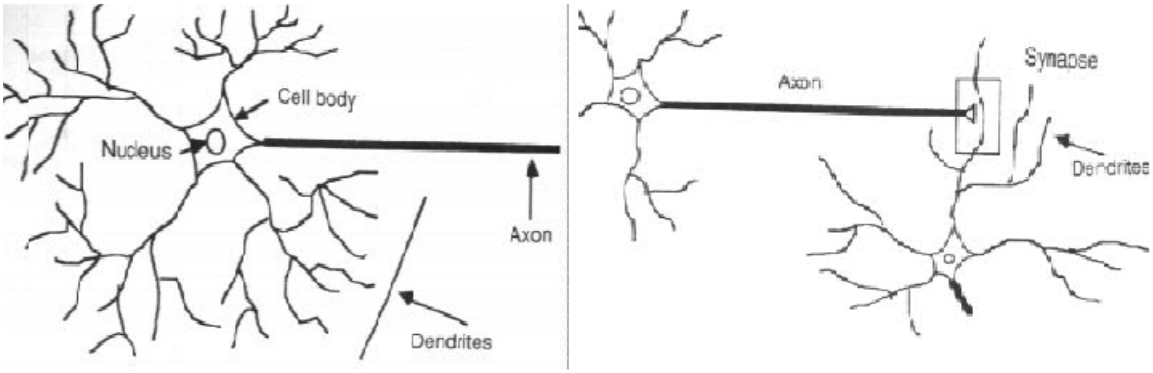



**Figure 4.1(a) Components of a nerve cell (neuron)** **Figure 4.1 (b) A Synapse**

A ***Neural Network***, similar to the human nervous system, consists of a multitude of highly interconnected simple processing elements called neurons as well. Each neuron is connected to certain of its neighbors with varying coefficients of connectivity. These represent the strengths or *weights* of these connections. *A weight represents the strength of association-that is, the concurrence of connected features, concepts, propositions, or events during a training period*. Rules for weights may change in neural networks.

*Learning* is achieved by adjusting these weights to facilitate the overall network to produce the desired results. Thus by adjusting the values of the weights between neurons, a neural network can be *trained* to execute a particular function. By and large, neural networks are adjusted or trained, so that a particular input leads to a specific target output[14]. This may be achieved in a number of steps, with the network being adjusted each time based on a comparison of the output and the target. When this has been accomplished the network is said to have converged. This type of training is called *supervised training* where normally numerous input/target pairs are used to train a network. Unsupervised training methods also exist but are comparatively less frequently used. Thus, it can be said that all knowledge in a neural network is encoded in the interconnection weights, with the learning process determining these weights.

The structure of a simple neuron with associated parameters is shown in Figure 4.2

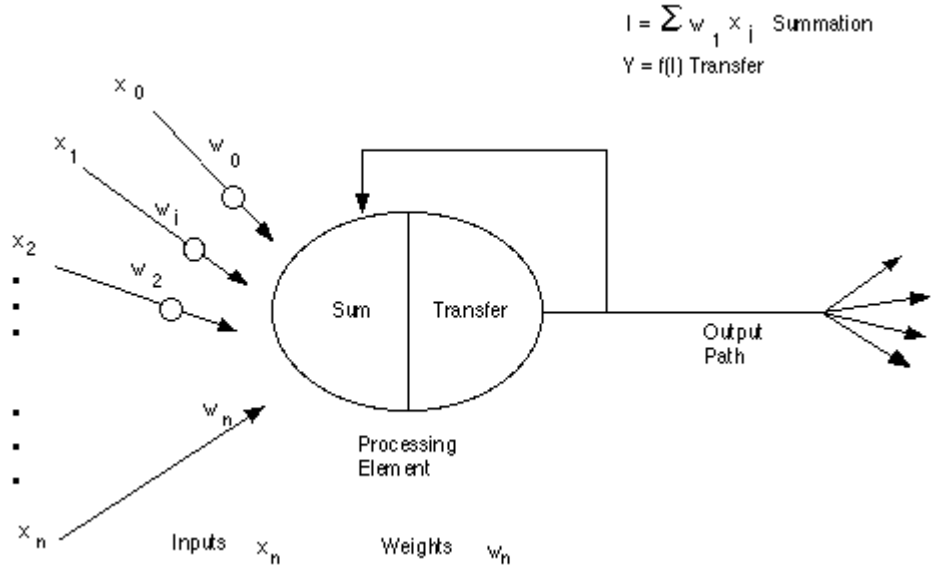



**Figure 4.2 A simple Neuron**

A neural network is typically organized in layers. These layers are made up of a number of interconnected 'nodes', which contain the activation function. Three types of layers exist, namely the hidden layer containing hidden units, and the input and output layers containing the input and output units respectively.

Input units represent the raw info that is fed into the network. The activities of hidden units are determined by the activities of the input units and the weights on the connections between the input and hidden units.

The output unit behavior depends on the activity of hidden units and the weights between the hidden and output units.

The behavior of a neural network is governed both by the weights and the input-output function (transfer function*)* that is specified for the units. This function, called the *activation function*, typically falls into one of three categories, namely, linear (or ramp), threshold and the *sigmoid*. Sigmoid functions are very useful activation functions. Here the output undergoes a continuous, non-linear and strict increase with the input. The logistic function and the hyperbolic tangent functions are the most commonly used

sigmoid functions. They are especially advantageous for use in neural networks trained by backpropagation, because the simple relationship between the value of the function at a point and the value of the derivative at that point reduces the computational burden during training.

Neural networks are either feedforward or feedback networks. *Feedforward networks* allow signals to travel one way only (from input to output) while feedback networks allow signals traveling in both directions.

### 4.2 The MLP Feedforward Back propagation Neural Network

The multi layer perceptron , constituted by solitary or a multiple number of hidden layers and employing the back propagation algorithm can tackle complicated problems that require building elaborate, non-linear decision boundaries in order to distinguish between the different pattern classes. Essentially, back propagation networks that have been properly trained tend to be less confused when presented with *new* inputs. Consequently, they have a higher probability of producing the right answers even when confronted with inputs they have never seen. Usually, a new inquiry would yield an output similar to the correct output with the result based upon a similar situation been faced by the network during training. This leads to an important, useful property: Without having to expose the network to the set of all possible inputs, we can get it to respond fairly correctly to almost any of our queries.
Figure 4.3 below shows the Multilayer Feedforward Network Structure

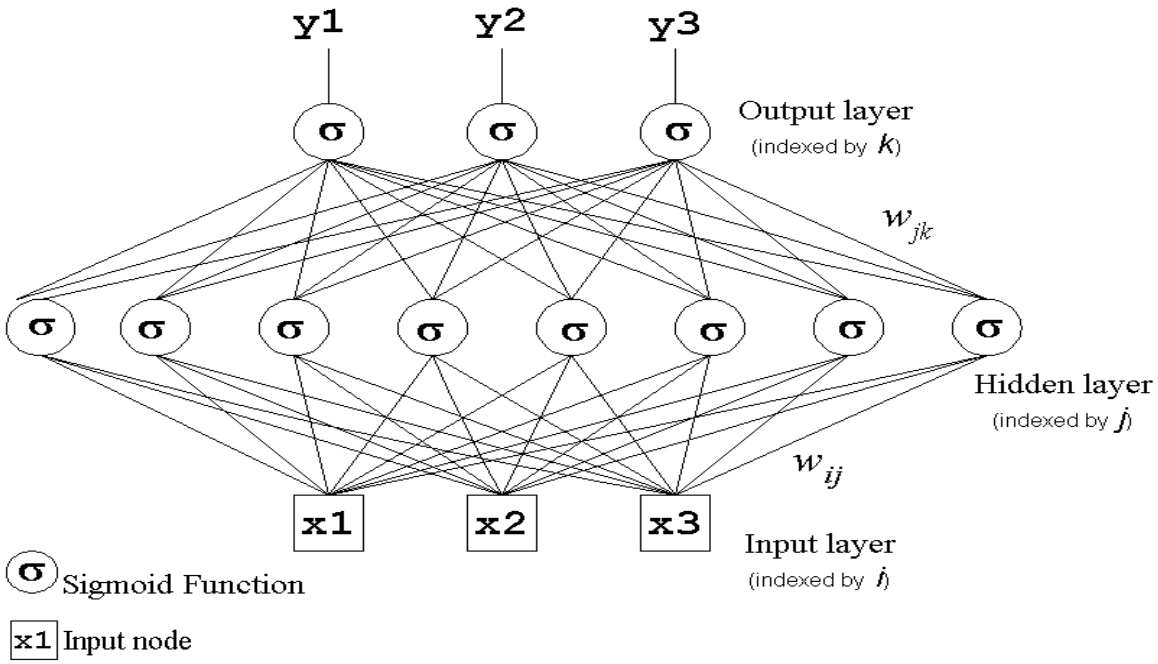



**Figure 4.3 The MLP feedforward network**

### 4.2.1 The Back-Propagation Algorithm

Back propagation algorithm uses the *gradient-descent* method to minimize the error on the training data by propagating errors backwards through the network layers starting at the output layer and working backwards towards the input layer. The connections between the layers are then adjusted accordingly, resulting in an enhancement of performance.

A suitable error function or cost function is selected. The values of this error function are dependent upon the actual and desired outputs associated with the network and the network parameters such as weights and thresholds. The gradient-descent method is an iterative process using repetitive calculation of the gradient value at different points of the weight space, moving in the direction of negative gradient. This ultimately leads to the *global minimum* although there can be exceptions. The goal, therefore, is to change the weights in such a way that they "move towards" the global minimum, reducing the error each time. Resultantly the error function is minimized.

The back propagation training consists of two computational passes: a *forward* and a *backward* pass. In the forward pass computation, as shown in Figure, given the

selected input vectors $Y_k^{(p)}$ and weight $w_{jk}$ between hidden nodes and input nodes, each hidden node $j$ will give the output

$$y_j^{(p)} = f(\sum_k w_{jk} y_k^{(p)}) \quad (4.1)$$

Where $f(.)$ is the transfer function. Then each output node will produce the final output.

$$y_j^{(p)} = f(\sum_k w_{ij} f(\sum_k w_{jk} y_k^{(p)})) \quad (4.2)$$

In the backward pass, we observe that the cost function is calculated as the mean square error (MSE) over all output units and over all patterns $p$ is given by the following expression

$$E = \frac{1}{2} \sum_p \sum_l (d_i^{(p)} - y_i^{(p)})^2$$

Or

$$E = \frac{1}{2} \sum_p \sum_l (d_i^{(p)} - f(\sum_k w_{ij} f(\sum_k w_{jk} y_k^{(p)})))^2 \quad (4.3)$$

Now, the gradient descent method is applied to find the gradient descent rule for the hidden layer to output layer connections as follows

$$\Delta w_{ij} = -\eta(\partial E / \partial w_{ij}) \quad (4.4)$$

Where $\eta$ is a constant that determines the rate of learning. By using the chain rule, we find

$$\Delta w_{ij} = \eta \sum_p (d_i^{(p)} - y_i^{(p)}) f'(x_i^{(p)}) y_i^{(p)} = \eta \sum_p \delta_i^{(p)} y_i^{(p)} \quad (4.5)$$

In the same manner the gradient descent rule for the input layer to hidden layer connections can be found as follows.

$$\Delta w_{jk} = -\eta(\partial E / \partial w_{jk}) \quad (4.6)$$

By again using the chain rule, we get

$$\Delta w_{jk} = \eta \sum_p \sum_i (d_i^{(p)} - y_i^{(p)}) f'(x_i^{(p)}) w_{ij} f'(x_i^{(p)}) y_k^{(p)} \quad (4.7)$$

Or

$$\Delta w_{jk} = \eta \sum_{p} \delta_j^{(p)} y_k^{(p)}$$

Where

$$\delta_j^{(p)} = f'(x_j^{(p)}) \sum_i \delta_i^{(p)} w_{ij} \qquad (4.8)$$

From the equations above, it can be seen that the transfer functions must be differentiable. If a sigmoid function is chosen as the transfer function, that is

$$f(x) = \frac{1}{1+e^x} \qquad (4.9)$$

Also we can find that

$$f'(u) = f(u) - [1 - f(u)] \qquad (4.10)$$

### 4.3 Applications of Neural Networks

Neural Networks have found extensive use in a variety of fields including science, business and medicine. With the passage of time their application is being extended to newer and more diverse areas.

Neural networks have broad applicability to real world business problems. In fact, they have already been successfully applied in many industries. Since neural networks are best at identifying *patterns* or trends in data, they are well suited for prediction or forecasting needs including sales forecasting, industrial process control, customer research, data validation, risk management as well as target marketing. In medicine, neural networks are turning out to be a valuable asset. For instance, they are being used experimentally to model the human cardiovascular system[15].

Neural networks have found effective use in telecommunication networks. Traffic trends analysis in ATM networks is done with the help of neural networks. Neural networks are also used for alarm correlation in cellular phone networks. Alarm handling, especially alarm correlation are necessary to manage large telecom networks. Control of dynamic channel allocation for mobile radio networks is yet another application of neural networks in the telecom domain.

Moreover, neural networks can also be used as a resource allocation tool in telecommunication networks. The allocation of capacity in an SDH network and the implementation of an optimal call acceptance policy for a mobile network are two examples of such applications.

More particular examples include the use of neural networks in the following specific paradigms*: recognition of speakers in communications*; diagnosis of hepatitis; recovery of telecommunications from faulty software; interpretation of multi meaning Chinese words; undersea mine detection; texture analysis; three-dimensional object recognition; hand-written word recognition; and facial recognition.

# CHAPTER 5

# RESULTS

## 5.1 The Results Obtained –An Analytical Discussion

The most crucial stage in the process of classification in an automatic speaker is the extraction and selection of the appropriate features that carry useful discriminatory information between the classes. When such features are obtained the problem of classification becomes simplified since we can use many useful algorithms that can classify both linearly and non-linearly separable classes.

In our case we first tried the Fast Fourier transform (FFT) coefficients as features to be tested by the classification stage. The results for the FFT features were about 95 % for four speakers, which was not very good at all. The reason for this bad performance is that FFT is not capable of analyzing the time varying signals i.e. the speech signals. The time localization of various frequency components present in the signal is lost and thus Discrete Fourier transform coefficients did not contain enough discriminatory information to distinguish the four classes from each other.

The discrete wavelet Transform (DWT) coefficients were obtained using the subband-coding algorithm. As mentioned in the earlier chapters DWT is a very powerful signal analysis tool that works like a magnifying glass for analyzing time varying and quasi stationary signals. The subband-coding algorithm gives very good time resolution high frequencies (with a relatively poor frequency resolution) and a good frequency resolution is achieved at lower frequencies. Due to these powerful characteristics of DWT coefficients were chosen as features after carefully choosing the useful frequency band. These coefficients are then applied to MLP back propagation neural network. The results obtained for identification and verification for a closed set of four speakers are

shown in table 5.1. The results obtained are the best as obtained with the available closed data set. The classification performance is tested for various combinations of training/testing data and different neural network architecture as explained in the chapter no 4. The classifier performance varied from 3-5 % with different sets of training/testing data. Also, neural networks with one or two hidden layer were performing well for the 108 point feature vector and introducing a third hidden layer did not improve the performance instead it resulted in increased training time. In our system we observed that the best results were given by neural network with 27 nodes in one hidden layer and the average training time was just 2 sec.

Table 5.2 shows the verification results obtained from 6 speakers and it can be observed that the misclassification results are very good when the classifier verifies a new speaker.

To enhance the robustness of the speaker verification system a telephone recorded speech database was collected and as seen in table 5.3 the results for the closed set telephone speech classification results are also very good.

### 5.2 Discussions (Strengths and Weaknesses)

It is clear that the speaker verification technology is indeed ready for use but it is not a universal solution. The main strength of speaker verification technology is that it relies on a signal that is natural and unobstructive to produce and can be obtained easily from almost anywhere using the familiar telephone network with no special user equipment or training. This technology has prime utility for applications with remote users and applications already employing a speech interface. Additionally, speaker verification is easy to use, has low computation requirements and given appropriate constraints, has high accuracy.

Some of the flexibility of speech actually lends to its weaknesses. Speech is behavioral signal that may not be consistently reproduced by a speaker and can be affected by a speaker's health (e.g. cold). Second the varied microphones and channels that people use can cause difficulties since most speaker verification systems rely on low level spectrum features susceptible to transducer effects. Robustness to channel variability is the biggest challenge to current systems. Spoofing of a system is often cited as a weakness, but there have been many approaches developed to thwart such attempts. Some of these weaknesses may be overcome by combination with a complementary biometrics, like face recognition. [16]

**Table 5.1 Summary of classification results for four speakers for closed set**

| SR NO | TOTAL SAMPLES | TRAINING/ TESTING SAMPLES | VALIDATION SAMPLES | PERFORMANCE (%) | |
|---|---|---|---|---|---|
| | | | | VERIFICATION | IDENTIFICATION |
| SPEAKER 1 (MALE) | 300 | 200/40 | 60 | 100 % | 100 % |
| SPEAKER 2 (MALE) | 300 | 200/40 | 60 | 100 % | 98.33 % |
| SPEAKER 3 (FEMALE) | 300 | 200/40 | 60 | 100 % | 100% |
| SPEAKER 4 (FEMALE) | 300 | 200/40 | 60 | 100 % | 100% |
| TOTAL | 1200 | 800/160 | 240 | 100 % | 99.58% |

| SR NO | TOTAL SAMPLES | TRAINING TESTING SAMPLES | VALIDATION SAMPLES | PERFORMANCE (%) | |
|---|---|---|---|---|---|
| | | | | VERIFICATION | IDENTIFICATION |
| SPEAKER1 (MALE) | 30 | 24 | 6 | 100% | 100% |
| SPEAKER2 (MALE) | 30 | 24 | 6 | 100% | 83% |
| SPEAKER3 (FEMALE) | 30 | 24 | 6 | 100% | 100% |
| SPEAKER4 (FEMALE) | 30 | 24 | 6 | 100% | 100% |
| TOTAL | 120 | 96 | 24 | 100% | 95% |

**Table 5.2 Summary of classification results for four speakers open set**

| SR NO | TOTAL | VALIDATION SAMPLES | PERFORMANCE (%) VERIFICATION |
|---|---|---|---|

| | SAMPLES | | |
|---|---|---|---|
| SPEAKER 1 (MALE) | 30 | 30 | 93.33% |
| SPEAKER 2 (MALE) | 30 | 30 | 93.33% |
| SPEAKER 3 (MALE) | 30 | 30 | 90% |
| SPEAKER 4 (FEMALE) | 30 | 30 | 93% |
| SPEAKER 5 (FEMALE) | 30 | 30 | 90% |
| SPEAKER 6 (MALE) | 30 | 30 | 96.67% |
| TOTAL | 180 | 180 | 92.72% |

**5.3 Summary of classification of four speakers closed set telephone speech**

| SR NO | TOTAL SAMPLES | TRAINING/ TESTING SAMPLES | VALIDATION SAMPLES | PERFORMANCE (%) (Verification) |
|---|---|---|---|---|
| SPEAKER 1 (MALE) | 130 | 80/20 | 30 | 97% |
| SPEAKER 2 (MALE) | 130 | 80/20 | 30 | 100% |
| SPEAKER 3 (FEMALE) | 130 | 80/20 | 30 | 97% |
| SPEAKER 4 (FEMALE) | 130 | 80/20 | 30 | 100% |
| TOTAL | 520 | 320/80 | 120 | 98.5% |

### 5.3 Future trends

Speaker recognition continues to be data-driven field, setting the lead among other biometrics in conducting benchmark evaluations and research on realistic data.

**5.3.1 Focus on Real World Robustness:** The continued ease of collecting and making available speech from real applications means that researchers can focus on more real-world robustness issues that appear. Obtaining speech from a wide variety of handsets, channels and acoustic environments will allow examination of problem cases and development and application of new or improved compensation techniques.

**5.3.2 Emphasis On Unconstrained Tasks**: With text-dependent systems making commercial headway, R&D effort will shift to the more difficult issues in unconstrained situations. This includes variable channels and noise conditions, text-dependent speech and the tasks of speaker segmentation and indexing of multi-speaker speech.

### 5.4 Future Work

The system developed should be tested on more users. Along with this, potential work can be done to incorporate more users in to the system, or even make it a variable user system that can add users according to the requirement of work. This would make the system very appealing for the corporate sector especially, as most companies require an Access Allowed/Denied system for a multitude of employees and at various levels depending upon the hierarchical structure existing within the company.

Another avenue of future work can be to strengthen system robustness. The real-world scenario presents us with a myriad of different and often difficult situations where noise can be a very disturbing factor. This is important specially as noise can significantly degrade system performance. The fact that there are so many sources that add noise further augments the emphasis in this respect.

Different types of mother wavelets should be researched to create a mother wavelet function that will give an informative, efficient and useful description of the signal of interest.

Speech recognition should be incorporated in the proposed system to make it a complete representative system.

Techniques should be used towards the development of machine-driven text independent systems. This is important as with the accelerated pace of technological development, system having such attributes would be needed one time or another in the future.

## References

[1] R. Mammone, X. Zhang, and R. Ramachandran, “Robust speaker recognition—A feature-based approach,” IEEE Signal Processing Mag., vol. 13, no. 5, pp. 58–71, 1996.

[2] D. Reynolds and R. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” IEEE Trans. Speech Audio Processing, vol. 3, no. 1, pp. 72–83, 1995.

[3] A. Higgins, L. Bahler, and J. Porter, “Speaker verification using randomized phrase prompting,” Digital Signal Processing, vol.1, no. 2, pp. 89–106, 1991.

[4] Bourlard, Morgan, “Speaker Verification - A Quick Overview” (1998).

[5] G. R. Doddington, “Speaker recognition—Identifying people by their voices,” Proc. IEEE, vol. 73, pp. 1651–1664, Nov. 1985.

[6] Douglas A. Reynolds and Larry P. Heck, “Automatic Speaker Recognition, Recent Progress, Current Applications, and Future Trends” Presented at the AAAS 2000 Meeting.

[7] T. Matsui and S. Furui, “Concatenated phoneme models for text-variable speaker recognition.” In ICASSP, pages 391--394.

[8] D. Reynolds and B. Carlson, “Text-dependent speaker veri-fication using decoupled and integrated speaker and speech recognizers,” in Proc. EUROSPEECH, Madrid, Spain, 1995, pp. 647–650.

[9] R. Gnanadesikan and J. R. Kettenring, “Discriminant analysis and clustering,” Statistical Sci., vol. 4, no. 1, pp. 34–69, 1989.

[10] S. Theodoridis, K.Koutroumbas, “Pattern Recognition”, Academic Press, 1999.

[11] I. Daubechies, “The Wavelet Transform, time-frequency localization and signal analysis” IEEE Transactions on Information theory, Vol. 36, pp. 961-1005, 1990.

[12] R.Polikar, “The Wavelet Tutorial” , last viewed April 24,2003. Available at http://www.engineering.rowan.edu/~polikar/WAVELETS/WTtutorial.html

[13] S.G.Millat, “Multiresolution approximation and Wavelet orthonormal bases of L2” Transactions of American Mathematical Society, Vol.315, pp. 69-87 , September 1989.

[14] C.G.Looney, “Pattern Recognition using neural networks: Theory and algorithms for engineers and scientists,” Oxford University Press, 1997.

[15] L. Udpa, S. S. U dpa, “Neural networks for the classification of non destructive evaluation signals” IEEE Proceedings on Radar and Signal Processing Vol. 138, pp. 41-45, February 1991.

[16] Douglas A. Reynolds, “An Overview of Automatic Speaker Recognition Technology”, IEEE Proceedings, pp.4072-4075, 2002

[17], www.speech.cs.cmu.edu/comp.speech, 20 March, 2003.