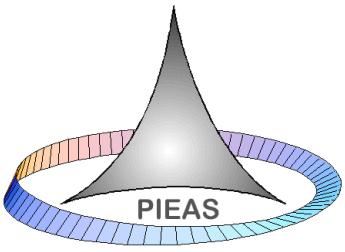
# Automatic Speaker Recognition System



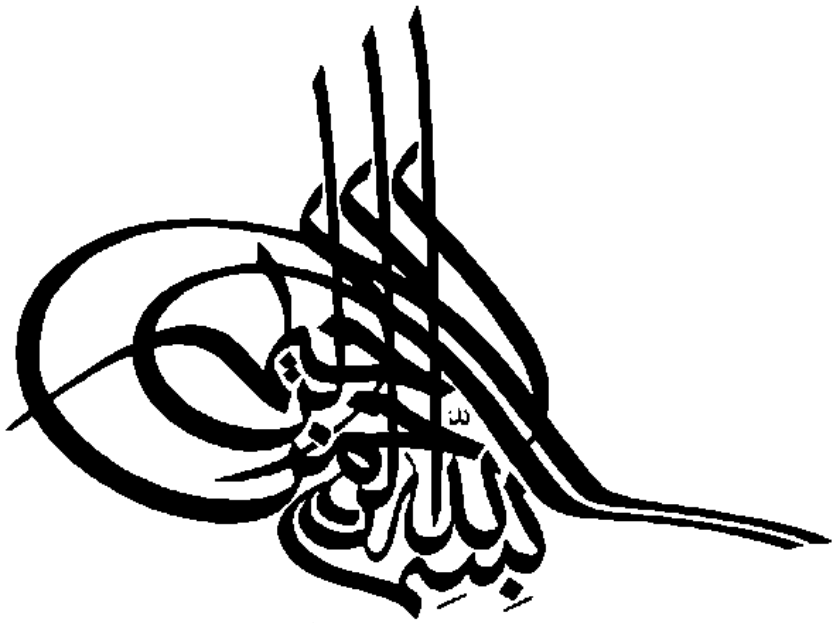

**Sidra Malik**

(Thesis submitted in partial fulfillment of the requirements
for the BS Degree in Computer and Information Science)

**Pakistan Institute of Engineering & Applied Sciences,**

**Nilore-45650, Islamabad.**

**June, 2009**



*"Truly my prayer and my service of sacrifice, my life and my death, are (all) for Allah, the Cherisher and Sustainer of the Worlds."*

# Certificate of Approval

Certified that the work contained in this report entitled

**"Automatic Speaker Recognition System"**

was carried out by Sidra Malik under my supervision and that in my opinion, it is fully adequate, in scope and quality, for the project of BS. (Computer Information Sciences.)

Approved By:

Signature: _____

Supervisor: Fayyaz-ul-Amir Afsar Minhas

June, 2009

*Dedicated to My Beloved Parents*

# Acknowledgements

All thanks to Allah Almighty for his blessings and strength he gave us to fulfill our aims. He is for sure the one to give us determination, understanding, joys and sorrows. On completion of this project and all the success I have achieved, I am truly grateful to Him. All respects for our dear Holy Prophet Muhammad (SAW), the last Prophet of Allah & great benefactor of mankind.

This work would not have been carried out successfully without the support of many people. First of all, I am extremely grateful to my supervisor, Fayyaz-ul-Amir Afsar Minhas for his support. His has always been dedicated to his work and I admire his very, very patient guidance. I am thankful to my parents, especially to my mother for her sweet prayers and care. Special thanks to my sisters for keeping me motivated always.

All my regards for my friends and especially those who contributed in collecting a database for my project and lastly, I am grateful to myself for choosing this project.

# Table of Contents

# List of Figures

# List of Tables

# Abstract

The objective of this project is to develop a text independent Automatic Speaker Recognition System titled as TAURUS.

Initially, the database was collected and a technique mentioned in a research paper titled “Robust speech features based on wavelet transform with application to speaker verification” [6] was implemented. According to the algorithm mentioned in [6], DWT (Discrete wavelet Transform) based Feature Extraction technique, LPC (Linear Predictive Coding), had been studied and implemented. These LPCCs were based on the coefficients given by the DWT (Discrete Wavelet Transform) of the signal. K-means clustering was used as a vector quantizer to standardize the dimensions of the features for variable length speech data. Classification was carried out using nearest neighbor classification algorithm. Moreover the results were taken on Non Telephonic Database. Recognition accuracy for DB1 (Database-1) was 90.38% whereas on DB2, it was 92%. Effect of reducing sample length to half was observed, accuracy was reduced to 85.01% and 85.7% for DB1 and DB2 respectively. Further, recognition rate was observed separately on male and female speakers. On males, accuracy obtained was 88% whereas on females it was 99.5%. Lastly, few approximations on recognition time were also calculated and approximated time for each sample was 0.3s.

The above technique has been applied to telephonic data (both in training and testing) and results have been collected which gave the accuracy of 86.84% on DB1 (Database-1) and 92.11% on DB2 (Database-2). Also, the decomposition levels for non telephonic database have been increased from one to three in order to analyze the effect of decomposition levels on recognition rate. Best results are obtained on level 1 and level 3 of DB1 and DB2 respectively. Next, I have used Non Telephonic speech for training and Telephonic for testing, i.e. mismatched conditions recognition rate has been reduced by 50-90% and error analysis phase in this regard is being carried out.

After that the algorithm for voice activity detection was revised and errors that resulted due to voice activity detection previously, were resolved. Errors that occurred in mismatched conditions were further analyzed and one of the channel normalization techniques, Cepstral Mean Subtraction was implemented. Results showed that the

error was probably not much due to channel noise but the variance of microphone used in different handsets.

Secondly, effect of number of LPCs was observed on Non Telephonic and Telephonic databases. For Non Telephonic as well as Telephonic speech data, appropriate number of LPCs came out to 12.

Thirdly, another feature extraction method, Mel-Frequency Cepstral Coefficients was used rather LPCs. MFCCs have exhibited better performance than LPCs. For this purpose, first task was to find appropriate number of MFCCs which unlike LPCs was different in case of Telephonic and Non Telephonic speech. In Non Telephonic, 24-MFCCs work best whereas in case of Telephonic, again there are variations regarding handset type. For Handset 1, the appropriate number is 36 whereas for Handset 2, it is 34. Generally, for Telephonic, 36-MFCCs have been used here. On selected parameters, MFCCs

Results were observed by using simple MFCCs and MFCCs with wavelet transform and were compared with those of LPCs. Simple MFCCs and DWT followed by MFCC showed somewhat similar results on DB1. On DB1-Non Telephonic, accuracy was 97.11%, 92.11% on Handset 1 and 94.74% on Handset 2 which is much better than that of DWT-LPC which is 89.47% on Non Telephonic, 86.82% and 92.11% on Handset 1 and Handset 2 respectively. In contrast to DB1, significance of using DWT along MFCC can be clearly observed on DB2. On DB2 Non-Telephonic, 99.23% was regarding simple MFCCs, 99.92% using DWT-MFCCs and 91.92% using DWT-LPCs. On Handset 1, it was 83.8% (MFCC) , 91.54% (DWT-MFCC), and 73.86% (DWT-LPC) whereas 91.54% (MFCC) , 93.85% (DWT-MFCC) and 76.15% ( DWT-LPC) on Handset 2.

Moreover, effect of decomposition levels was being observed on MFCC. Unlike LPC, MFCCs showed better results on level 1. The accuracy was reduced as the number of levels was increased from 1 to 5.

Effect of Wavelet type has also been observed by using different types of wavelets like Daubechies, Haar, Symlets, Discrete Meyer. They overall do not cause much difference in accuracy. On DB1, Daubechies-3 and Symlets performance is better than all, whereas on DB2, Discrete Meyer has resulted in better accuracy.

Finally, Results were being computed on mismatched conditions. There is a considerable improvement by using this new technique which uses MFCCs. When the system was trained on Non Telephonic Speech and tested on Handset 1 and Handset 2

(DB1) there was approximately 20% increase in accuracy in case of Handset 1 and 30% increase in Handset 2. Similarly, when system was trained on one handset and tested on other, accuracy was 50.37% on DB1 which is 50% greater than the one resulted using previous technique.

Recognition rate achieved is quite good when population size was increased i.e. 96.25% for non telephonic and 86.77% for telephonic speech for 64 speakers. Moreover for the proposed technique, best parameter selection was analyzed and for PIEAS Speech Database, 38 MFCCs based on wavelet type Symlets 7 and decomposition level 1 have proven best. The performance of proposed method is comparable to approaches for feature extraction based on wavelet transform, LPCs and LPCCs as well as to speaker modeling techniques like VQ.

# Chapter 1
# Introduction

Human race has come a long way since its inception in small primitive societies where every person knew the other but in today's geographically mobile world where the societies are connected and grow electronically, the problem of person's identification remains a challenge.

In current era of Information Technology, the person's presence is not essential for transaction means, or in other fields where the person itself is needed to do provide a password, personal identification numbers (PINs) etc. Moreover, these conventional means of identification can be stolen, reused and replicated which is a threat to crucial environments where a chance on person's identity cannot be taken. To solve this problem, biometrics is seen as one of the best aspirants. From the increasing importance of biometrics, it is believed that Biometrics will become a significant component of information technology because of the increasing strength of biometrics and the continuous fall in prices of biometric sensors.

## 1.1 Biometrics

Biometrics deals with the person's identification on the basis of behavioral and biological characteristics. The field of Biometrics consists of automated methods which use features to be measured; face, fingerprints, hand geometry, handwriting, iris, retinal, vein, and voice. The reason for these methods to become a foundation for identification is that they are highly secure because person's physical or behavioral traits cannot be stolen. Biometric-based solutions are able to provide confidential financial transactions and personal data privacy. Thus, in electronic media of today, this method is reliable and safe to be used. Classification of Biometric Traits

Biometric characteristics can be divided in two main classes i.e. Physical biometrics and behavioral biometrics. The former is related to natural shape of a body like fingerprints and the latter is related behavior of a person, for example signature.

### 1.1.1 Parameters of Biometric System

For any of the human physiological or behavioral characteristics, following parameters must be fulfilled.

**a. Universality**

All persons should possess a particular characteristic.

**b. Uniqueness**

A person has its own unique trait i.e. no two persons can have similar characteristics.

**c. Permanence**

The observed characteristics should be invariant of time.

**d. Collectability**

Collectability means that the characteristics can be measured quantitatively.

**e. Performance**

Performance of biometric systems is achievable identification accuracy, robustness and the consumption of resources.

**f. Acceptability**

Acceptability is the extent to which people are willing to accept the biometric system.

**g. Circumvention**

Circumvention is a measure of how easily a system can be fooled with frauds and forgeries.

### 1.1.2 Different Biometric Technologies

**a. Face**

Face recognition analyzes facial characteristics. This technique has attracted considerable interest because it is the most natural biometric for distinctiveness verification. Certain factors like aging, mood, make-up, pose, and lightning condition are the great source of error in this case.

**b. Infrared Facial and Hand Vein Thermograms**

Heat from any human body can be sensed by an infrared sensor which acquires an image indicating the heat emanating from different parts of the body as shown in

Figure 1.1. These images are called thermograms. The absolute importance of this method lies in for the diagnosis of drug usage but this method is expensive.

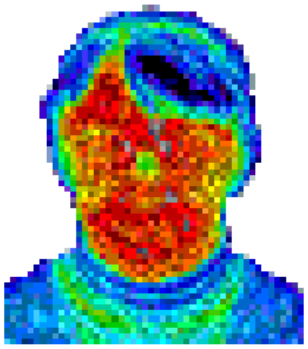


**Figure 1.1 Identification based on facial thermograms [37]**

### c. Fingerprint

Fingerprints are graphical ridges present on human fingers and they are believed to be unique to every person. This is one of the mature biometric technologies especially in forensic divisions. Typically, a fingerprint image is captured in one of two ways: (i) scanning an inked impression of a finger or (ii) using a live-scan fingerprint scanner as shown in Figure 1.2.


**Figure 1.2 (a) Inked-fingerprint (b) Live scan fingerprint [37]**

### d. Iris

Just like fingerprints, Iris formation is An iris image is typically captured using a non-contact imaging process. An example of such an image can be seen in Figure 1.3.
The identification error rate using iris technology is believed to be extremely small.

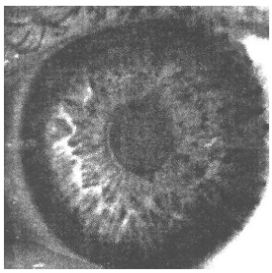


**Figure 1.3 Identification based on Iris (visual texture of Iris) [37]**

**e. Hand and Finger Geometry**

Hand geometry has become a very popular access control biometrics which has captured almost half of the physical access control market [37].

One of the great advantages of this technology is that the representational requirements of the hand are very small (9 bytes) which is an attractive feature for bandwidth and memory limited systems. The disadvantage of such systems is that the hand geometry is not unique and cannot be performed well when the identification is performed from a large database.

**f. Voice**

Voice is a physical as well as behavioral trait but is not expected to be a unique characteristic of an individual because it tends to change according to environment, stress, age, communication channel, and microphone. Using this biometric is challenging because speech can be mimicked, production of same speech can alter time to time.

**Other Biometric Technologies**

There is ongoing research on other biometric technologies including signature, retinal patterns, DNA typing, hand vein , keystroke dynamics, ear shape, gait, lip shape and ear shape for authentication of person.

## 1.2 Architecture of Biometric Systems

A biometric system is a pattern matching system, which makes an identification or verification decision by analyzing one or more biometric characteristic of a person. It is a combination of hardware which senses a particular trait that is interconnected with software modules.

The different logical modules in biometric system are acquisition, enrollment and test module.

**a. Acquisition Module**

This module is to capture all the important information regarding a particular biometric feature; hence it is the interface between the outside world and the system. Examples include fingerprint scanners, signature tablets, cameras, microphones etc. the basic block diagram of a biometric system can be seen in Figure 1.4.

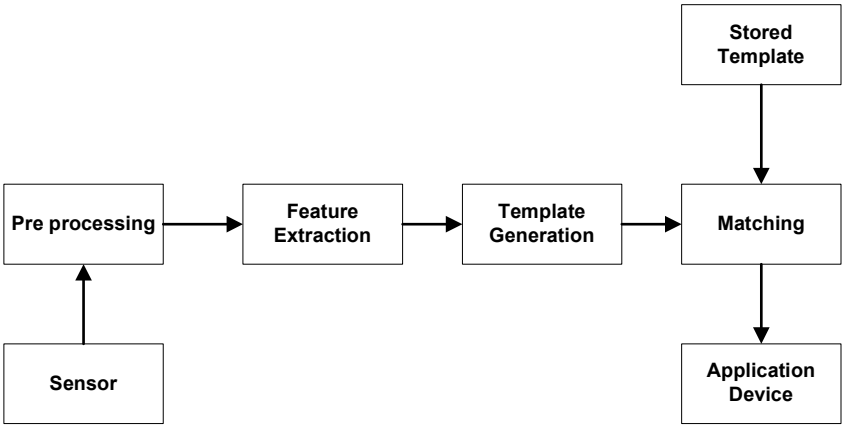



**Figure 1.4: The basic block diagram of biometric system**

**b. Enrollment Module**

During the enrollment, a system is to store biometric information of an individual. This block performs all the tasks starting from capturing a feature from an individual, performing all the preprocessing needed, choosing the efficient method to extract features and model the subject and in the end to form a template. A template contains all the extracted features necessary for authentication without any loss of information. After that this template is being stored in system's database for further use.

**c. Testing Module**

During testing, biometric information are provided as a test sample of any individual which also requires preprocessing. Feature extraction of that test sample is carried out and matched with templates stored in a database. Distance is estimated between them using any matching or classification techniques (e.g. Minimum Distortion). A verification system authenticates the identity claim of a person by performing a 1-1 comparison of the captured biometric feature with the subject's own pre-stored biometric template. An identification system conducts one-to-many comparison in order to search for the given biometric characteristic in the template database.

## 1.3 Speaker Recognition System and its Types

Automatic speaker recognition is the most economical biometric used to solve problems regarding unauthorized use of computers and multi level access control. The motivation for using this biometric lies in its cost effectiveness as voice can easily be captured through an existing large telephone network and microphones, the only cost is for the software itself.

There are few different types of speaker recognition systems used for authentication purpose.

### 1.3.1 Single Pass Phrase System

This is a phrase dependant authentication system in which the speaker is required to utter the same phrase as it was uttered during the enrollment in other words, text dependant speaker recognition system. This type of authentication has an advantage that it can still perform well provided with a little speech data and the intruder needs a correct pass phrase along with the speaker's voice in order to sneak in.

### 1.3.2 Text Prompt System

This kind of system allows the speaker to utter the text provided by the system. This text can be the one for which that particular speaker is enrolled or different from that text. The advantage of such systems is that it makes replay-attacks as in case of single pass phrase systems, more difficult because speakers are to utter a different text each time. One of the major disadvantages of such a system is that it requires a longer utterance for recognition.

### 1.3.3 Speaker authentication along with dialog system

In this third type of speaker verification system, speaker verification is integrated within a speech recognition system. For example, in a bank if the speaker is asked to utter his account number and password, first the speech recognition system authenticates the password being spoken and then the speaker is verified by the voice recognition system by matching the voice sample to template stored in a system. This is rather a strong than above two.

## 1.4 Project Objectives

The main objectives of the project are given below:

i. Literature Survey to get an overview of some of the techniques which are widely used Automatic Speaker Recognition (ASR), Moreover, to study techniques that are efficient to recognize a speaker in clean as well as noisy environments, especially when a telephonic channel is used.
ii. The first phase is enrollment of subjects. For this purpose, the data are collected which includes clean as well as telephonic speech samples. Next step is to extract features using an effective feature extraction technique and form a template database.
iii. After the feature extraction module, the suitable classification technique according to literature must be used.
iv. In the implementation phase MFCC and LPC have been used as features extraction techniques and VQ as feature modeling method. Moreover, for classification minimum distortion measure has been used.
v. After all the above, some other experiments based on selection of different parameters (like the appropriate number of features, wavelet type, decomposition levels), increasing population size, gender, and time have been carried out.

## 1.5 Thesis Organization

The thesis is organized as follows: Chapter-2 gives insight to speech perception, production and speech acoustics. Chapter-3 describes the various processes and terminologies that are involved in a speaker identification and verification system along with the applications of speaker recognition. Chapter-4 details about implemented techniques, their introduction and description. In Chapter-5, the implementation results are given with Telephonic and Non Telephonic data. Chapter 6 gives conclusion and future recommendations in this Project.

# Chapter 2
# Speech Production, Acoustics and Perception

Speech is classified into three perspectives, i.e. articulatory, acoustic and perceptual perspective. In articulatory approach, the description of production of speech based on anatomy and physiology of speech organs is given. Acoustic perspective gives the acoustic properties of signal itself. In perceptual approach, anatomy and physiology of hearing mechanism of humans is examined.

## 2.1 Articulatory Approach

Speech production consists of numerous parts which include message formulation, coding of a message into a language code, mapping of the language code to neuromuscular commands and the realization of those neuromuscular commands. Voice production organs are shown in Figure 2.1. Normally, the physiological part of voice production system is time-varying and consists of

**Sub glottal Component**: it is related to lungs and associated respiratory organs.

**Larynx:** contains vocal folds

**Supralaryngeal vocal tract:** consists of pharyngeal, oral and nasal cavities.

### 2.1.1 The Subglottal Respiratory System

Subglottal process empowers and air stream which helps in speech production. During the respiration process, when the air is inhaled into the lungs, they expand in their volume and the energy is stored in these elastic expansions. When the air is exhaled, this energy is spontaneously released and the airstream flows through windpipe or trachea to the larynx.

### 2.1.2 The Larynx

The larynx is responsible for different phonation mechanisms [7], which produces acoustic energy which acts as a input to vocal tract. Larynx has also an important

function of blocking trachea and opening esophagus during swallowing. Along with larynx, vocal folds and glottis have also an important role in speech production. Glottis is a small, triangular region between vocal folds. When the air from lungs passes through glottis and to vocal tract, the vocal folds determine the type of phonation which are voicelessness, whisper and voicing [7].

The difference between whisper and voiceless phonation is determined by the degree of the glottal opening. In whisper, the glottal area is smaller which results in a turbulent airstream, generating the characteristic “hissing” sound of whispering whereas in voiceless phonation, the area of the glottis will be larger and the airstream is only slightly turbulent when it enters the vocal tract. An example of voiceless phonation is the initial [h] in the Finnish word ‘hattu’ (a hat) [7]. Voicing on the other hand is a phenomenon of opening and closing of vocal folds periodically. This mechanism is more complex than voiclessness and whisper.

### 2.1.3 Vocal Tract

The vocal tract is the most important in the speech production process. Vocal tract refers to voice organs above the larynx. The three main cavities of the vocal tract are the pharyngeal, oral and nasal cavities which are responsible for producing vowels and nasal sounds. For example, nasal sounds are the ones produced by ‘m’ and ‘n’.

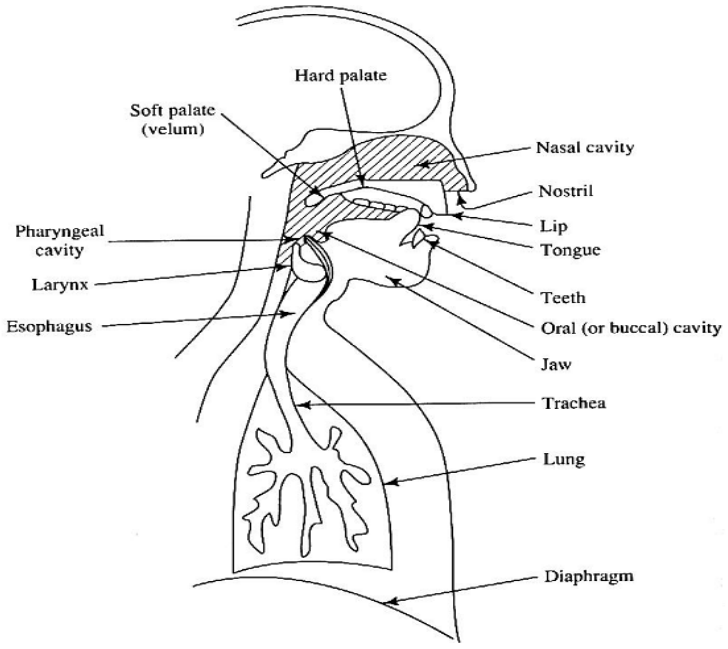



**Figure 2.1 Human voice production system [7].**

Other parts that contribute in shaping a sound wave are given below [36]:

**Velum (Soft Palate):** operates as a valve, opening to allow passage of air (and thus resonance) through the nasal cavity. Sounds produced with the flap open include m and n.

**Hard palate:** a long relatively hard surface at the roof inside the mouth, which, when the tongue is placed against it, enables consonant articulation.

**Tongue:** flexible articulator, shaped away from the palate for vowels, placed close to or on the palate or other hard surfaces for consonant articulation.

**Teeth:** another place of articulation used to brace the tongue for certain consonants.

**Lips:** can be rounded or spread to affect vowel quality, and closed completely to stop the oral air flow in certain consonants (p, b, and m).

## 2.2 Acoustic Approach

Acoustic approach aims to find the correlation of physiology and behavioral aspects of speech production organs. This analysis is either being done in frequency domain or time domain.

### 2.2.1 Spectrographic Analysis of Speech

Speech can graphically be represented in two important ways, i.e. as a speech waveform and as a spectrogram. A waveform represents air pressure variations whereas spectrogram shows magnitude of different frequencies present in a speech signal at different intervals of time. The Spectrogram as shown in Figure 2.2 depicts the dense areas as areas of high magnitudes of frequency at a certain time.

There are two types of spectrograms: wideband and narrowband spectrograms. In wideband spectrograms, the bandwidth of the analysis filter is around 300 Hz and thus the time spacing is approximately 1/300 s = 3.33 ms. For narrowband analysis, the bandwidth is around 50 Hz and thus the time spacing is around 1/50 s = 20 ms Wideband spectrograms are suitable for tracking vowel formants whereas the narrowband spectrograms can be used in Fundamental frequency (F0) estimation [7].

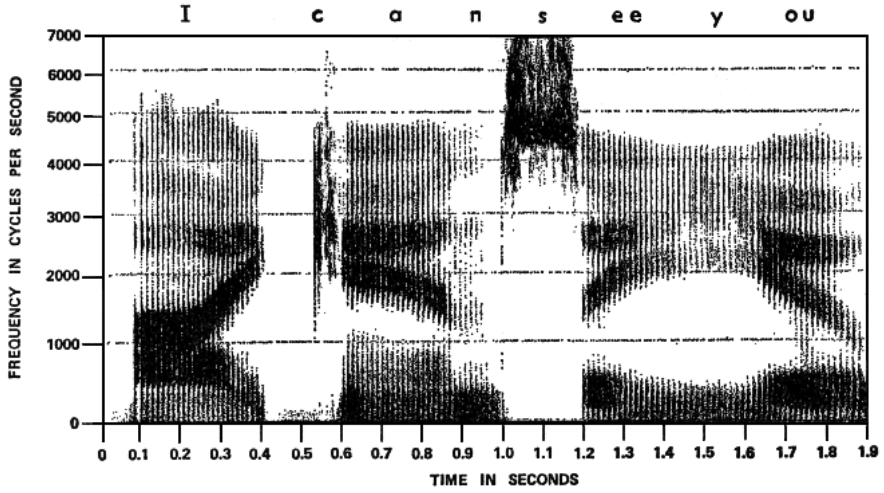



**Figure 2.2 Spectrogram of a speech signal**

### 2.2.2 The Source-Filter Model

Speech production can be modeled by source-filter model [7]. According to this model, voice production is a combination of voice source and acoustic filter that is why it is named ass source-filter model. The "source" refers to the airstream generated by the larynx and the "filter" refers to the vocal tract [7]. They both are inherently time-varying and assumed to be independent of each other.

### 2.2.3 Segmentals and Suprasegmentals

The terms are related to the span of the acoustic analysis. Segmental measurements are done for a short segment of speech, e.g. for a single phoneme. The order of segmental measurements is in milliseconds.

Suprasegmental parameters are also known as prosodic parameters and they are spread over several segments. They are responsible for controlling the intonation, stress, and rhythmic organization of the speech [7].

## 2.3 Perceptual Approach

Perceptual approach is related to how human listening mechanism responds to speech sounds. The sound perception mechanism is referred to as psychoacoustics. This

discipline gives insight to techniques that can be adopted to reduce amount of irrelevant data. Some of the physical attributes and their counterparts are described in Table 2.1.

**Table 2.1 Physical attributes and their perceptual counterparts**

| **Physical Attribute** | **Perceptual Attribute** |
|---|---|
| Intensity | Loudness |
| Fundamental Frequency | Pitch |
| Spectral Shape | Timbre |
| Onset/Offset Time | Timing |
| Phase difference in binaural hearing | Location |

Intensity of the sound is not proportional to Loudness. The relationship between them is defined by Decibel.

The relative amplitudes of different frequencies determine the overall spectral shape [7]. Timbre is perceptual attribute of spectral shape and is known to be an important feature in speaker recognition. For example, the widely used mel-cepstrum feature set measures the perceptual spectral shape. The question for importance of perceptual perception lies in the fact that human ear is the optimal recognizer, which keeps the useful information and discards the unwanted, i.e. higher frequencies.

## 2.4 Speaker Individuality

Speaker individuality is a complex phenomenon which builds up from both the anatomy of the speaker's vocal organs, as well as learned traits. There has been a debate that which among the anatomy and traits is more dependent on speaker. According to Nolan [7], vocal organs are not fixed but they can be changed intentionally which defines a speech signal to be a different biometric because it varies with time unlike fingerprint.

### 2.4.1 The voice source

The larynx is unique for every speaker and the vocal folds of children and females are smaller which results in higher pitch than males. The shape of the glottal pulse affects the overall downward slope of the spectrum but the glottal flow is difficult to

compute. Overall, it appeared that although the glottal features were observed to contain useful speaker-related information, but they are difficult to compute specially in noisy environments.

### 2.4.2 The vocal Tract

The length of the vocal tract differs in a way that if it is assumed that the articulatory configurations of two speakers are the same and the only difference is the length of the vocal tract (measured from glottis to lips), then the acoustic theory predicts that the formant frequencies are inversely scaled by the ratio of the speakers' vocal tract lengths [7]. Oral and pharyngeal parts of the vocal tract are varying from speaker to speaker. Studies show that both the length and the shape of the vocal tract are individual.

## 2.5 Types of Speech Recognition

Speech recognition systems can be classified into numerous types based on a fact that in what particular aspect they are efficient to recognize. This is due to the reason that most of the recognition system cannot judge the start and end of an utterance. Most packages can fit into more than one class, depending on which mode they're using. Some classes of ASR are described below:

**a. Isolated Words**

Isolated word recognizers generally entail each utterance to have silent pause on BOTH sides of the sample window. They require a single utterance at a time. Often, these systems have "Listen/Not-Listen" states, where they require the speaker to wait between utterances. Isolated Utterance can be a better name for this class.

**b. Connected Words**

Connect word systems consent to separate utterances to be 'run-together' with a minimal pause between them.

**c. Continuous Speech**

Continuous recognition is the next step. Recognizers with continuous speech capabilities are some of the most difficult to create because they must utilize special methods to determine utterance boundaries. Continuous speech recognizers let users to speak more or less naturally, while the computer decides the content. Basically, it's computer dictation.

**d. Spontaneous Speech**

Definition of spontaneous speech varies, at a basic level; it is taken as a speech that is natural sounding and not rehearsed. Thus spontaneous speech recognizer must have abilities to cope with the intricacies of natural speech feature

## 2.6 Auditory Perception: Hearing Speech

The ability of human auditory processing system to overcome challenges like speech variability and effect of noise suggests that auditory-based recognition are to superior to systems based on acoustics and signal progressing. Human auditory processing is somehow tuned to speech. The human ear, for example can detect frequencies from 20 Hz to 20,000 Hz, but it is most sensitive to the frequency range critical for speech: 1000 Hz to 6000 Hz [38].

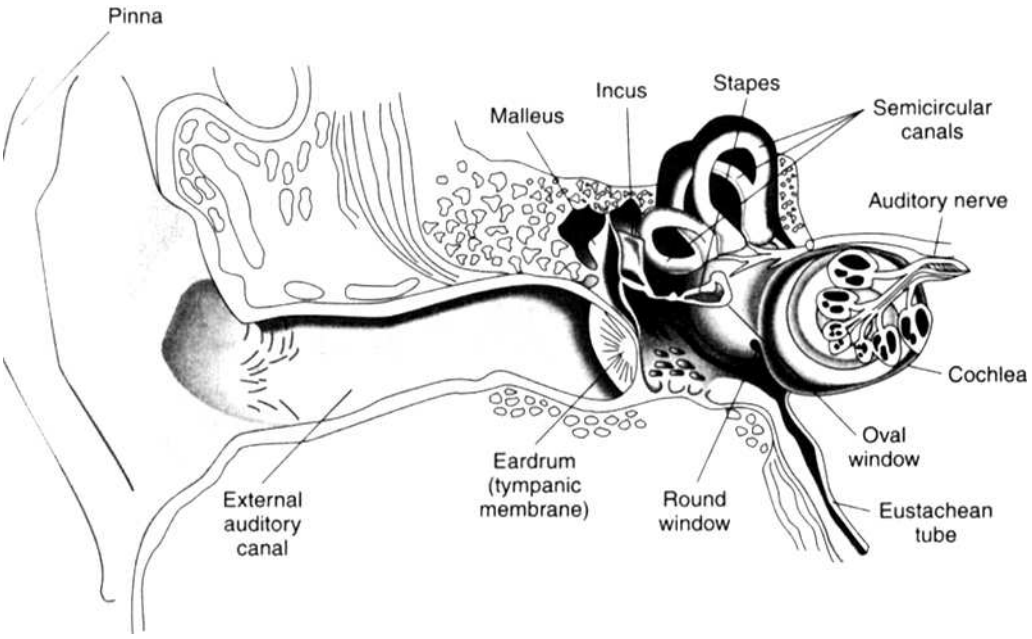



**Figure 2.3 Peripheral Auditory System of Humans**

The major divisions of auditory system include the outer ear, middle ear and the inner ear as shown in Figure. Sound enters through pinna to outer ear which actually localizes the sound. Sound then travels through auditory canal resulting in vibration of eardrum. Eardrum connects the outer ear to middle ear which acts as a transformer to efficiently transport the vibrations to the inner ear. The important part of the inner ear is cochlea, a coiled tube filled with fluid. Vibrations of the eardrum result in

movement of oval window and this window produces compression sound wave in cochlear fluid. This compression wave causes vertical vibration of basilar membrane.

When the ear is excited by an input stimulus, different regions of basilar membrane respond maximally to different frequencies. In other words, frequency tuning occurs in basilar membrane. This response is due to a bank of cochlear filters along basilar membrane. Measurements show a logarithmic increase in bandwidth of these filters. Also, a simple model of the inner ear front-end auditory processing is that of a wavelet transform along the vertically oscillating basilar membrane. This wavelet representation of the cochlear fluid was introduced by Yang, Wang, and Shamma [39].

The cochlear filter bank provides a range of analysis window durations and bandwidths with which to analyze the signal at different frequencies. Also, rapidly varying signal components (for e.g. Plosives) are better analyzed with shorter windows than those of low frequency harmonics.

# Chapter 3
# Principles of Speaker Recognition

---

This chapter includes few of the most important concepts that should be taken in account regarding speaker recognition task before stepping towards the actual implementation phase.

## 3.1 Speaker identification vs. Verification

Recognition is based on either identification or verification. It is a process of determining which speaker, if any, in a group of known speakers, closely matches an unknown speaker. The identification may be closed set, where it is assumed that the unknown is in the set of known speakers; or open set, where the unknown speaker may or may not be in the set of known speakers.

Speaker verification on the other hand, is the process of accepting or rejecting the identity claim of a speaker.

## 3.2 Text- Independent vs. Text-Dependant

The project will include speaker verification phase, which will pursue text independent methods.

Text dependant in which, the unknown speakers must speak the same prescribed text that was used for training and text independent methods in which, speaker models capture characteristics of somebody's speech which show up irrespective of what one is saying. It allows the user to read any text during both training and testing. In general, text dependant methods are more accurate as the recognizer is precise about the text.

## 3.3 Open set and Closed set verification

Speaker identification can either be open set or closed set. If the target speaker is assumed to be one of the registered speakers, this is a closed set problem

whereas if target speaker is not among the one registered; this states to open set problem.

Verification task is the special case of open-set identification with only one speaker in database. Open-set identification is more challenging than closed-set. In closed-set system has to make a compulsion for selecting best matched speaker regardless of the fact how poor the speaker matches whereas in open-set problem, system must have predefined threshold so that the similarity measure between the unknown speaker and best matched speaker must lie within this threshold value.

## 3.4 Normalization and Adaption Techniques

Normalization in speaker recognition is adopted to remove intra speaker variability introduced due to channel or multiple recording sessions. These variations can be raised from speaker himself or from noise as well. To have a speaker recognition system adapt these variations, normalization methods of two types are used, i.e. parameter domain and likelihood/similarity domain.

### 3.4.1 Parameter Domain Normalization

This is also known as blind equalization method which is effective in reducing linear channel effects and long term spectral variations. This method is more effective for text-dependant speaker recognition applications which use long utterances. For this purpose, Cepstral Mean Subtraction (CMS) is used which is fairly an effective method but this method removes some text-dependant and speaker specific features; also it is ineffective for only short term utterances.

### 3.4.2 Likelihood Normalization

The likelihood ratio is the ratio of conditional probability of the claimed identity is correct to the probability of claimed identity being an imposter. A positive likelihood shows a valid claim and negative likelihood shows existence of an imposter. This method is unrealistic due to its large computational cost in calculating all the probabilities. Thus for this purpose, a small set of speakers which are representatives of population distribution near the claimed speaker are chosen for calculating the normalization term.

## 3.5 Types of Distortions in Speaker Recognition

Errors in Speaker Recognition come from different sources. Some are caused by speaker itself and some are due to technical conditions. In general, Distortion can be mentioned along two independent dimensions:

- Deliberate versus non deliberate
- Electronic versus non electronic

### 3.5.1 Deliberate-electronic:

It is the use of electronic scrambling devices to alter the voice. This is often done by radio stations to conceal the identity of a person being interviewed.

### 3.5.2 Non deliberate-electronic:

This includes, for example, all of the distortions and alterations introduced by voice channel properties such as the bandwidth limitations of telephones, telephone systems, and recording devices.

Poor-quality microphones introduce nonlinear distortion to the true speech spectrum [7]. Quatieri [20] demonstrate, by comparing pairs of same speech segment recorded with good- and poor-quality microphones, that poor-quality microphones introduce several spectral artifacts.

If the speech is transmitted through a telephone network, it is compressed using lossy techniques which might have added noise into the signal. Speech coding can degrade speaker recognition performance significantly [20].

Above mentioned parameters such as Environmental acoustics mismatch, mismatch in type and amount of background noise, microphone type mismatch and recording quality mismatch; all correspond to "*Mismatched conditions*". This is recognized as the most serious error source in speaker recognition [7].

### 3.5.3 Deliberate-non electronic:

It includes use of falsetto, teeth clenching, etc.

### 3.5.4 Non deliberate-non electronic:

Alterations that result from some involuntary state of the individual such as illness, use of alcohol or drugs (the effects are involuntary), or emotional feelings [1]. Types of Distortions are shown in Table 3.1

**Table 3.1 Types of Distortions**

| | | |
|---|---|---|
| **Electronic** | Electronic scrambling, etc | ***Channel distortions, etc.*** |
| **Non-Electronic** | Speaking in a falsetto, etc | **Hoarseness, intoxication, etc** |

But speaker verification over telephone network presents the following challenges:

- Variations in handset microphones which result in severe mismatches between speech data gathered from these microphones.
- Signal distortions due to the telephone channel.
- Inadequate control over speaker/speaking conditions. [Speaker Verification over Long Distance telephone]

This project focuses on signal cell in above table, Non Deliberate - Electronic, i.e. distortion in telephone channel only. Other distortion levels will not be treated.

## 3.6 Channel Noise

Telephonic mode of access is used in various areas that include banking, voice activated access in control of data entry in medical or dark room, telephone shopping, voice mail, security control and criminal detection in forensic but this fairly introduces channel noise. Channel noise refers to effect of speech input device on spoken input. These devices are characterized into two

Microphones

Telephones

The spoken word or speech is converted into analog signals before being transmitted via channel. Signal distortion is first introduced by microphones and then channel contributes electrical noise. Designers of speaker-independent systems for use over the telephone are careful to collect speech samples over telephone networks that will

be used by the application [38]. Moreover, technology developers create a separate set of models for land line and for cellular telephones in order to accommodate different channel noise conditions.

### 3.6.1 Microphones

Each brand and model of microphone produces a different and a unique configuration of distortion and additive electronic noise. Since, it is extremely difficult to remove the microphone noise characteristic from the signal; they are generally encoded in reference models of an application or the system [38].

Microphones vary in quality and type. *Omnidirectional* microphones have uniform pickup patterns i.e. they pick up speech from all the directions whereas *directional* microphones are designed to respond to specific direction. *Directional* microphones are best suited for speech and speaker recognition because they can isolate sound coming from specific direction.

Noise cancelling microphones are preferred in highly noisy environments because they cancel the effect of sound coming from distant sources. One extensive experiment performed at IBM in understanding the relationship between microphone type and noise and Directional microphones were found to be less sensitive to background noise of all types. Moreover, Rabiner & Juang, the IBM researchers found that it is not advisable to train with one microphone and use another in this field [38].

### 3.6.2 Effect of Telephone Channel in Speaker Recognition

Various researches shows that performance of speaker recognition can be degraded dramatically when the recognizer is applied in an environment that is different from the environment in which it was trained.

The aim is to gauge the performance loss incurred by transmitting the speech over the telephone network. Factors that add the degradation to speech over a telephonic channel are:

- Band limitations: The Voice limitations from 300-3400 Hz over Telephonic Channel
- Spectral Shaping (filtering): The spectral shaping appears to be from the carbon-button microphone.

- Noise addition: The environmental and channel Noise.

Performance factor can be significantly influenced by the frequency character of the communication channel. It has been reported that the error rate of a speech recognizer can increase from 1.3 to 44.6%when the testing data are filtered by a pole/zero filter modeling a long-distance telephone line [4].

First, there are three types of microphones used in standard telephone equipment:

- Electric speaker phone
- Carbon button handset
- Electret handset

Electret and Carbon button microphones are both

## 3.7 General Structure

General Architecture of Speaker Recognition is shown in Figure 3.1

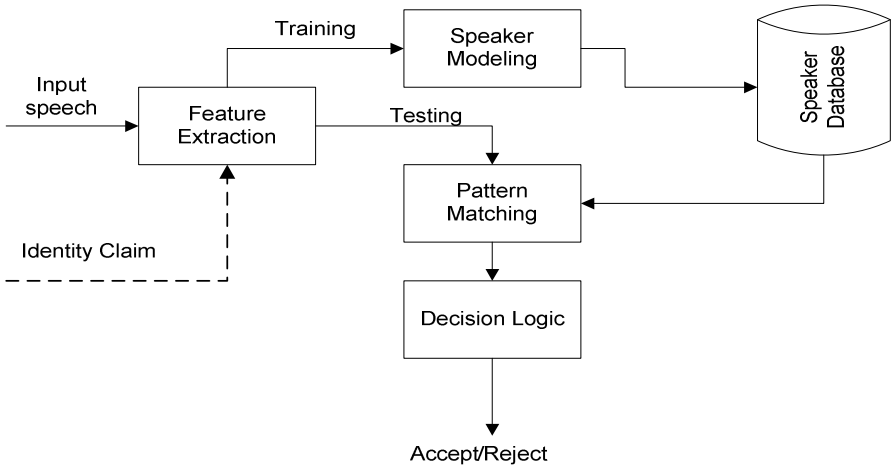



**Figure 3.1 General architecture of TAURUS**

First, speaker is enrolled via recorded speech sample, i.e. non-telephonic and telephonic. Then feature extraction is the first phase carried out. After extracting features, we transform these features to create a model for each speaker and store it. Then comes Patten matching, for each testing speaker, we match the model for the unknown speaker to template we have already stored. Decision is based on how closely model for an unknown speaker matches with the stored ones.

### 3.7.1 Training Phase: Feature Extraction and Speaker Modeling

During the first phase i.e., speaker enrollment phase, speech samples are collected from the speakers, and they are used to train their models. The collection of enrolled models is also called a speaker database.

#### Feature Extraction

After collecting speaker database, Features are extracted. Feature extraction is a phase of acquiring compact and speaker dependant information out of original data. The main reason of this step is to perform data reduction while retaining speaker discriminative information because the amount of data, generated during the speech production, is quite large while the essential characteristics of the speech process change relatively slowly and therefore, they require less data.

Feature Extraction consist of three sub processes:

a. Some form of speech activity detection is performed to remove non-speech portions from the signal. This is known as voice activity detection
b. Next, features conveying speaker information are extracted from the speech. The techniques mostly used for speaker verification are :
    1) Mel-frequency cepstrum coefficients (MFCC)
    2) Linear predictive coding (LPC) (also known as auto regressive modeling or AR modeling).
    3) Linear Predictive Cepstrum Coefficient (LPCC)
    4) Perceptual Linear Prediction (PLP)
c. The final process in feature extraction is some form of channel compensation.
    1) Parallel Model Compensation (PMC)
    2) CMS, Cepstral Mean Subtraction
    3) RASTA filtering
    4) the Gaussian dynamic cepstrum representation

CMS is widely used in speaker recognition tasks for channel compensation and speech enhancement.

## Speaker Modeling

There are two main approaches for estimating the class of feature distributions: parametric (stochastic) and non-parametric (template) approaches [7]. In the parametric approach, a certain type of distribution is tied to the training data by searching the parameters of the distribution that maximize some criterion. The non-parametric approach, on the other hand, makes minimal assumptions about the distribution of the features.

Statistical Modeling techniques that are used for speaker verification are:

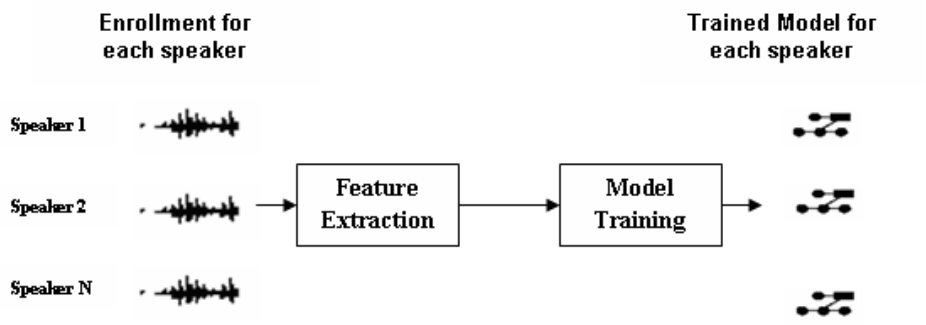
1) Gaussian mixture speaker models (GMM).
2) Vector quantization (VQ).
3) Hidden Markov Models (HMM)
4) Nearest Neighbor (NN)
5) Artificial Neural Network (ANN).

Two approaches widely adopted for text independent speaker verification task are Vector Quantization (VQ) and Gaussian Mixture Models (GMM). VQ is a non-parametric method whereas GMM is a parametric method.

Both methods are used to generate a model for each speaker; which is stored in the database. This process is represented in Figure

## Enrollment Mode 1

Enrollment mode 1 has speakers' data recorded on microphone on computer. The process is shown in Figure 3.2.




1

**Figure 3.2 Speaker enrollment mode 1**

### Enrollment Mode 2

Enrollment Mode 2 is data recording through handset over wireless network. Process is shown in Figure 3.3

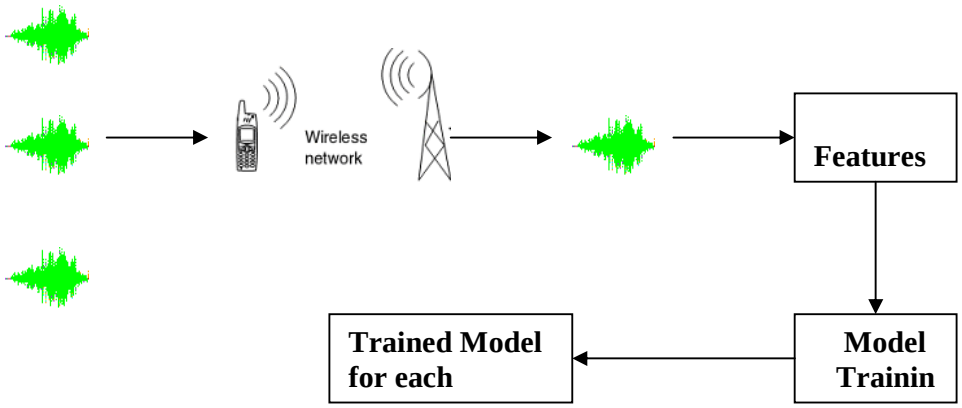



**Figure 3.3 Speaker enrollment mode 2**

## 3.7.2 Testing Phase

In the second phase, speaker verification phase, a test sample from an unknown speaker is compared against the speaker database which contains model set for each speaker.

Feature extraction is the main part of both training as well as testing. Speaker modeling is carried out for an unknown identity and then this model is compared with background models to find the best match, i.e. the process of decision making, which is to simply accept an unknown speaker as of enrolled subjects or to reject it. The process is described in Figure 3.4

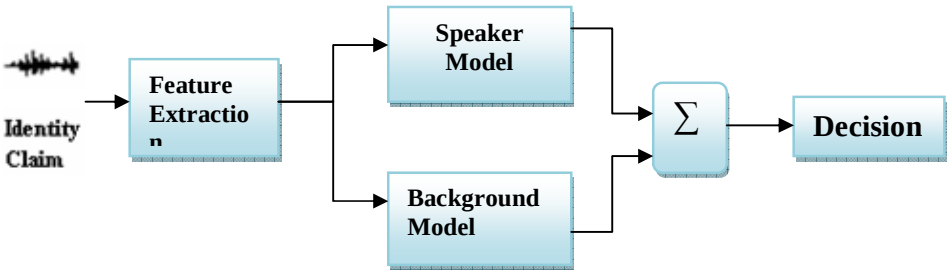



**Figure 3.4 Speaker verification phase**

## 3.8 Applications of Speaker Recognition System

Although any task that involves interfacing with a computer can potentially use SR, the following applications are the most common right now. This system is of great value in the following fields

### Access Restriction

Access restriction is the area in which speaker recognition technology has had the greatest impact. While access to secured areas can be restricted with the use of keys, magnetic cards, and lock combinations, all three can be lost or stolen. Telephonic Speaker recognition can provide an alternative or supplemental means of entry.

### Forensic

The use of telephonic speaker recognition in law enforcement is becoming common place where evidence is in the form of voice recordings of the suspects. Such cases might include bomb threats, ransom negotiations, undercover tape recordings, wire taps, etc

### Computer-human Interaction

There is an increasing need for computer human interaction in the world of information, and in applications ranging from telephones to mobile devices and robotics. Some new cellular phones include C&C speech recognition that allows utterances such as “Call Homes”.

### Industry

In the industries where the need for security is quite real, there real time telephonic speaker recognition may result in great mean of access control.

### Telephony

Some PBX/Voice mail systems allow callers to speak commands instead of pressing buttons to send specific tones.

## 3.9 Literature Survey

This section includes overview to various techniques that have been used and implemented in literature for speaker recognition systems.

### 3.9.1 Voice Activity Detection

Voice Activity Detection is one of the essential steps in speech pre processing. It is a process of extracting speech from a noisy or non speech signal. This aids in reducing the bandwidth requirement in communication as well as helps in speech recognition areas.

Most of the techniques adopted for VAD are based on energy estimation. This is because it is simple and regardless of noise assessment. However; drawback lies in its sensitivity to noise variation statistics. The misfortune is in a fact that energy estimation methods mask the unvoiced speech area with noise. The main problem turns up when one has to compute decision threshold for these methods. Research shows computation of more than one threshold in order to extract the voiced and unvoiced part of speech. But energy based methods are successful only where it is a high signal to noise ratio. Performance measure of these methods degrades dramatically with low SNR.

In order to overcome this scenario, Entropy measure has been adopted for detection of speech. Spectrograms of very noisy speech signal show that speech regions are more organized than noisy ones [21]. An appropriate metric to measure the organization of the signal is Shannon's entropy. Observation shows that approach based on the entropy of the magnitude spectrum of the signal, always outperforms the energy-based method for the estimation of both the means and variances of the noise. It has also been stated that the entropy based approach obtains significantly better performance in non stationary noise like factory noise than energy based method [21].

Minimum phase Group Delay Functions is another approach to automatically segment continuous speech. The algorithm for segmentation is based on processing the short-term energy function of the continuous speech signal which can be processed in a manner similar to that of the magnitude spectrum. This method is used to estimate an effective boundary for speech estimation and error for this has been estimated less than 20% [22].

DWT based VAD has also been adopted in the year 2002 followed by an improved method in the year 2005. By means of the multi-resolution analysis property of the DWT, the voiced, unvoiced, and transient components of speech can be distinctly discriminated. TEO, Teager Energy Operation is then applied to the DWT coefficients of each subband. Experimental results show that the proposed method can extract the speech activity under poor SNR conditions and that it is also insensitive to variable-level of noise [23].

Recent research on other hand uses hard c-means clustering for voice activity detection. Moreover Self-organized Feature Mapping (SOM) and Learning Vector Quantization (LVQ) network have also been used in recent years in order to detect voice activity.

### 3.9.2 Speech Enhancement

Speech enhancement in past few years has been used to lessen the additive background noise in speech. This noise particularly reduces the quality of speech; by quality it means intelligibility or pleasantness of speech signal.

Research shows that there is no philosopher's stone discovered for noise removal in real world communication problems. The central methods of speech enhancement are:

- removal of background noise
- echo suppression
- process of artificially bringing certain frequencies into the speech signal.

All the speech enhancement methods aimed at suppressing the background noise are (naturally) based in one way or the other on the estimation of the background noise [27]. If noise is more stationary than speech, it is easy to estimate it in intervening pauses.

The oldest and simplest methods in removal of noise are spectral subtraction and wiener filtering. Weiner filtering is considered as a better approach than spectral shaping because spectral shaping results in "musicality". This musicality is caused by rapid coming and going of sine waves over successive frames [29]. To avoid these annoying fluctuations within speech, wiener filtering is used to smooth the speech signal.

Other approaches in Literature that are adopted in order to compensate channel in telephone lines are Cepstral Mean Subtraction, RASTA filtering, Gaussian Dynamic Cepstrum Representation, GDCR.

Referring to [30], it has been shown that CPS and RASTA outperform GDCR. Morever, CPS has been proved superior to RASTA filtering due to phase distortion introduced by RASTA filter. Studies show that RASTA filtering can be improved by its phase correction but phase corrected RASTA has similar performance to that of CPS.

According to [31], another approach has been found useful in reduction of word error rate. This approach suggests using combination of both CMS, as well as vocal tract length normalization (VTLN) for speaker normalization.

Some other set of experiments on Cepstral Mean Subtraction (CMS) and RASTA-filtering were carried out in Phoneme Recognition [32]. The CMS-method has been reported to increase the performance of Automatic Speaker Recognition over Telephone.

### 3.9.3 Feature Extraction

Feature extraction is important phenomenon in order to reduce the effect of curse of dimensionality as well as increases computational complexity.

There are two techniques for carrying out feature extraction over a frequency band.

- Full band technique
- Multiband technique

A full band technique is a conventional technique in which speaker recognition is performed by recognizing an extracted set of feature vectors over a full frequency band of input speech signal. The major drawback of this approach is that even partial band-limited noise corruption affects all the feature vector components. The multiband approach deals with this problem by performing acoustic feature analysis independently on a set of frequency sub bands. The motivations for multi band technique in speaker recognition are as follows [4, 5]:

- Human speech perception is multiband in nature
- If speech is corrupted by some additive noise, a band-limited noise signal does not spread over the entire feature space
- It is suited for parallel architectures

The multiband technique used in this project is via Discrete Wavelet Transform as a tool for decomposing the speech signal into different sub bands. The main reason that supports the use of DWT is that speech is a non stationary signal. Moreover time-frequency resolution characteristics of DWT resemble with those of human ear.

## Recognition Rates by using Dwt based Features

Table 3 shows the accuracy achieved by applying features based on DWT in Speaker Recognition field. Research shows that DWT based feature extraction strategy is being experimented in late 1990's -2008. Few of the Results are summarized in Table 3.1

**Table 3.2 DWT Feature Based Recognition in Literature**

| | Technique | Accuracy | Source |
|---|---|---|---|
| **Speaker Recognition in Noisy Environments** | LFCCs (dwt based) - HMM | 99.2% | W.Khaldi,W.Fakhar, W.Hamdy in 2002 |
| **Multiband Approach to Robust Features in SI** | LPCCs – FCGMM | 94.96% (with band level 3) 88.07% (with band level 4) 88.04% (with full band) | Wan-Chen Chen , Ching-Tang Hsieh , and Eugene Lai in 2004 |
| **Robust Speech Features with application to Speaker Recognition** | LPCCs-VQ(LBG) | 96.68% | Wan-Chen Chen , Ching-Tang Hsieh , and Eugene Lai in 2002 |
| **De-noising with Novel DWT-MFCC for Speaker Recognition** | MFCC-GMM | 95.8% (Matched Conditions) 80.1% (mismatched conditions) | Zhengquan Qiu1 , Junxun Yin in 2006 |

Kinnumen Tomi [7], has made a good comparison for these features. He divided the most commonly used features in speaker recognition into following sets:

- Filterbanks
- FFT-cepstral features
- LPC-derived features

- Delta features

All of the stated feature sets represent somewhat same information but some of them are better than others because they tend to be more robust to quantization and Euclidian distance classifier. Studies show that LPC features do not give satisfactory results when Vector Quantization is used as classifier with Euclidian distance measure. LPCCs are far most efficient than other feature extraction techniques. The two techniques that have been reported to give good results are LPCCs and LSF, Line Spectral Frequencies. LPCCs due to their better performance and LSF due to its good quantization properties it is well suited to Vector Quantization modeling. However, there is a debate which of the two is better. LPCCs are considered to be more efficient whereas LSFs as more accurate.

According to [7], Mel-Cepstrum is the most commonly used feature set but it may not give best cepstrum representation. The mel-scale simulates human hearing, but there is no reason to assume that the human ear resolves frequency bands optimally in respect to the speaker recognition task [7].According to [33], and many of other research studies, high and medium frequencies are more important in speaker recognition than low frequencies which is generally a contradiction to mel-cepstrum which focuses low frequencies.

Next come the delta features. Literature shows that static feature set gives better performance than dynamic. Regression methods have been reported better than differentiation methods. Regression Analysis highly depends on number of frames used in derivative estimation and that of the original feature set. If the order is high (long temporal span), rapid spectral changes that might be speaker dependent, are smoothed out; On the other hand, if the order is low (short temporal span), it is likely that the feature trajectories become noisy [7].

### 3.9.4 Classification

Classification deals with recognition of an unknown speaker based on pattern matching Pattern Matching depends upon the similarity measure between an unknown speaker and speakers that are enrolled. There are number of pattern matching algorithms which include Gaussian mixture speaker models (GMM), Vector quantization (VQ), Hidden Markov Models (HMM), Nearest Neighbor (NN) and Artificial Neural Network (ANN) are most popular in speaker recognition. *Classifier ensembles* (committee classifiers) have become also popular in the past few years.

Many of combination strategies have also been introduced. The basic idea behind them is to model each feature set with the modeling technique best suited for it and then to combine the final score. An overview and comparison of several combinations strategies for speaker recognition is given in [35]. Neural Networks have an advantage that they use less parameter but they have a profound drawback; the network is needed to be trained every time the new speaker has to be added which effects system efficiency and performance. Similarly, HMMs have also been found to be computationally expensive.

GMM (Single State HMM) and VQ are widely used and accepted techniques in Pattern Matching. GMM is parametric (stochastic) technique], which uses K Gaussian distributions [26] whereas VQ is non- parametric (template) technique which use VQ codebooks consisting of a small number of representative feature vectors [25].

GMM algorithm is computationally complex as compared to VQ but research proves it to be better than VQ. VQ approach has no computational complexities but its downside lie in non-overlapping clusters which results in its discontinuity. On the other hand GMM forms overlapping clusters and known to be extension of VQ.

In [26], author has proved GMMs to be best classifier in Robust Text Independent Speaker identification as it outperforms other techniques.

Lately another approach described in [24], makes use of combining both VQ and GMM for text dependant Speaker Recognition over Telephone channels.

In general, GMM has high success rates as Pattern Patching technique in Speaker Recognition followed by Vector Quantization.

# Chapter 4
# Feature Extraction and Classification Technique

In chapter 3, we have seen that there are number of techniques that exist for each phase of Speaker Recognition but all of them do not lie in scope of this project.
In this project, few of the most suited and widely used techniques have been applied and their comparison has been made. This chapter includes detail description of each technique that has been implemented.

## 4.1 Speech Pre-Processing

In speech Pre-processing, voice activity detection is being carried out followed by channel normalization technique Cepstral mean Subtraction.

### 4.1.1 Voice Activity Detection, VAD

Voice Activity Detection is used to separate speech and non-speech data from a speech signal. Non- Speech data is pre utterances, post utterances and silence between words [21]. Voice Activity Detection includes two stages:

- **Parameter extraction:** Significant parameters are extracted from the speech signal. Parameters containing discriminative variation allow a good detection of the speech regions
- **Thresholding:** A threshold level is applied split signal into speech non-speech segments. This threshold can be fixed or adaptive.
  For a robust voice activity detector, this threshold value must be adaptive.

**VAD based on energy Thresholding**

As described in chapter 2, VAD algorithms are extensively based on energy estimation phenomenon due to its simplicity. However, these methods are sensitive to noise to a great expense.

First, the signal is being framed and short time energy of each frame is then computed. The energy of *mth* frame of length is given by

$$E(m) = \sum_{n=m.N}^{m.N+N-1} y^2(n) \qquad (4.1)$$

Where *n* is the time index.

Generally, two thresholds are used to form a hysteresis, to avoid switches when the energy level is near to the threshold [21]. Let these values be $\lambda_1$ and $\lambda_2$ . $\lambda_1$ is threshold estimation in speech segments whereas $\lambda_2$ determines threshold level in noisy segments. When the energy of a speech signal is greater than the speech threshold, speech is detected and when the energy is lower than the noise threshold speech pause is detected. The use of two thresholds defines a hysteresis and reduces the problem of fast changes in the detection which are obtained if a single threshold is used [21].

This Energy based method works effectively in areas where noise varies slower over the signal and the speech segment energy is considered greater than noise level.

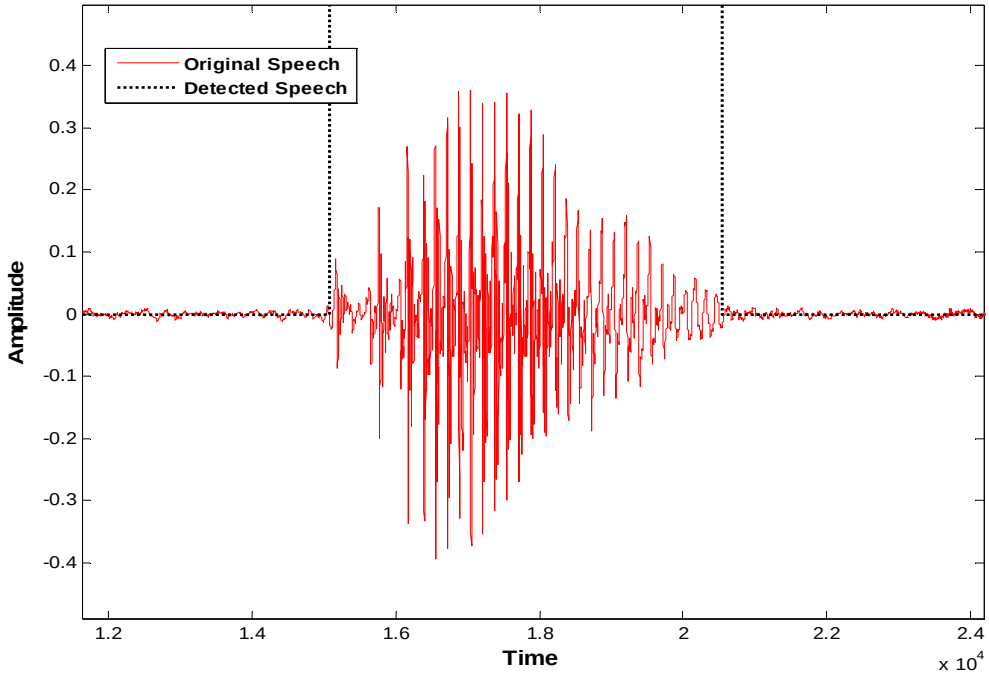



**Figure 4.1 Voice Activity Detection**

Figure above shows the effect of voice activity detection which removes the noise and pause in original speech, keeping only the spoken words.

### 4.1.2 Speech Enhancement: Cepstral Mean Subtraction

Acquaintance of environmental robustness is a very important issue in speech recognition. As this project aims to telephonic mode of speaker recognition, channel normalization is an important step in preprocessing.

Normalization method influences the signal such that the important information is retained and unwanted information is cancelled out. Widely adopted technique by these models is Cepstral Mean Subtraction, CMS.

Convolutional Distortion becomes multiplicative in spectral domain and additive in log spectral domain, when a signal passes through a linear time invariant channel [28]. This log spectrum can be treated same as cepstrum because cepstrum is linear transformation of log spectrum. Hence it is assumed that this property of spectral noise is true in a short time estimate of a signal as well. Measured spectrum of speech after short time analysis is given by $Y_t(w)$

$$Y_t(w) = C(w).S_t(w) \tag{4.2}$$

And cepstrum is given by

$$y_t = c + s_t \tag{4.3}$$

Where C(w) corresponds to constant channel , St(w) is speech spectrum and t shows time independence. It is supposed that channel is constant; therefore, if a mean of measured spectrum is subtracted from it, the effect of channel can be compensated.

This subtraction is known as cepstral mean subtraction and the obtained result is cepstral mean subtracted feature zt.

$$z_t = y_t - \overline{y_t} = c + s_t - (c + \overline{s_t}) = s_t - \overline{s_t} \tag{4.4}$$

Where $\overline{y_t}$ is mean of cepstrum.

Thus, this technique has an advantage of simplicity because it doesn't require signal assumptions but only mean noise power estimation. But this is weakness as a same time due to the fact speech mean $\overline{s_t}$ is also subtracted; resulting in loss of local static components of speech signal.

## 4.2 Feature Extraction

This project focuses the use of dwt based LPCs, MFCCs and LPCCs. This section will briefly describe the use and description of these feature extraction techniques.

### 4.2.1 DWT – An Introduction

A discrete wavelet transform (DWT) is any wavelet transform for which the wavelets are discretely sampled. It is easy to implement and reduces the computation time and resources required.

The DWT of a signal x is calculated by passing it through a series of filters. First the samples are passed through a low pass filter with impulse response g resulting in a convolution of the two:

$$y[n] = (x * g)[n] = \sum_{k=-\infty}^{\infty} x\,[k]\,g[n-k] \qquad (4.5)$$

The signal is also decomposed simultaneously using a high-pass filter *h*. The output gives the detail coefficients (from the high-pass filter *h*) and approximation coefficients (from the low-pass *g*). The two filters are related to each other and they are known as quadrature mirror filters.

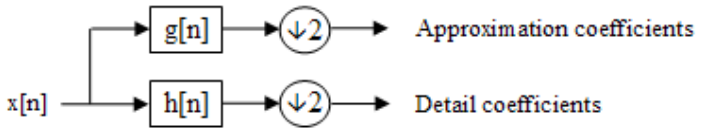



**Figure 4.2 Single level wavelet decomposition**

Since half the frequencies of the signal now been removed, half the samples can be discarded according to Nyquist's rule. The filter outputs are then subsampled by 2. [3] Approximations give characteristics of lower frequencies whereas details give information about higher frequency characteristics.

The Approximation coefficients at each level can be used for another level decomposition and this can be extended to multiple levels. Figure 3.2 shows decomposition up to three levels.

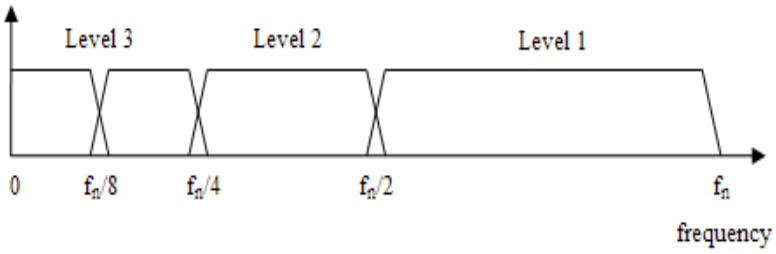



**Figure 4.3 Frequency domain representation of DWT**

There are different types of wavelet transforms that can be used like Daubechies, Symlets, Haar, Discrete Meyer and etc. Daubechies-3 has been used in this project.

### 4.2.2 LPC- An Introduction

LPC is based upon principles that have been derived from basic principles of sound production. Whenever a buzz is produced by buzzer placed at the end of tube, glottis produces that buzz. This buzz is actually characterized by its intensity and frequency.

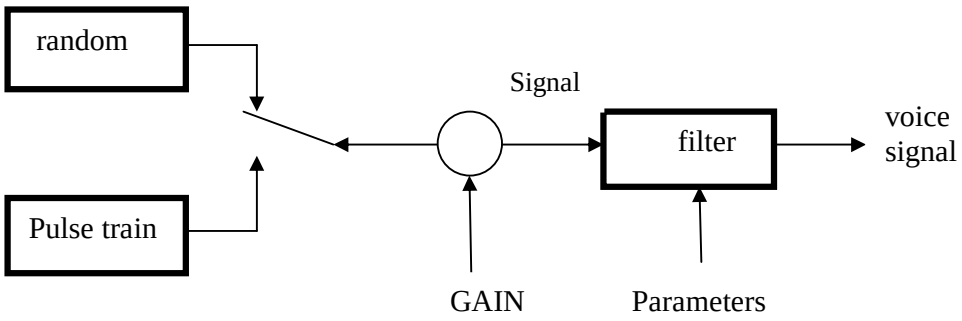



**Figure 4.4 LPC model**

LPC model assumes that speech signal can be modeled by an excitation, passed through a time varying all-pole filter derived from speech produced by human voice box. The model used in LPC for speech generation is shown in Figure 3.4

LPC analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal is called the residue [7].

The principle of a linear predictor is that, it predicts each samples of speech wave from as a linear sum of the past samples.

$$x[n] \approx \sum_{k-1}^{p} a[k]x[n-k] \tag{4.6}$$

Where $p$ is the order of predictor. The aim of LP is to find predictor coefficients keeping the average prediction error minimum.

The prediction error of nth sample is given by

$$e[n] = x[n] - \sum_{k-1}^{p} a[k]x[n-k] \tag{4.7}$$

This equivalently gives,

$$E = \sum_{n} e[n]^2$$

$$= \sum_{n} \left( x[n] - \sum_{k-1}^{p} a[k]x[n-k] \right)^2 \tag{4.8}$$

When prediction error is small, approximation of signal *x [n]* is performed well. The problem of finding the optimal predictor coefficients results in solving of so-called (Yule-Walker) Auto Regression (AR) equations [8].

According to [10], there are two methods for solving AR equations, autocorrelation method and covariance method. Out of these two, autocorrelation is preferred because it gives somewhat stable filters and is computationally efficient [10]. AR equations for autocorrelation method are:

$$Ra = r \tag{4.9}$$

Where $R$ is Toeplitz matrix, $a$ is the vector of the LPC coefficients and $\boldsymbol{r}$ is the autocorrelation. The autocorrelation sequence is given by

$$R[k] = \sum_{n-0}^{N-1-k} x[n]x[n-k] \tag{4.10}$$

The redundancy of AR equations can efficiently be computed by Levinson Derbin Regression [10].

Any signal can be approximated with the LP model with an arbitrary small prediction error 0. The optimal model order depends on what kind of information one wants to extract from the spectrum. The main purpose of using LPC within all approximation channels (obtained from DWT) in order to capture characteristics of individual speaker is that these parameters give good representation on the envelope of speech spectrum of vowels. Moreover they are simple [6].

### *Steps in Computing LPC*

In spectral estimation, an AR system driven by a white noise is used to model a wide sense stationary random signal. In speech coding, the driving signal (excitation signal) is instead a quasi-periodic impulse train. However, we can still use the Yule-Walker equations to estimate the coefficients *a (k).*

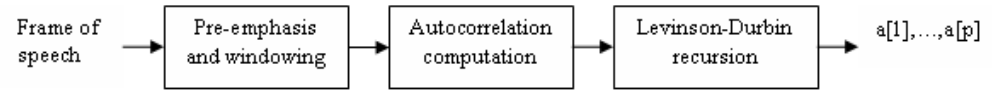



**Figure 4.5 LPC computation steps**

LPC Speech compression consists of steps mentioned below and shown in Figure 4.5.

1. Segment the sampled speech signal into short intervals (10-30 milliseconds long). These segments are called frames and can be overlapping or non-overlapping.
2. Compute the autocorrelation of the frame
2. For each frame, compute the LPC parameters (*a (k)* for $1 \leq k \leq p$) from the data. This can be done by solving the Yule-Walker equations or AR(Auto Regression) equations, or by other related methods.
3. Compute the excitation signal.
4. Model the excitation with a small number of parameters (its pitch and amplitude during the frame).
5. Quantize and code the parameters:
    (a) the LPC coefficients (*a(k)* for $1 \leq k \leq p$)
    (b) the parameters of the excitation signal
    (c) the parameters of the secondary excitation signal

## Time Windows in Linear Prediction Analysis of Speech

Windowing is done in order to get frames of data. Each frame is then used in calculation of autocorrelation sequence.

In frequency domain, the effect of windowing can be seen as a convolution of frequency response of signal with frequency response of window. Important is, the

selection of width of a window because convolution smears frequency features and they are dependent upon width of the main lobe of window frequency response.
One important factor is window length, In Speech Analysis, a window length of 30 ms (240 samples at a sampling rate of 8 kHz) has been found to be a reasonable compromise in terms of the dynamics of speech production [16].

### 4.2.3 MFCC- An Introduction

Mel-cepstrum is one of the most commonly used feature extraction technique used in both speech and speaker recognition [7].
MFCC technique is based on the known variation of the human ear's critical bandwidth frequencies with filters that are spaced linearly at low frequencies and logarithmically at high frequencies to capture the important characteristics of speech [12].
MFCC is composed of five phases. First phase is of framing. Speech waveform is divided into more or less frames of 30 milliseconds. The next step involves windowing of each frame. This minimizes the discontinuities at start and end of each frame. Then windowed speech signal is converted from time domain to frequency domain by taking FFT. Once converted to frequency domain, signal is passed through Mel-frequency wrapping block. The purpose of the mel-bank is to simulate the critical band filters of the hearing mechanism. The filters are evenly spaced on the mel-scale, and usually they are triangular shaped [7]. This can be viewed in Figure 3.4.

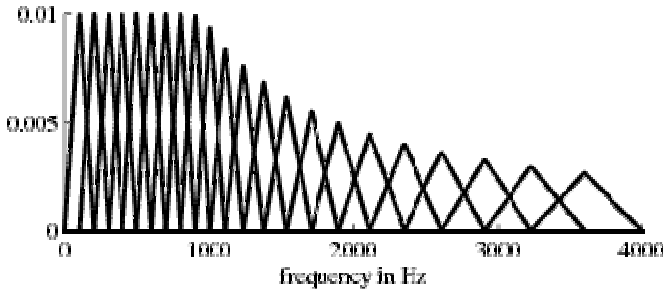



**Figure 4.6 Mel-spaced filter bank**

The main purpose of the MFCC processor is to mimic the behavior of the human ears. Studies have shown that human hearing does not follow the linear scale but rather the Mel-spectrum scale which is a linear spacing below 100 Hz and logarithmic scaling

above 100 Hz [18]. In the final phase, The triangular filter outputs Y (i), i = 1, . . . ,M are compressed using logarithm, and discrete cosine transform (DCT) is applied [7]:

$$c[n] = \sum_{i=1}^{M} \log Y(i) \cos\left[\frac{\pi n}{M}(i - \frac{1}{2})\right] \qquad (4.11)$$

The resultant matrices are referred to as Mel-Frequency Cepstrum Coefficients. This spectrum provides a fairly simple but unique representation of the spectral properties of the voice signal.

Important property of cepstral coefficients is that they are fairly uncorrelated with each other [7].

### *Steps in Computing MFCCs*

The steps in computation of MFCCs are shown in Figure 4.7 below.

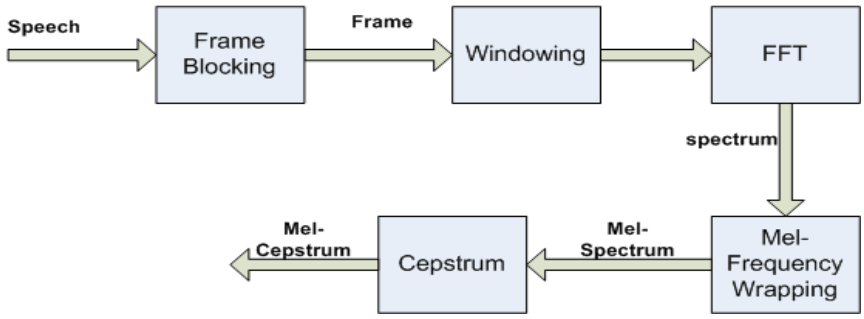



**Figure 4.7 Steps in computing MFCCs**

## Step 1: Framing of Speech Signal

Speech signal is due to its non stationary nature is divided into frames. Figure 4.8 shows the original speech signal. Frame size taken is 256 samples.After framing, Figure 4.9 shows the framed speech signal.

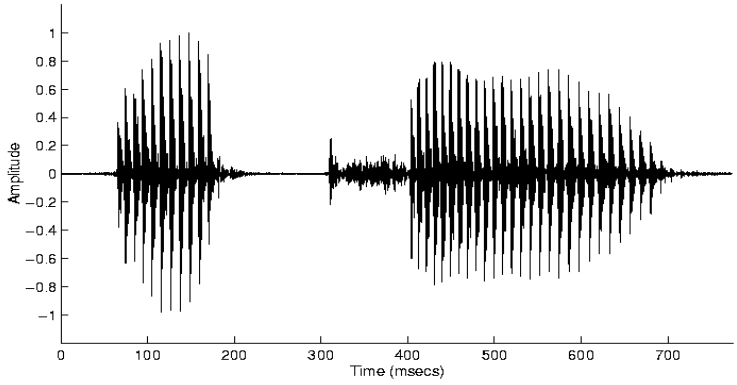



**Figure 4.8 Original Speech Signal in MFCC**

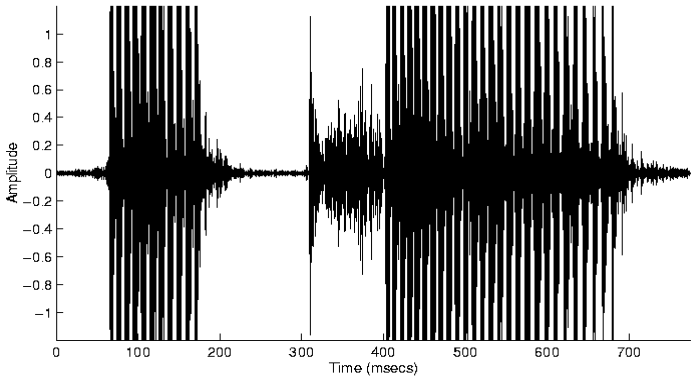



**Figure 4.9 Framed Speech Signal in MFCC**

## Step 2: Windowing

Individual frames are windowed in order to minimize signal discontinuities and spectral distortion. The Hamming window (shown in Figure ) of size 256 is used to decrease the signal to zero at the beginning and end of each frame.

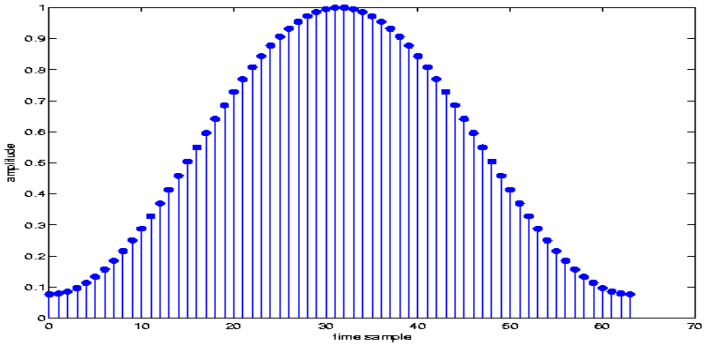



**Figure 4.10 Hamming window**

Figure 4.10 shows hamming window applied to each frame and Figure 4.11 shows windowed speech signal.

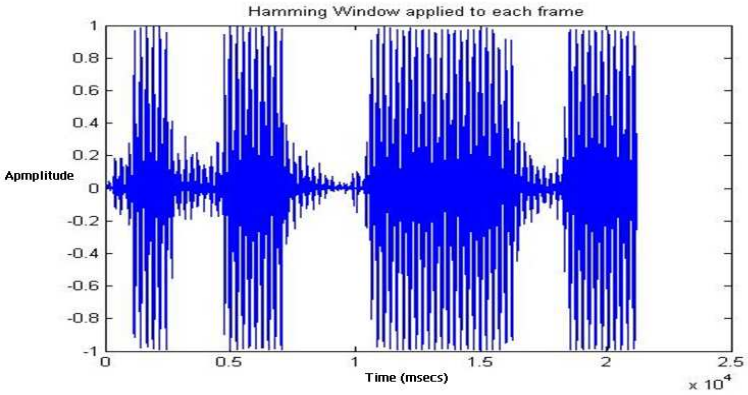



**Figure 4.11 Windowed Speech Signal**

## Step 3: Fourier Transform of Windowed Speech Signal

FFT is used to convert each frame of N samples from the time domain into the frequency domain.512-point FFT is used to give a spectrum which gives information about the frequencies present in a signal. Signal after taking FFT is shown in Figure 4.12.

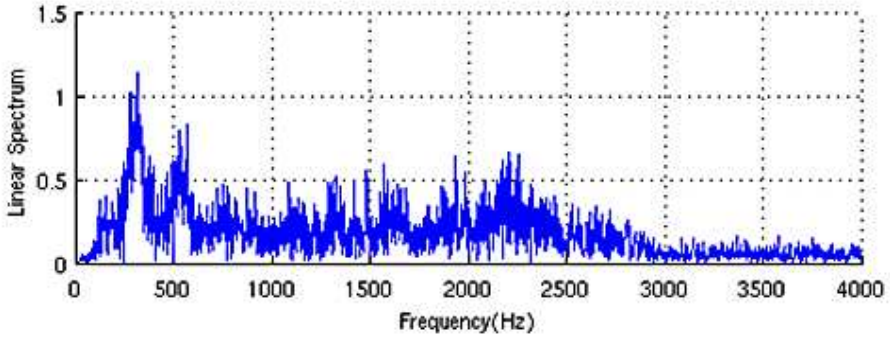



**Figure 4.12 Speech signal after taking FFT**

## Step 4: Mel-Frequency Wrapping

Signal is passed through Mel-frequency wrapping block which emphasizes low frequencies and de-emphasizes high frequencies similar to behavior of human ear.

## Step 5: Discrete Cosine Transform (DCT)

In this final step, log of spectrum is taken to overcome dynamic ranges and after that DCT is taken to compress the information in spectrum.

## Importance of MFCC Derivatives

While extracting MFCCs, it had been assumed that each spectral vector is representation of stationary signal. But it is not absolutely stationary as the articulators are continuously changing their positions with a certain rate. These variations are reflected as changes in formant frequencies and bandwidths [7]. Some of these spectral dynamics left ignored, do contain indicators of speaker itself.

Thus, encoding dynamic information of spectral features can result in improved performance of speaker recognition. Use of this dynamic information is known as delta – features [7]. First, the time derivatives of feature vectors are estimated and then appended with actual feature vectors which lead to higher dimensional feature vectors. Often, the time derivatives of the delta features are also estimated, yielding so-called delta-delta parameters. These parameters are again appended to previously

appended feature vectors, resulting in even higher dimensional feature space than before.

Now there are two arguments regarding this. First, feature space formed by concatenation of static and dynamic features leaves no interpretation; Second, Dimensionality of feature space has increased twice or thrice which needs more training data in order to get reliable speaker models.

As far as computation of these delta features is concerned, there are two methods:

- Differentiating
- Fitting a polynomial expansion

Differentiating is simple but as it serves as high pass filtering method, it somehow enhances the effect of noise. Hence fitting a polynomial is a better approach which is known as regression analysis in statistics.

It has been noted that among both of these methods, regression analysis gives smoother estimates.

On the whole, these derivatives when appended to feature vectors give information about the dynamics of spectral parameters and that of the speaker but on the expense of a large dimensional feature set.

### 4.2.4 LPCC, Linear Predictive Cepstral Coefficients

LPC coefficients are observed to be highly correlated that is why they are used very seldom as features [7]. Their efficiency can be increased when they are replaced with less correlated set of features. A well known feature set in this regard is linear predictive cepstral coefficients.

Given LP coefficients, $a_k$, Cepstral coefficients are computed from following equation [7]

$$c[n] = \begin{Bmatrix} a[n] + \sum_{k=1}^{n-1} \frac{k}{n} c[k]a[n-k], 1 \le n \le p \\ \sum_{k=n-p}^{n-1} \frac{k}{n} c[k]a[n-k], n \succ p \end{Bmatrix} \quad (4.12)$$

LPC cepstrum has been used effectively in both speech and speaker recognition. The thing that needs consideration is that LP coefficients are finite in number whereas

LPC cepstrum sequence is infinite. The magnitude of *c[0]* approaches to zero fast, thus it needs a very small number of coefficients to form a model.

When performance of LPCC parameters is being compared to that of LPC, LPCC cepstral coefficients perform better.

## 4.3 Speaker Modeling

There are various classification techniques as mentioned in previous section. But we have used Vector Quantization based on Minimum distance classifier for classification. Moreover, Open-Set classification has also been performed based on threshold measure.

### 4.3.1 Vector Quantization- An Introduction

Ideally, as much feature information used in matching or classification, more is the chance to obtain accuracy but storing all that information is not realistic. Vector Quantization is a technique to compress information (feature space) in such a way that it maintains the most important or prominent characteristics. It is used to map vectors from a vector space (of test samples) to fixed regions in that space. These regions are called clusters and represented by their central vectors or centroids and a set of centroids, which represents the whole vector space, is called a codebook. This codebook is generated via VQ. Figure 3.5 shows an example diagram of VQ codebook.

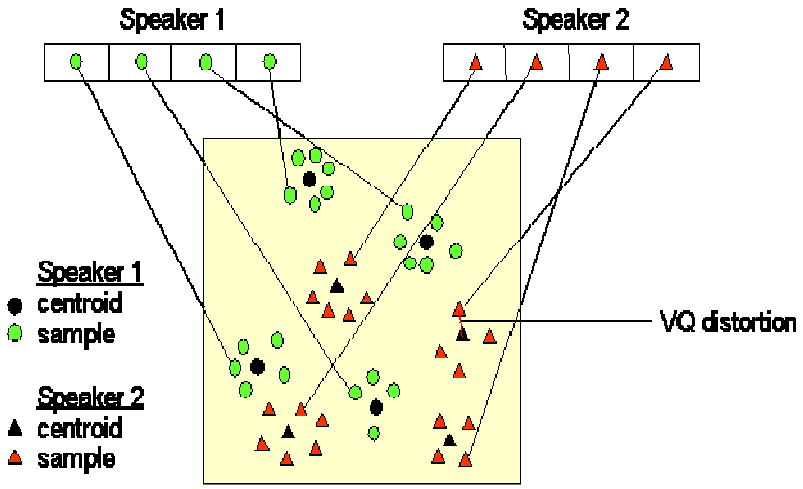



**Figure 4.13 Conceptual diagram of VQ codebook [13]**

There are two important factors that must be considered while implementing VQ.

i. Algorithm to generate codebook
ii. Size of codebook

**Algorithm to generate codebook: K-means Clustering**

K-means clustering is an algorithm to group subjects/objects based on attributes/features into K number of groups (K is positive integer number). The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid [18]. Thus the purpose of K-mean clustering is to classify the data. Steps in K means clustering are given below

**Step 1**. Select k, number of clusters

**Step 2**. Put any initial partition that classifies the data into k clusters either by randomly assigning the samples or taking first k training samples as single-element clusters.

**Step 3**. Take each sample in sequence and compute its distance from the centroid of each of the clusters. If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample [18].

**Step 4.** Repeat step 3 until convergence is achieved or there is no new assignment to be made.

**Size of Codebook in Speaker Recognition**

Codebook size is a tradeoff between time and accuracy. Size of codebook in most of the cases is dependent upon the data we want to classify. Efficiency by greater size of codebooks can be increased but at cost of time and computational efficiency.

Researches show that incase of Speaker recognition task , a codebook of size 32 is optimum. Reason behind this selection is that at approximate level there are 60 phonemes in English Language. Out of these 60, 30-40 can be uttered by a speaker on average. So in few of the experiments as in [6], codebook of size 32 has been suggested optimum.

## 4.4 Classification Technique

There are various classification techniques as mentioned in chapter 3, but in this project, Minimum distortion measure is used.

### 4.4.1 Minimum Distortion Classifier

In testing phase we need to match a testing utterance which is simply a set of non compressed features, to stored models of all the speakers. This matching is done using distortion measure [4] given in equation (2).

$$D = \frac{1}{N_F} \sum_{i=1}^{N_F} \min_{1 \le n \le N_{book}} [d(\overline{x_n}, x_i)] \tag{4.13}$$

Where $N_F$ denotes the number of speech frames in testing utterance, $x_n$ denotes nth reference codebook , $x_i$ is the testing utterance, $N_{book}$ represents total number of codebooks, $d$ is the distance between $x_n$ and $xi$ and $D$ is the overall distortion measure between test utterance and model. The identity of each speaker is then established according to average distortion $D$ for each speaker [4] and the matching speaker is the one who has minimum distortion measure.

### Open-Set Classification using threshold mechanism

Decision making in Classification step can either be closed-set or Open-set. We assume to have a speaker database S containing speaker model of $N$ Speakers such that $S = (S_i, S_{i+1}....S_N)$.

Feature extraction and modeling is same for both unknown and enrolled speaker. In distance based classifiers, it depends on score or distance, how much an unknown speaker $X$ matches with that of $S$. Greater the score, much higher is the similarity measure between the unknown speaker and one of the Speakers in $S$.

In closed-set speaker identification task, the decision is simply the speaker index which gives maximum score.

$$i^* = \arg\max_j score(X, S_j) \tag{4.14}$$

Verification can be performed by setting verification threshold. Let $\theta_j$ be the verification threshold. This threshold defines the boundary for speakers that lie within the database and those who are unknown i.e. to determine *false acceptance rate* and *false rejection rate.* The former means accepting an impostor speaker and the latter means rejecting a true speaker. There is a trade- off between the two errors, when the decision thresholds $\theta_j$ are increased, false acceptance error decreases but false rejection error increases, and vice versa [7].

This decision process can be explained by the following equation

$$score(X, S_j) \begin{Bmatrix} \geq \theta_j, accept \\ \leq \theta_j, reject \end{Bmatrix} \tag{4.15}$$

Open set identification task is defined as follows [7]

$$decide \begin{Bmatrix} i^*, if, i^* = \arg\max_j score(X, S_j) \wedge score(X, S_{i^*}) \geq \theta_j \\ none, otherwise \end{Bmatrix} \tag{4.16}$$

In this case we have two checks in order to characterize the speaker as true speaker. Best matching speaker against *X* is being measured and then this score is being matched with decision threshold. If the speaker is above decision threshold it is accepted as true speaker else it is characterized as an imposter.

# Chapter 5
# Experimental Results and Analysis

As speech interaction with computers becomes more persistent in activities such as financial services and information retrieval from speech databases, the utility of automatically recognizing a speaker based entirely on vocal characteristics increases. Given a speech sample, speaker recognition is concerned with extracting clues to the identity of the person who was the source of that utterance.

There have been numerous approaches aimed at understanding the underlying process involved in the perception and production of speech. These approaches involve disciplines as diverse as pattern classification and signal processing to physiology and linguistics. The interdisciplinary nature of the problem is one thing that makes speech recognition such a complex and fascinating problem. This chapter gives the detailed description of the database which has been made for speaker recognition system for evaluating the various algorithms. It also contains the experiments performed on database to check the performance of classifiers in real time application.

## 5.1 An Overview of Standard Speech Corpora

There are numerous corpora for speech recognition. The most popular bases are: *TIMIT* and its derivatives, *Polycost*, and *YOHO*.

### 5.1.1 TIMIT and Derivatives

The ***TIMIT*** corpus of read speech has been designed to provide the evaluation of automatic speech recognition systems [3]. Although it was primarily designed for speech recognition, it is also widely used in speaker recognition studies. It contains 630 speakers' voice messages (438 M/192 F), and each speaker reads 10 different sentences. It is a single-session database recorded in a sound booth with fixed wideband headset. The derivatives of *TIMIT* are: *CTIMIT*, *FFMTIMIT*, *HTIMIT*, *NTIMIT*, *VidTIMIT*. They were recorded by playing different recording input devices, such as telephone handset lines and cellular telephone handset. It was recorded into 3 sessions with around one week delay between each session. It can be useful for

research involving automatic visual or audio-visual speech recognition or speaker verification.

### 5.1.2 Polycost

Establishing the *Polycost* corpus was an activity of the so called COST 250 European project. It includes both native and non-native English from 134 speakers (74 M/60 F) from 13 European countries. Therefore it can not only be used in speaker recognition, but language and accent recognition as well. It has more than 5 sessions recorded over weeks in home/office environment by variable telephone handsets through digital ISDN.

### 5.1.3 YOHO

The *YOHO* corpus was designed for evaluating speaker verification in text-dependent situation for secure access applications. It consists of 138 speakers' speech messages (106 M/32 F). It was recorded in multi sessions over a three months period by fixed high-quality handset in the office environment. The text read was prompted digit phrases.

## 5.2 Data Description

Speech databases are most commonly classified into single-session and multi-session. Multi-session databases allow estimation of temporal intra-speaker variability. Combination sets are also possible including single-session recording with a larger set of speakers and multi-session recordings with a smaller set of speakers. [2]

With respect to input devices the most common means of recording are microphones or telephone handsets, the latter can be modified by being over local or long distance telephone lines, GSM or over multiple microphones. According to the acoustic environment, databases are recorded either in noise free environment, such as in the sound booth, or with office/home noise.

Our speech database is single session.

### 5.2.1 Subjects

Database consists of 64 speakers in total. It is divided into subsets namely PIEAS Speech Database1 and PIEAS Speech Database2. PSDB1 contains 38 speakers, 29 males and 9 females whereas PSDB2 contains 26 speakers, 16 males and 10 females.

The database overall has an age distribution from 14 to 36 years.

### 5.2.2 Samples

Each Speaker has 30 recorded samples. Out of which, 10 are recorded from Microphone (Mono), 10 from one handset and 10 from other handset. The reason to choose 10 samples over a single mode of recording is to assess intra-speaker variability.

Different handsets have been chosen in order to analyze the performance measures when different handsets are used.

Overall, the database consists of approximately 2000 samples.

### 5.2.3 Telephonic Data Recording

Samples recorded over telephone calls consist of two modes. In PSDB1, 10 samples for each handset are recorded in a single call whereas in PSDB2 each sample of a particular speaker is recorded in a separate call in order to get more insight into analysis phase of channel distortion.

### 5.2.4 Text

The recorded phrases include short sentences and digit strings of various lengths. The text spoken during enrollment and testing by each speaker is as follows:

- Sequence of digits ( i.e. 1-6, 98,27,33,11)
- Assalamulaikum
- Allah Hafiz
- Joe took father's green shoe bench out
- He eats several light tacos

### 5.2.5 Sampling Rate

Samples have been recorded at 16 KHz with bit rate of 16bps.

### 5.2.6 Equipment Used in Recording

The equipment used for recording speech samples are

- Microphone
- Handset 1
  - Nokia 1100
- Handset 2
  - Nokia N70

**Table 5.1 Summary of Database**

| | **PSDB1** | | **PSDB2** | |
|---|---|---|---|---|
| | **Non Telephonic** | **Telephonic** | **Non Telephonic** | **Telephonic** |
| **Speakers** | 38 (29m, 9f) | 38 (29m, 9f) | 26(16m, 10f) | 26 (16m, 10f) |
| **Samples** | 380 | 380 | 260 | 260 |
| **Sampling Rate** | 16 KHz, 16 bits | 16 KHz, 16 bits | 16 KHz, 16 bits | 16 KHz, 16 bits |
| **Recording time** | 400 min @ | 570 min @ | 300 min @ | 540 min @ |
| **Recording Software** | Sony Sound Forge 9.0 | Call recorder without beep for S60 mobiles v 1.0 | Sony Sound Forge 9.0 | Call recorder without beep for S60 mobiles v 1.0 |
| **Recording Equipment** | A4Tech Headset Mic 1.Nominal Impedance: 32 ohm 2.Sensitivity 97dB | Nokia 1100 Handset<br><br>Nokia N70 Handset | A4Tech Headset Mic 1.Nominal Impedance: 32 ohm 2.Sensitivity 97dB | Nokia 1100 Handset<br><br>Nokia N70 Handset |
| **Size of Digitized Data** | 251 Mb | 451Mb | 163 Mb | 330 Mb |
| **Data format** | .wav | .amr (converted later to .wav) | .wav | .amr (converted later to .wav) |
| **Samples per Call** | - | 10 | - | 1 |
| **Environment** | Office Noise | Channel +Office Noise | Office Noise | Channel + Office Noise |

### 5.2.7 Sample Plots

## Non-Telephonic

Speaker uttering first four digits in text. The sample is recorded through Microphone.

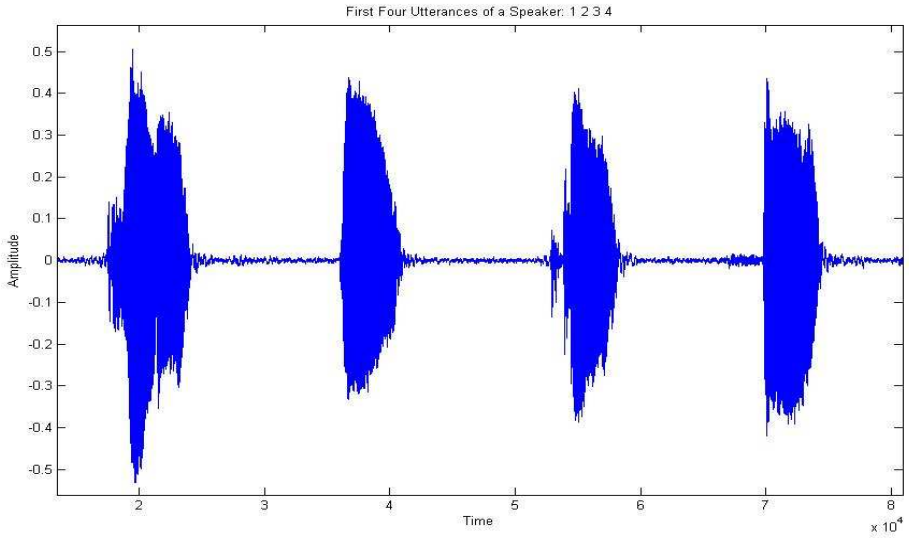



**Figure 5.1 Non telephonic speech sample**

## Telephonic-Handset 1

Speaker uttering first four digits in text. The sample is recorded through Handset 1.

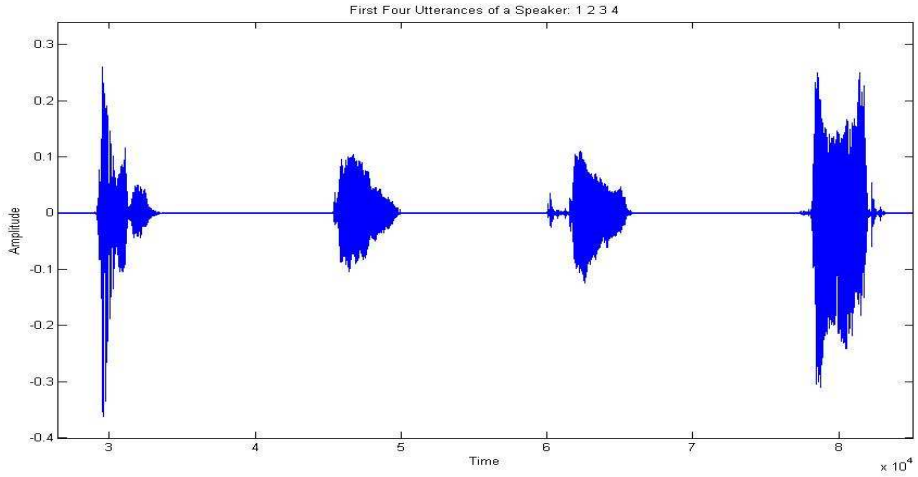



**Figure 5.2 Telephonic speech sample, handset1**

## Telephonic-Handset 2

Speaker uttering first four digits in text. The sample is recorded through Handset 2.

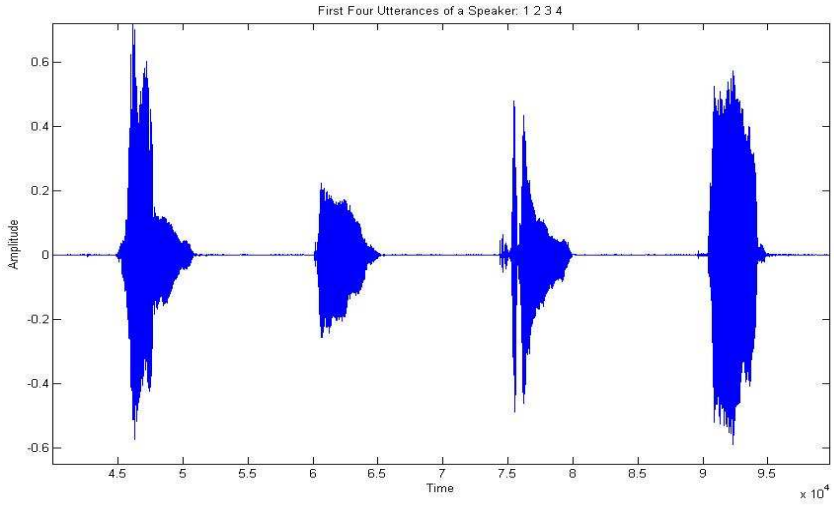



**Figure 5.3 Telephonic speech sample, handset2**

## 5.3 Empirical Results

This section includes all the empirical results based on provided telephonic and non telephonic data. A new approach based on MFCC as feature extraction technique was being used which resulted in improved accuracy as compared to LPC. Analysis on selection of proper number of LPCs and MFCCs has been carried out. Moreover, after selecting appropriate number of features, Results have been taken on Telephonic and Non Telephonic data for matched and mismatched conditions. Effect of using different wavelet has also been observed

The recognition accuracy in case of sample length, gender, Analysis on the basis of single and multilevel decompositions, wavelet type, choosing appropriate technique among MFCC and LPC and analysis on choosing appropriate number of LPCs and MFCCs has also been carried out.

### 5.3.1 Results based on Sample Length

In this case, results on non-telephonic databases were collected according to following training and testing parameters. Moreover, the efficiency of the approach was being observed by reducing the sample length.

**Training:**

| Training Parameters | Values |
|---|---|
| Technique used | DWT-LPC |
| No. of features | 12-order LPCs |
| DWT-Type | Daubechies-3 |
| Speech type | Non-Telephonic |
| Speaker Modeling | K-means, codebook size= 32 |

**Testing:**

| Testing Parameters | Values |
|---|---|
| Classifier | Minimum Distortion |

**Results and Comments:**

Results using above technique on full length samples is given in Table 5.2.

| Database | Type | Technique | No. of Samples | Accuracy % | Error % |
|---|---|---|---|---|---|
| **DB1** | Non Telephonic | LPC-12 | 380 | 91% | 9 % |
| **DB2** | Non Telephonic | LPC-12 | 270 | 92% | 8 % |

**Table 5.3 Recognition Rate After Reducing Length of Sample**

| Database | Type | Length of Sample | Accuracy% | Error% |
|---|---|---|---|---|
| **DB1** | Non Telephonic | Reduced by half | 85.01% | 14.9% |
| **DB2** | Non Telephonic | Reduced by half | 85.7% | 14.3% |

Results obtained by reducing the sample lengths by half are shown in Table 5.3 which show that the performance has been reduced by 5%.

### 5.3.2 Results based on Gender

Training and Testing parameters are given as follows:

**Training:**

| Training Parameters | Values |
|---|---|
| Technique used | DWT-LPC |
| No. of features | 12-order LPCs |
| DWT-Type | Daubechies-3 |
| Speech type | Non-Telephonic |
| Speaker Modeling | K-means, codebook size= 32 |

**Testing:**

| Testing Parameters | Values |
|---|---|
| Classifier | Minimum Distortion |

## Results and comments:

Accuracy of speaker recognition system over both databases was computed on gender bases which are given in Table 5.4 and Table 5.5. The performance of LPC over female voice is better than on male.

**Table 5.4 Recognition Accuracy for Male Speakers**

| Database | Type | Technique | No. of Male Samples | Accuracy % | Error % |
|---|---|---|---|---|---|
| **DB1** | Non Tel. | LPC-12,VQ | 280 | 88% | 12% |
| **DB2** | Non Tel. | LPC-12,VQ | 160 | 88% | 12% |

**Table 5.5 Recognition Accuracy for Female Speakers**

| Database | Type | Technique | Female | Accuracy % | Error % |
|---|---|---|---|---|---|
| **DB1** | Non Telephonic | LPC-12,VQ | 80 | 100% | 0% |
| **DB2** | Non Telephonic | LPC-12, VQ | 110 | 99% | 1% |

### 5.3.3 Selection of No. of LPC

In previous results, 12 order LPCs have been used according to reference to previous researches. Now, the effect of changing number of LPCs has been observed. Training and Testing parameters are given as follows:

**Training:**

| Training Parameters | Values |
|---|---|
| Technique used | DWT-LPC |
| No. of features | 8 to 24 order LPCs |
| DWT-Type | Daubechies-3 |
| Speech type | Non-Telephonic and telephonic |
| Speaker Modeling | K-means, codebook size= 32 |

**Testing:**

| Testing Parameters | Values |
|---|---|
| Classifier | Minimum Distortion |

## Results and Comments:

***Non-Telephonic:***

The best accuracy for DB1 has been achieved at 10. However, on DB2 best accuracy is achieved on $10^{th}$ and $12^{th}$ order of LPC. Thus, according to previous research and the analysis above, LPC of order 12 is more efficient and hence chosen as appropriate number.

Figure 5.4 shows comparison of accuracy on both databases. It is clear that LPCs have performed well in case of DB2 rather DB1. Overall, the approximate accuracy using 12-order LPCs is 90-92%

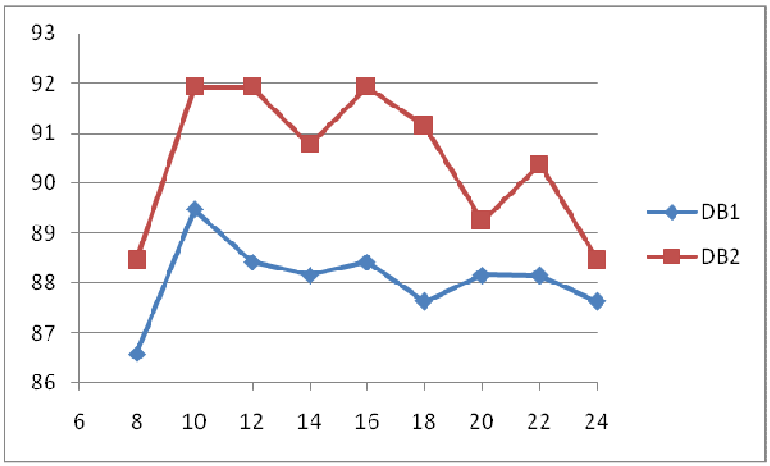



**Figure 5.4 Comparison of DB1 and DB2(Non-Tel) for No. of LPCs**

***Telephonic***

The significance of using 12- order LPC can clearly be seen in Figure 5.4 which shows results on telephonic data. The best accuracy is achieved on 10 and 12 order LPCs in both handsets, i.e. H1 and H2. Though, non telephonic results also show that 10-order LPCs are not consistent in efficiency, we have chosen 12-order LPCs to be an appropriate number for telephonic database as well.

On DB2-H1 as shown in Figure 5.5, best accuracy is approximately 76% whereas on Handset 2, it is about 81%. This also shows that the performance of LPC in Telephonic data is decreased from 10-15%. This is possibly due to sensitivity of LPCs in noisy environments.

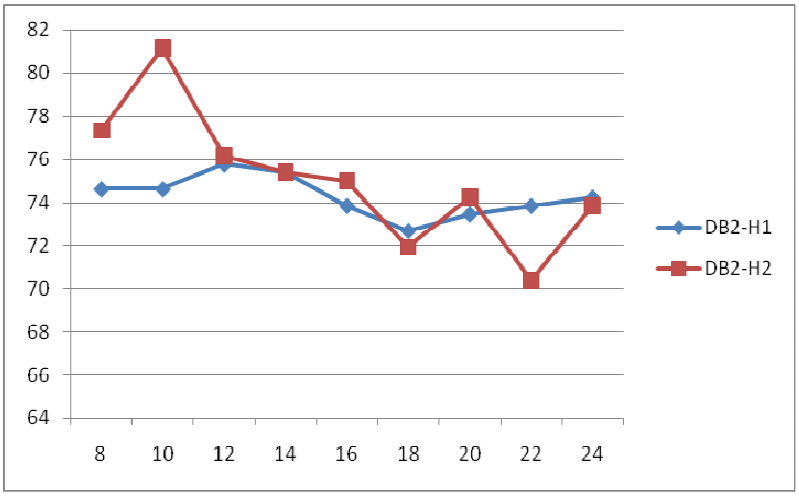



**Figure 5.5 Comparison of DB2-TEL for No. of LPCs**

### 5.3.4 Selection of Number of MFCC

As the analysis for LPCs was being carried out, in this part, analysis on number of MFCCs was being carried out for both telephonic and non telephonic data. Training and Testing parameters are given as follows:

**Training:**

| Training Parameters | Values |
|---|---|
| Technique used | DWT-MFCC |
| No. of features | 10 to 40 order MFCCs |
| Filter Bank for MFCCs | 40 |
| DWT-Type | Daubechies-3 |
| Speech type | Non-Telephonic and telephonic |
| Speaker Modeling | K-means, codebook size= 32 |

**Testing:**

| Testing Parameters | Values |
|---|---|
| Classifier | Minimum Distortion |

## Results and Comments:

***Non-Telephonic***

While analyzing on non telephonic data, best results are achieved when 22 and 24 number of MFCCs have been used. Figure 5.6 shows that there is a rapid improvement in results till 16 MFCCs, i.e. 16 MFCCs must be used and the accuracy increases till 24, but starts decreasing after that. So in our case, 24 is selected as appropriate number of MFCC for non telephonic data.

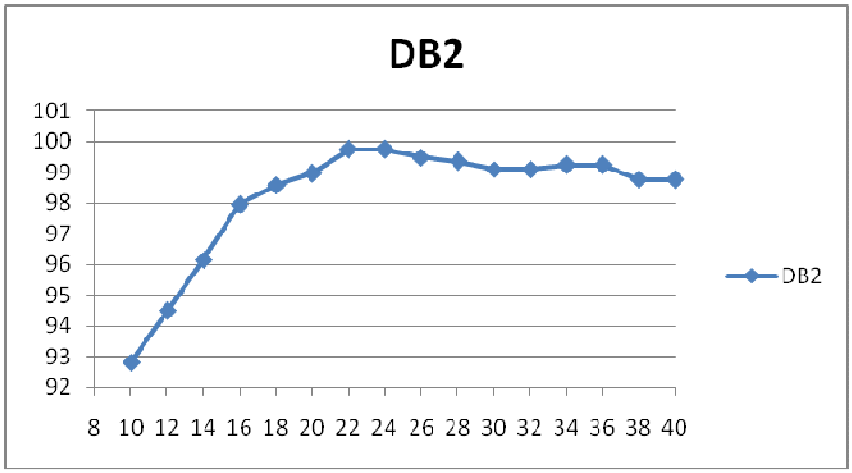



**Figure 5.6 Effect of No. of MFCCs on DB2- Non Telephonic**

***Telephonic***

For telephonic data, it has been observed that the number of MFCCs varies greatly in performance from that of non telephonic. Figure 5.7 shows that approximately 93% accuracy is being achieved in case of Handset 1 by using 12 and 18 number of MFCCs whereas by using 36 numbers of MFCCs, we obtained little better results. Similarly, above analysis when performed for Handset 2, 18 MFCCs didn't worked best but 34.

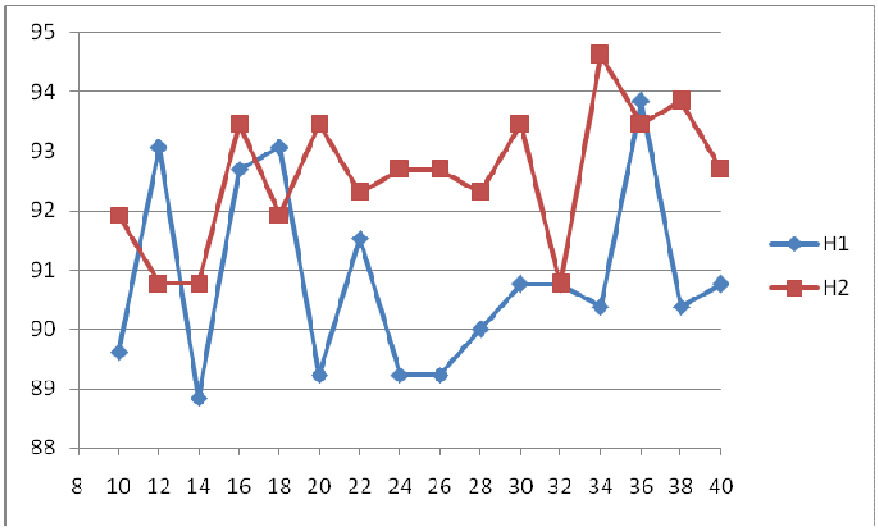



**Figure 5.7 Effect of No. of MFCCs on DB2 Handset-1 and Handset-2**

Figure 5.7 also shows that number of MFCCs fluctuate a lot at start. As after 30 MFCCs, these fluctuations are not that abrupt, 36 MFCCs have been chosen in case of telephonic data for the above dataset.

### 5.3.5 Comparison of Techniques

After the appropriate number for LPCs and MFCCs have been chosen, this section encapsulates overall accuracy on both the databases which also includes the effect of DWT in case of MFCCs. The comparison is being made between simple MFCCs, DWT-MFCCs and DWT-LPCs.

**Training:**

| Training Parameters | Values |
|---|---|
| Technique used | DWT-LPC<br>MFCC<br>DWT-MFCC |
| No. of features | 24 order MFCCs for Non-Tel<br>36 order MFCCs for Telephonic<br>12-order LPC |
| Filter Bank for MFCCs | 40 |
| DWT-Type | Daubechies-3 |
| Speech type | Non-Telephonic and telephonic |
| Speaker Modeling | K-means, codebook size= 32 |

**Testing:**

| Testing Parameters | Values |
|---|---|
| Classifier | Minimum Distortion |

#### *Non Telephonic*

MFCCs have been proved to work best here. As shown in Figure 5.8, MFCC and DWT-MFCC do not differ in results. This effect of DWT can be observed in results discussed later.

As compared to MFCC, LPCs have been proven less efficient here as the result is decreased from 7-8%.In DB2, again MFCCs have outperformed LPCs. Also, a slight increase in performance can be observed when DWT is being used. Again, the difference between performance levels of LPCs and MFCCs is 7-8%. Overall, Results for DB2 are 2-3% greater in performance than that of DB1.

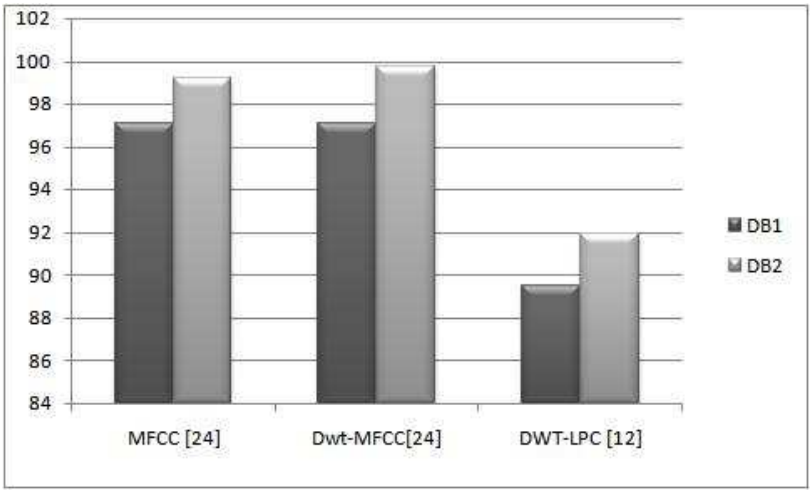



**Figure 5.8 Comparison of techniques on both databases (Non-Telephonic)**

Hence, DWT-MFCCs has been proved best technique until now for non telephonic data.

## Telephonic

This section summarizes which of the three techniques was best in Telephonic case. Considering DB1, Figure 5.9 shows similar trend that was observed in DB1 non telephonic. MFCCs show better results than LPCs.

Handset 2 also showed similar trend as of handset 1, but the overall efficiency of DB1 on handset 2 is greater than that of handset 1. This is due to the handset variability and the microphone distortion in handsets.

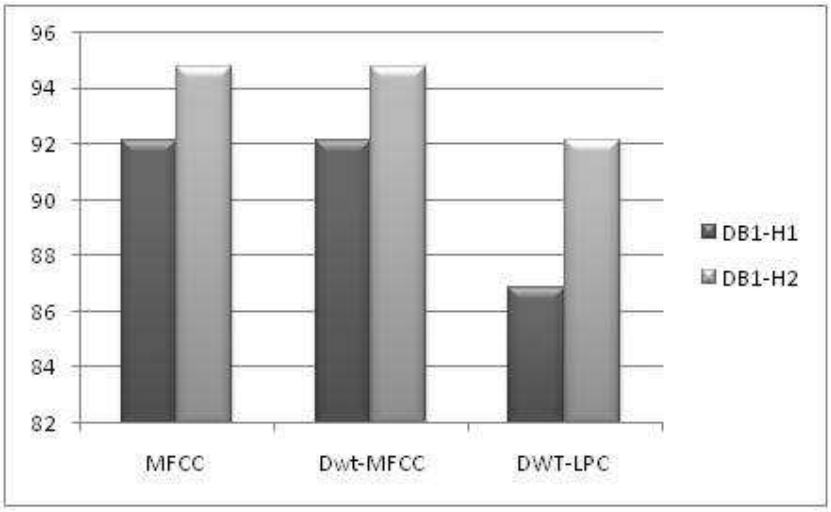



**Figure 5.9 Comparison of techniques on both Handset 1 and Handset 2 of DB1**

DB2 on the other hand shows significance of DWT in case of MFCCs more clearly than that of DB1, but the overall efficiency here is decreased.

Figure 5.10 below shows results on Handset 1 and Handset 2 of DB2, which depicts that DWT- MFCCs have resulted better than simple MFCCs by an increase of approximately 10% accuracy. Overall results as compared to non telephonic are though decreased. Again, MFCCs have outperformed LPCs.

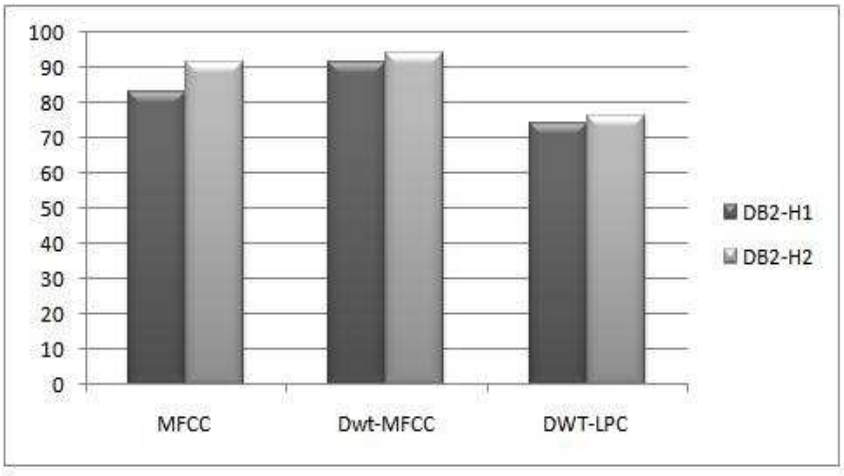



**Figure 5.10 Comparison of techniques on both Handset 1 and Handset 2 of DB2**

## 5.3.6 Effect of Decomposition Levels

In this section, the effect of decomposition levels on technique that performed best (i.e. DWT-MFCC) was observed. Training and Testing parameters are given as follows:

**Training:**

| Training Parameters | Values |
|---|---|
| Technique used | DWT-MFCC |
| No. of features | 24 order MFCCs |
| Filter Bank for MFCCs | 40 |
| DWT-Type | Daubechies-3 |
| Speech type | Non-Telephonic |
| Speaker Modeling | K-means, codebook size= 32 |

**Testing:**

| Testing Parameters | Values |
|---|---|
| Classifier | Minimum Distortion |

.

**Results and Comments**

While analyzing the effect of DWT levels, it has been observed that MFCC decreases in efficiency as the number of levels are increased.

Figure 5.11 shows the best accuracy was achieved at level 1.

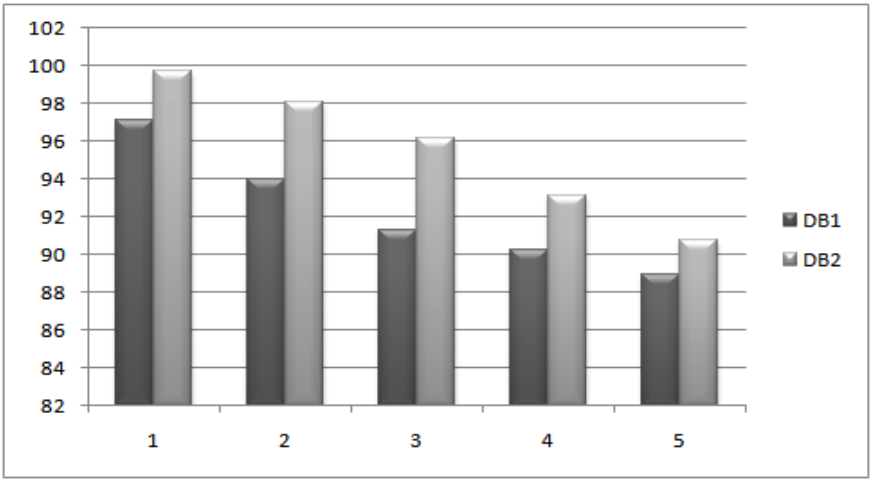



**Figure 5.11 Effect of decomposition levels using MFCCs on DB1 and DB2 (Non Telephonic)**

## 5.3.7 Effect of Wavelet Type

In this project, previously Daubechies-3 is being used as reference to [6]. Now the analysis on wavelet type was carried out.

Training and Testing parameters are given as follows:

**Training:**

| Training Parameters | Values |
|---|---|
| Technique used | DWT-MFCC |
| No. of features | 24 order MFCCs for Non-Tel<br>36 order MFCC for Tel |
| Filter Bank for MFCCs | 40 |
| DWT-Type | Daubechies-1,2,3<br>Haar<br>Symlets<br>Discrete Meyer |
| Speech type | Non-Telephonic |
| Speaker Modeling | K-means, codebook size= 32 |

**Testing:**

| Testing Parameters | Values |
|---|---|
| Classifier | Minimum Distortion |

The effect of other wavelet types has been observed on Non Telephonic Database and results are shown in Figure 5.12.

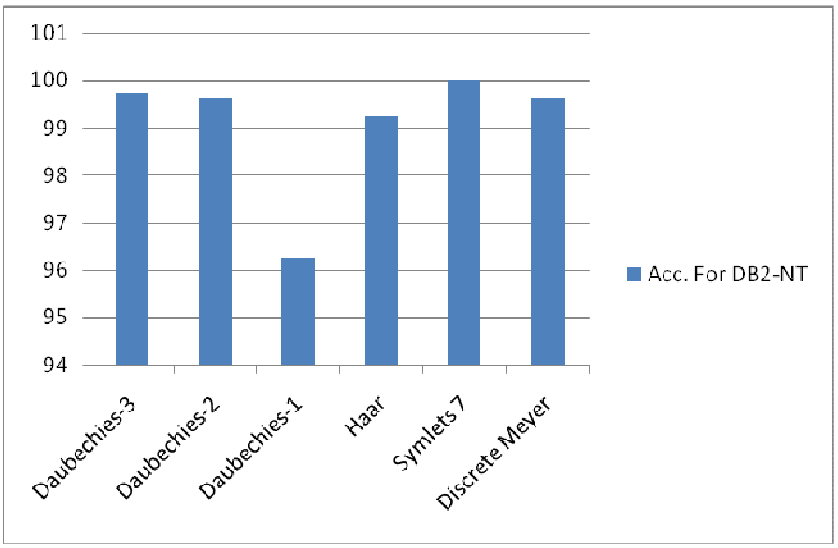



**Figure 5.12 Effect of wavelet type on DB2-Non Telephonic**

On DB2, Symlets and Daubechies-3 gave best accuracy which is shown in Figure 5.12. Whereas on DB1, Discrete Meyer has performed best.

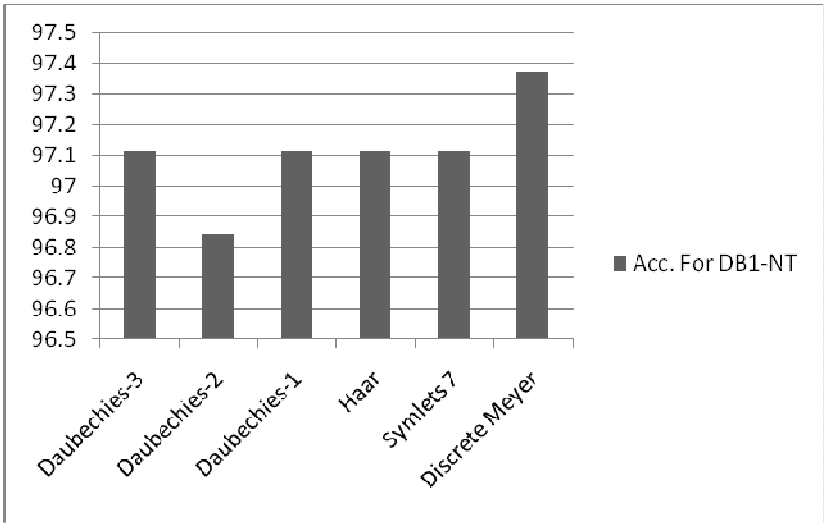



**Figure 5.13 Effect of wavelet type on DB1 Non-Telephonic**

Overall, from Figure 5.12 and 5.13 it can be seen that there is not much difference in results by changing wavelet. Moreover, Daubechies-3's performance is stable and better than Daubechies-1 and Daubechies 2.

On the whole, Symlets' performance is better than all others.

### 5.3.8 Mismatched Conditions

Mismatched conditions are those in which testing and training data is recorded in different environments. These are most likely conditions when it comes to Speaker Recognition.

**Training:**

| Training Parameters | Values |
|---|---|
| Technique used | DWT-MFCC |
| No. of features | 36 order MFCCs |
| Filter Bank for MFCCs | 40 |
| DWT-Type | Daubechies-3 |
| Speech type | Non-Telephonic , Telephonic |
| Speaker Modeling | K-means, codebook size= 32 |

**Testing:**

| Testing Parameters | Values |
|---|---|
| Classifier | Minimum Distortion |

**Results and Comments**

Figure 5.14 below, shows the results for mismatched conditions in DB1. When the system is trained on Non Telephonic speech and tested on Telephonic, Accuracy is reduced about 40-50%.

Whereas when the system is being trained on telephonic speech from another handset and tested on other, then MFCCs have resulted in improved accuracy.

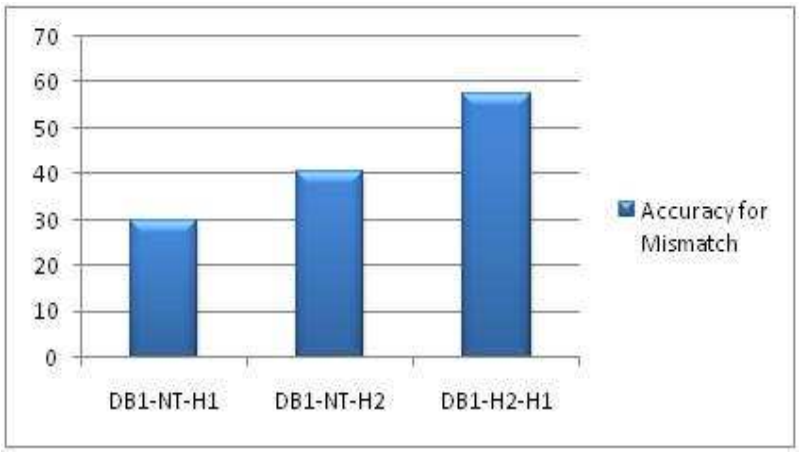



**Figure 5.14 Results for mismatched conditions in DB1**

Figure 5.15 shows results for mismatched conditions for DB2. This also shows that Handset 2 has some advantage over Handset 1. Also, if a system is trained on a telephonic speech, testing can give better results.

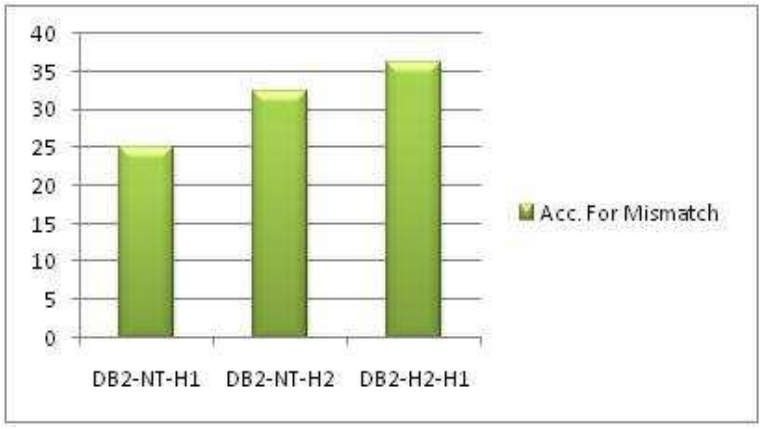



**Figure 5.15 Results for mismatched conditions in DB2**

Error in this case is possibly due to microphone distortion. As mentioned in chapter 3, carbon button in microphones is mainly a reason for not getting satisfactory results.

### 5.3.9 Increasing Population Size for the Best Technique

In this case, the population size was increased and results were again computed using best technique i.e. DWT-MFCC.

#### *a. Comparison of MFCCs with LPCs*

Results for proposed technique were compared with that in which LPCs were used. Experiments show that MFCCs along with wavelet transform perform better than LPCs with wavelet transform, especially in telephonic data as shown in Figure .5.16.

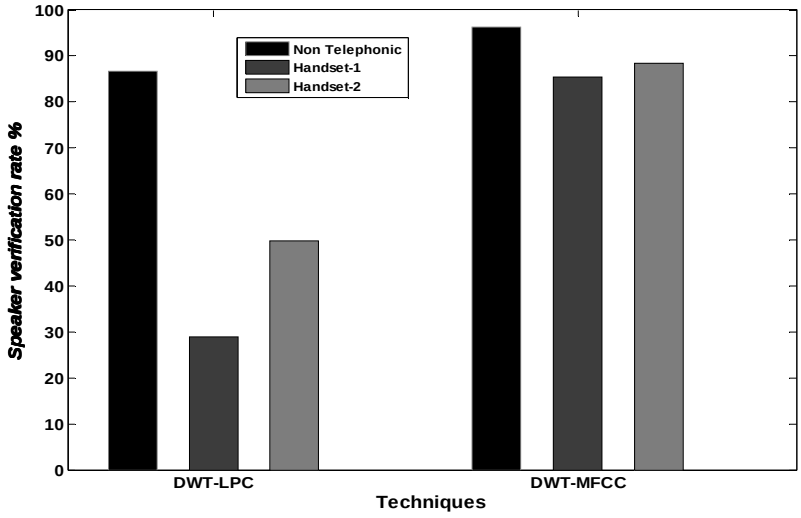



**Figure 5.16 Comparison of recognition rate using DWT with LPC and MFCC**

### *b. Effect of Number of MFCCs*

As previous results show that proposed technique performed better, we then analyzed effect of number of MFCCs for both telephonic and non telephonic data.
Fig. 5.17 shows that as the number of MFCCs are increased, recognition rate increases rapidly in start and then varies gradually. In case of non-Telephonic speech, 18 MFCCs whereas in case of Telephonic 20 MFCCs must be used. Overall, the best number of MFCCs is suggested to be 38 which gives 96% recognition accuracy on non telephonic data and 86% recognition accuracy on telephonic data.

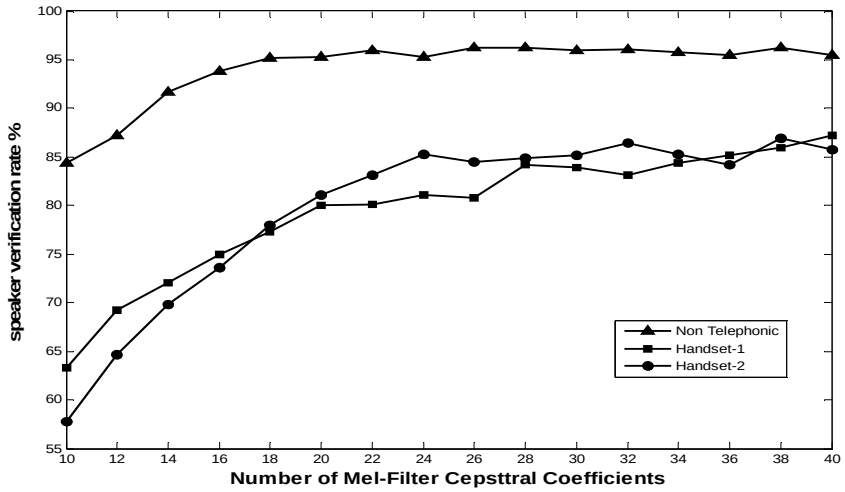



**Figure 5.17 Effect of increasing number of Mel-Coefficients**

### *c. Effect of Decomposition Levels*

As number of decomposition levels is increased, further information from a signal can be extracted in some cases, but increasing number of levels not only increases computational complexity but also introduces redundant data which do not contain any further information. Thus, choosing an appropriate number of decomposition levels has become a significant problem.

Fig. 5.18 shows results for increasing number of decomposition levels which were computed up to three levels and experiments show that highest performance is achieved on level 1. This is possibly due to the reason that speech signal loses its characteristics as the decomposition levels are increased.

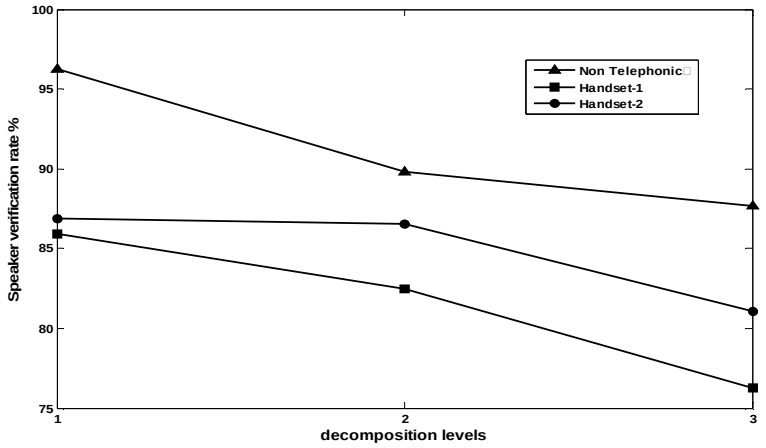



**Figure 5.18 Effect of Decomposition Levels**

### *d. Effect of Wavelet Type*

Wavelet type has also been a part of analyzing. For this purpose we have used Daubechies wavelet i.e. DB1-DB3, Haar, Symlets and Discrete Meyer.

Results show that Symlets has performed better in both non telephonic and telephonic speech as shown in Table 1, so we have used Symlets throughout this technique.

**Table 5.2 Effect of changing wavelet type**

| Wavelet Type | Non Telephonic | Handset-1 | Handset-2 |
|---|---|---|---|
| **Db1** | 95.47 | 86.41 | 87.03 |
| **Db2** | 95.94 | 85.78 | 85.78 |
| **Db3** | 95.47 | 84.69 | 87.19 |
| **Haar** | 95.94 | 85.47 | 85.78 |
| **Sym7** | 96.25 | 85.94 | 86.88 |
| **Discrete Meyer** | 96.09 | 85.47 | 86.88 |

### *e. Overall proposed parameters and procedure*

These are most likely conditions when it comes to Speaker Recognition.

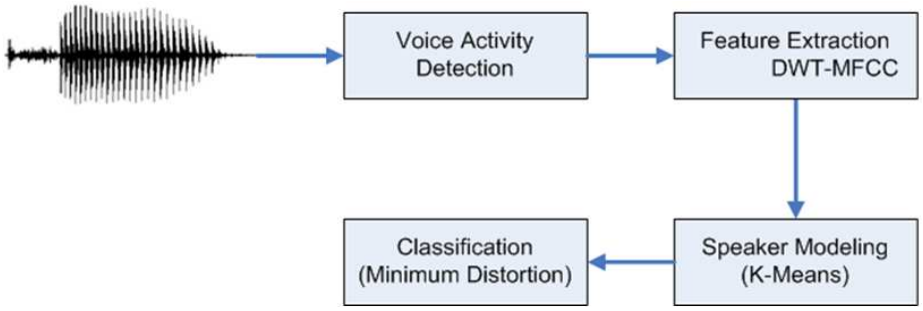
**Training:**

| Training Parameters | Values |
|---|---|
| Technique used | DWT-MFCC |
| No. of features | 38order MFCCs |
| Filter Bank for MFCCs | 40 |
| DWT-Type | Symlets 7 |
| | |
| Speaker Modeling | K-means, codebook size= 32 |

**Testing:**

| Testing Parameters | Values |
|---|---|
| Classifier | Minimum Distortion |

**Procedure:**



## 5.3.10 Appendix-A: Results

**5.3 Effect of Increasing No. of LPCs on DB1-Non Telephonic**

| Database | Technique | Number of LPC | Samples | Accuracy % | File Name |
|---|---|---|---|---|---|
| **DB1-NT** | Dwt-LPC | 8 | 380 | 86.58 | DB1NT_exAll_8lpc.m |
| **DB1-NT** | Dwt-LPC | 10 | 380 | 89.47 | DB1NT_exAll_10lpc.m |
| **DB1-NT** | Dwt-LPC | 12 | 380 | 88.42 | DB1NT_exAll_12lpc.m |
| **DB1-NT** | Dwt-LPC | 14 | 380 | 88.16 | DB1NT_exAll_14lpc.m |
| **DB1-NT** | Dwt-LPC | 16 | 380 | 88.42 | DB1NT_exAll_16lpc.m |
| **DB1-NT** | Dwt-LPC | 18 | 380 | 87.63 | DB1NT_exAll_18lpc.m |
| **DB1-NT** | Dwt-LPC | 20 | 380 | 88.16 | DB1NT_exAll_20lpc.m |
| **DB1-NT** | Dwt-LPC | 22 | 380 | 88.16 | DB1NT_exAll_22lpc.m |
| **DB1-NT** | Dwt-LPC | 24 | 380 | 87.63 | DB1NT_exAll_24lpc.m |

**Table 5.4 Effect of Increasing No. of LPCs on DB2-Non Telephonic**

| Database | Technique | Number of LPC | Samples | Accuracy % | File Name |
|---|---|---|---|---|---|
| **DB2-NT** | Dwt-LPC | 8 | 260 | 88.46 | DB2NT_exAll_8lpc.m |
| **DB2-NT** | Dwt-LPC | 10 | 260 | 91.92 | DB2NT_exAll_10lpc.m |
| **DB2-NT** | Dwt-LPC | 12 | 260 | 91.92 | DB2NT_exAll_12lpc.m |
| **DB2-NT** | Dwt-LPC | 14 | 260 | 90.77 | DB2NT_exAll_14lpc.m |
| **DB2-NT** | Dwt-LPC | 16 | 260 | 91.92 | DB2NT_exAll_16lpc.m |
| **DB2-NT** | Dwt-LPC | 18 | 260 | 91.15 | DB2NT_exAll_18lpc.m |
| **DB2-NT** | Dwt-LPC | 20 | 260 | 89.23 | DB2NT_exAll_20lpc.m |
| **DB2-NT** | Dwt-LPC | 22 | 260 | 90.38 | DB2NT_exAll_22lpc.m |
| **DB2-NT** | Dwt-LPC | 24 | 260 | 88.46 | DB2NT_exAll_24lpc.m |

**Table 5.5 Effect of Increasing No. of LPCs on DB2- Telephonic**

| Database | Technique | No. of LPC | Samples | Accuracy | File Name |
|---|---|---|---|---|---|
| **DB2-H2** | DWT-LPC | 8 | 260 | 77.31 | DB2H2_exAll_8lpc.m |
| **DB2-H2** | DWT-LPC | 10 | 260 | 81.15 | DB2H2_exAll_10lpc.m |
| **DB2-H2** | DWT-LPC | 12 | 260 | 76.15 | DB2H2_exAll_12lpc.m |
| **DB2-H2** | DWT-LPC | 14 | 260 | 75.38 | DB2H2_exAll_14lpc.m |
| **DB2-H2** | DWT-LPC | 16 | 260 | 75.0 | DB2H2_exAll_16lpc.m |
| **DB2-H2** | DWT-LPC | 18 | 260 | 71.92 | DB2H2_exAll_18lpc.m |
| **DB2-H2** | DWT-LPC | 20 | 260 | 74.23 | DB2H2_exAll_20lpc.m |
| **DB2-H2** | DWT-LPC | 22 | 260 | 70.38 | DB2H2_exAll_22lpc.m |
| **DB2-H2** | DWT-LPC | 24 | 260 | 73.85 | DB2H2_exAll_24lpc.m |

**Table 5.6 Effect of Increasing No. of LPCs on DB2-Telephonic Handset-1**

| Database | Technique | No. of LPC | Samples | Accuracy | File Name |
|---|---|---|---|---|---|
| **DB2-H1** | DWT-LPC | 8 | 260 | 74.62 | DB2H1_exAll_8lpc.m |
| **DB2-H1** | DWT-LPC | 10 | 260 | 74.62 | DB2H1_exAll_10lpc.m |
| **DB2-H1** | DWT-LPC | 12 | 260 | 75.77 | DB2H1_exAll_12lpc.m |
| **DB2-H1** | DWT-LPC | 14 | 260 | 75.38 | DB2H1_exAll_14lpc.m |
| **DB2-H1** | DWT-LPC | 16 | 260 | 73.85 | DB2H1_exAll_16lpc.m |
| **DB2-H1** | DWT-LPC | 18 | 260 | 72.69 | DB2H1_exAll_18lpc.m |
| **DB2-H1** | DWT-LPC | 20 | 260 | 73.46 | DB2H1_exAll_20lpc.m |
| **DB2-H1** | DWT-LPC | 22 | 260 | 73.85 | DB2H1_exAll_22lpc.m |
| **DB2-H1** | DWT-LPC | 24 | 260 | 74.23 | DB2H1_exAll_24lpc.m |

**Table 5.7 Effect of Increasing No. of MFCCs on DB2-Non Telephonic**

| Database | Technique | No. of MFCC | Samples | Accuracy | File Name |
|---|---|---|---|---|---|
| **DB2-NT** | DWT-MFCC | 10 | 260 | 92.82 | DB2NT_exAll_10mfcc.m |
| **DB2-NT** | DWT-MFCC | 12 | 260 | 94.49 | DB2NT_exAll_12mfcc.m |
| **DB2-NT** | DWT-MFCC | 14 | 260 | 96.15 | DB2NT_exAll_14mfcc.m |
| **DB2-NT** | DWT-MFCC | 16 | 260 | 97.95 | DB2NT_exAll_16mfcc.m |
| **DB2-NT** | DWT-MFCC | 18 | 260 | 98.59 | DB2NT_exAll_18mfcc.m |
| **DB2-NT** | DWT-MFCC | 20 | 260 | 98.97 | DB2NT_exAll_20mfcc.m |
| **DB2-NT** | DWT-MFCC | 22 | 260 | 99.74 | DB2NT_exAll_22mfcc.m |
| **DB2-NT** | DWT-MFCC | 24 | 260 | 99.74 | DB2NT_exAll_24mfcc.m |
| **DB2-NT** | DWT-MFCC | 26 | 260 | 99.49 | DB2NT_exAll_26mfcc.m |
| **DB2-NT** | DWT-MFCC | 28 | 260 | 99.36 | DB2NT_exAll_28mfcc.m |
| **DB2-NT** | DWT-MFCC | 30 | 260 | 99.10 | DB2NT_exAll_30mfcc.m |
| **DB2-NT** | DWT-MFCC | 32 | 260 | 99.10 | DB2NT_exAll_32mfcc.m |
| **DB2-NT** | DWT-MFCC | 34 | 260 | 99.23 | DB2NT_exAll_34mfcc.m |
| **DB2-NT** | DWT-MFCC | 36 | 260 | 99.23 | DB2NT_exAll_36mfcc.m |
| **DB2-NT** | DWT-MFCC | 38 | 260 | 98.77 | DB2NT_exAll_38mfcc.m |
| **DB2-NT** | DWT-MFCC | 40 | 260 | 98.77 | DB2NT_exAll_40mfcc.m |

**Table 5.8 Effect of Increasing No. of MFCCs on DB2-Telephonic Handset-2**

| Database | Technique | No. of MFCC | Samples | Accuracy | File Name |
|---|---|---|---|---|---|
| **DB2-H1** | DWT-MFCC | 10 | 260 | 89.61 | DB2H1_exAll_10mfcc.m |
| **DB2-H1** | DWT-MFCC | 12 | 260 | 93.07 | DB2H1_exAll_12mfcc.m |
| **DB2-H1** | DWT-MFCC | 14 | 260 | 8.84 | DB2H1_exAll_14mfcc.m |
| **DB2-H1** | DWT-MFCC | 16 | 260 | 92.69 | DB2H1_exAll_16mfcc.m |
| **DB2-H1** | DWT-MFCC | 18 | 260 | 93.07 | DB2H1_exAll_18mfcc.m |
| **DB2-H1** | DWT-MFCC | 20 | 260 | 89.23 | DB2H1_exAll_20mfcc.m |
| **DB2-H1** | DWT-MFCC | 22 | 260 | 91.53 | DB2H1_exAll_22mfcc.m |
| **DB2-H1** | DWT-MFCC | 24 | 260 | 89.23 | DB2H1_exAll_24mfcc.m |
| **DB2-H1** | DWT-MFCC | 26 | 260 | 89.23 | DB2H1_exAll_26mfcc.m |
| **DB2-H1** | DWT-MFCC | 28 | 260 | 90.0 | DB2H1_exAll_28mfcc.m |
| **DB2-H1** | DWT-MFCC | 30 | 260 | 90.76 | DB2H1_exAll_30mfcc.m |
| **DB2-H1** | DWT-MFCC | 32 | 260 | 90.76 | DB2H1_exAll_32mfcc.m |
| **DB2-H1** | DWT-MFCC | 34 | 260 | 90.38 | DB2H1_exAll_34mfcc.m |
| **DB2-H1** | DWT-MFCC | 36 | 260 | 93.84 | DB2H1_exAll_36mfcc.m |
| **DB2-H1** | DWT-MFCC | 38 | 260 | 90.38 | DB2H1_exAll_38mfcc.m |
| **DB2-H1** | DWT-MFCC | 40 | 260 | 90.76 | DB2H1_exAll_40mfcc.m |

**Table 5.9 Effect of Increasing No. of MFCCs on DB2-Telephonic Handset-2**

| Database | Technique | No. of MFCC | Samples | Accuracy | File Name |
|---|---|---|---|---|---|
| **DB2-H2** | DWT-MFCC | 10 | 260 | 91.92 | DB2H2_exAll_10mfcc.m |
| **DB2-H2** | DWT-MFCC | 12 | 260 | 90.77 | DB2H2_exAll_12mfcc.m |
| **DB2-H2** | DWT-MFCC | 14 | 260 | 90.77 | DB2H2_exAll_14mfcc.m |
| **DB2-H2** | DWT-MFCC | 16 | 260 | 93.46 | DB2H2_exAll_16mfcc.m |
| **DB2-H2** | DWT-MFCC | 18 | 260 | 91.92 | DB2H2_exAll_18mfcc.m |
| **DB2-H2** | DWT-MFCC | 20 | 260 | 93.46 | DB2H2_exAll_20mfcc.m |
| **DB2-H2** | DWT-MFCC | 22 | 260 | 92.30 | DB2H2_exAll_22mfcc.m |
| **DB2-H2** | DWT-MFCC | 24 | 260 | 92.69 | DB2H2_exAll_24mfcc.m |
| **DB2-H2** | DWT-MFCC | 26 | 260 | 92.69 | DB2H2_exAll_26mfcc.m |
| **DB2-H2** | DWT-MFCC | 28 | 260 | 92.30 | DB2H2_exAll_28mfcc.m |
| **DB2-H2** | DWT-MFCC | 30 | 260 | 93.46 | DB2H2_exAll_30mfcc.m |
| **DB2-H2** | DWT-MFCC | 32 | 260 | 90.76 | DB2H2_exAll_32mfcc.m |
| **DB2-H2** | DWT-MFCC | 34 | 260 | 94.62 | DB2H2_exAll_34mfcc.m |
| **DB2-H2** | DWT-MFCC | 36 | 260 | 93.46 | DB2H2_exAll_36mfcc.m |
| **DB2-H2** | DWT-MFCC | 38 | 260 | 93.84 | DB2H2_exAll_38mfcc.m |
| **DB2-H2** | DWT-MFCC | 40 | 260 | 92.69 | DB2H2_exAll_40mfcc.m |

**Table 5.10 Comparison of Techniques on DB1-Non Telephonic**

| Database | Technique | Number of Features | Samples | Accuracy | File Name |
|---|---|---|---|---|---|
| **DB1** | MFCC | 24 | 380 | 97.11 | DB1NT_exAll_mfcc.m |
| **DB1** | Dwt-MFCC | 24 | 380 | 97.11 | DB1NT_exAll_dmfcc.m |
| **DB1** | DWT-LPCC | 12 | 380 | 89.47 | DB1NT_exAll_LPC.m |

**Table 5.11 Comparison of Techniques on DB2-Non Telephonic**

| Database | Technique | No. of Features | Samples | Accuracy | File Name |
|---|---|---|---|---|---|
| **DB2** | MFCC | 24 | 260 | 99.23 | DB2NT_exAll_mfcc.m |
| **DB2** | Dwt-MFCC | 24 | 260 | 99.74 | DB2NT_exAll_dmfcc.m |
| **DB2** | DWT-LPCC | 12 | 260 | 91.92 | DB2NT_exAll_LPC.m |

**Table 5.12 Comparison of Techniques on DB1-Telephonic Handset-1**

| Database | Technique | No. of Features | Samples | Accuracy | File Name |
|---|---|---|---|---|---|
| **DB1-H1** | MFCC | 36 | 380 | 92.11 | DB1H1_exAll_mfcc.m |
| **DB1-H1** | Dwt-MFCC | 36 | 380 | 92.11 | DB1H1_exAll_dmfcc.m |
| **DB1-H1** | DWT-LPC | 12 | 380 | 86.84 | DB1H1_exAll_LPC.m |

**Table 5.13 Comparison of Techniques on DB1-Telephonic Handset-2**

| Database | Technique | No. of Features | Samples | Accuracy | File Name |
|---|---|---|---|---|---|
| **DB1-H2** | MFCC | 36 | 380 | 94.74 | DB1H2_exAll_mfcc.m |
| **DB1-H2** | Dwt-MFCC | 36 | 380 | 94.74 | DB1H2_exAll_dmfcc.m |
| **DB1-H2** | DWT-LPC | 12 | 380 | 92.11 | DB1H2_exAll_LPC.m |

**Table 5.14 Comparison of Techniques on DB2-Telephonic Handset-1**

| Database | Technique | No. of Features | Samples | Accuracy | File Name |
|---|---|---|---|---|---|
| **DB2-H1** | MFCC | 36 | 260 | 83.08 | DB2H1_exAll_mfcc.m |
| **DB2-H1** | Dwt-MFCC | 36 | 260 | 91.54 | DB2H1_exAll_dmfcc.m |
| **DB2-H1** | DWT-LPCC | 12 | 260 | 73.86 | DB2H1_exAll_LPC.m |

**Table 5.15 Comparison of Techniques on DB2-Telephonic Handset-2**

| Database | Technique | No. of Features | Samples | Accuracy | File Name |
|---|---|---|---|---|---|
| **DB2-H2** | MFCC | 36 | 260 | 91.54 | DB2H2_exAll_mfcc.m |
| **DB2-H2** | Dwt-MFCC | 36 | 260 | 93.85 | DB2H2_exAll_dmfcc.m |
| **DB2-H2** | DWT-LPCC | 12 | 260 | 76.15 | DB2H2_exAll_LPC.m |

**Table 5.16 Comparison of Techniques on DB1-Mismatched Conditions**

| Database-Train-Test | Technique | No. of Features | Samples | Accuracy | File Name |
|---|---|---|---|---|---|
| **DB1-NT-H1** | Dwt-MFCC | 36 | 380 | 29.69 | DB1_exAll_NTH1.m |
| **DB1-NT-H2** | Dwt-MFCC | 36 | 380 | 40.38 | DB1_exAll_NTH2.m |
| **DB1-H2-H1** | Dwt-MFCC | 36 | 380 | 57.37 | DB1_exAll_H2H1.m |

**Table 5.17 Comparison of Techniques on DB2-Mismatched Conditions**

| Database-Train-Test | Technique | No. of Features | Samples | Accuracy | File Name |
|---|---|---|---|---|---|
| **DB2-NT-H1** | Dwt-MFCC | 36 | 260 | 25.00 | DB2_exAll_NTH1.m |
| **DB2-NT-H2** | Dwt-MFCC | 36 | 260 | 32.38 | DB2_exAll_NTH2.m |
| **DB2-H2-H1** | Dwt-MFCC | 36 | 260 | 36.15 | DB2_exAll_H2H1.m |

**Table 5.18 Effect of Decomposition Levels on DB1**

| No of DWT-Levels | Wavelet Type | Technique | DB1-NonTEL. Accuracy | File Name |
|---|---|---|---|---|
| **1** | Daubechies-3 | DWT-MFCC | 97.11 | DB1_exAll_lev1.m |
| **2** | Daubechies-3 | DWT-MFCC | 93.95 | DB1_exAll_lev2.m |
| **3** | Daubechies-3 | DWT-MFCC | 91.32 | DB1_exAll_lev3.m |
| **4** | Daubechies-3 | DWT-MFCC | 90.26 | DB1_exAll_lev4.m |
| **5** | Daubechies-3 | DWT-MFCC | 88.95 | DB1_exAll_lev5.m |

**Table 5.19 Effect of Decomposition Levels on DB2**

| No of DWT-Levels | Wavelet Type | Technique | DB2-NonTEL. Accuracy | File Name |
|---|---|---|---|---|
| **1** | Daubechies-3 | DWT-MFCC | 99.74 | DB2_exAll_lev1.m |
| **2** | Daubechies-3 | DWT-MFCC | 98.08 | DB2_exAll_lev2.m |
| **3** | Daubechies-3 | DWT-MFCC | 96.15 | DB2_exAll_lev3.m |
| **4** | Daubechies-3 | DWT-MFCC | 93.08 | DB2_exAll_lev4.m |
| **5** | Daubechies-3 | DWT-MFCC | 90.77 | DB2_exAll_lev5.m |

**Table 5.20 Effect of Changing Wavelet Type on DB1**

| Database | Technique | Wavelet Type | Accuracy | File Name |
|---|---|---|---|---|
| **DB1-NT** | DWT-MFCC | Daubechies-3 | 97.11 | DB2_exAll_db3.m |
| **DB1-NT** | DWT-MFCC | Daubechies-2 | 96.84 | DB2_exAll_db2.m |
| **DB1-NT** | DWT-MFCC | Daubechies-1 | 97.11 | DB2_exAll_db1.m |
| **DB1-NT** | DWT-MFCC | Haar | 97.11 | DB2_exAll_haar.m |
| **DB1-NT** | DWT-MFCC | Symlets 7 | 97.11 | DB2_exAll_sym.m |
| **DB1-NT** | DWT-MFCC | Discrete Meyer | 97.37 | DB2_exAll_dmey.m |

**Table 5.21 Effect of Changing Wavelet Type on DB2**

| Database | Technique | Wavelet Type | Accuracy | File Name |
|---|---|---|---|---|
| **DB2-NT** | DWT-MFCC | Daubechies-3 | 99.74 | DB1_exAll_db3.m |
| **DB2-NT** | DWT-MFCC | Daubechies-2 | 99.62 | DB1_exAll_db2.m |
| **DB2-NT** | DWT-MFCC | Daubechies-1 | 96.23 | DB1_exAll_db1.m |
| **DB2-NT** | DWT-MFCC | Haar | 99.23 | DB1_exAll_haar.m |
| **DB2-NT** | DWT-MFCC | Symlets 7 | 100.0 | DB1_exAll_sym.m |
| **DB2-NT** | DWT-MFCC | Discrete Meyer | 99.62 | DB1_exAll_dmey.m |

**Table 5.22 Comparison of Techniques on increased Non Telephonic Data**

| Non Telephonic (Samples) | Technique | Accuracy | File Name |
|---|---|---|---|
| **640** | Dwt-LPC | 86.72 | exAll_lpc.m |
| **640** | Dwt-MFCC | 96.25 | exAll_mfcc.m |
| **640** | DWT-MFCC-DER-2 | 95.47 | exAll_mfccd2.m |
| **640** | DWT-MFCC-DER-1 | 95.63 | exAll_mfccd1.m |

**Table 5.23 Comparison of Techniques on increased Telephonic (H1) Data**

| Telephonic –H1 | Technique | Accuracy | File Name |
|---|---|---|---|
| **640** | Dwt-LPC | 28.91 | exAll_lpc_h1.m |
| **640** | Dwt-MFCC | 85.31 | exAll_mfcc_h1.m |
| **640** | DWT-MFCC-DER-2 | 86.88 | exAll_mfccd2_h1.m |
| **640** | DWT-MFCC-DER-1 | 82.50 | exAll_mfccd1_h1.m |

**Table 5.24 Comparison of Techniques on increased Telephonic (H2) Data**

| Telephonic –H2 | Technique | Accuracy | File Name |
|---|---|---|---|
| **640** | Dwt-LPC | 49.88 | exAll_lpc_h2.m |
| **640** | Dwt-MFCC | 88.44 | exAll_mfcc_h2.m |
| **640** | DWT-MFCC-DER-2 | 86.41 | exAll_mfccd2_h2.m |
| **640** | DWT-MFCC-DER-1 | 85.78 | exAll_mfccd1_h2.m |

**Table 5.25 Effect of Number of MFCCs on increased Non Telephonic Data**

| Database | Technique | No. of MFCCs | Samples | Accuracy | Filename |
|---|---|---|---|---|---|
| **Non Tel** | Dwt-MFCC | 10 | 640 | 84.38 | CDB_exAll_10mfcc.m |
| **Non Tel** | Dwt-MFCC | 12 | 640 | 87.19 | CDB_exAll_12mfcc.m |
| **Non Tel** | Dwt-MFCC | 14 | 640 | 91.72 | CDB _exAll_14mfcc.m |
| **Non Tel** | Dwt-MFCC | 16 | 640 | 93.81 | CDB _exAll_16mfcc.m |
| **Non Tel** | Dwt-MFCC | 18 | 640 | 95.16 | CDB _exAll_18mfcc.m |
| **Non Tel** | Dwt-MFCC | 20 | 640 | 95.31 | CDB _exAll_20mfcc.m |
| **Non Tel** | Dwt-MFCC | 22 | 640 | 95.94 | CDB _exAll_22mfcc.m |
| **Non Tel** | Dwt-MFCC | 24 | 640 | 95.31 | CDB _exAll_24mfcc.m |
| **Non Tel** | Dwt-MFCC | 26 | 640 | 96.25 | CDB _exAll_26mfcc.m |
| **Non Tel** | Dwt-MFCC | 28 | 640 | 96.25 | CDB _exAll_28mfcc.m |
| **Non Tel** | Dwt-MFCC | 30 | 640 | 95.95 | CDB _exAll_30mfcc.m |
| **Non Tel** | Dwt-MFCC | 32 | 640 | 96.09 | CDB _exAll_32mfcc.m |
| **Non Tel** | Dwt-MFCC | 34 | 640 | 95.78 | CDB _exAll_34mfcc.m |
| **Non Tel** | Dwt-MFCC | 36 | 640 | 95.47 | CDB _exAll_36mfcc.m |
| **Non Tel** | Dwt-MFCC | 38 | 640 | 96.25 | CDB _exAll_38mfcc.m |
| **Non Tel** | Dwt-MFCC | 40 | 640 | 95.47 | CDB _exAll_40mfcc.m |

**5.26 Effect of Number of MFCCs on increased Telephonic (H1) Data**

| Database | Technique | No. of MFCC | Samples | Accuracy | File Name |
|---|---|---|---|---|---|
| **Handset 1** | DWT-MFCC | 10 | 640 | 63.28 | CDB_exAll_10mfcc.m |
| **Handset 1** | DWT-MFCC | 12 | 640 | 69.22 | CDB_exAll_12mfcc.m |
| **Handset 1** | DWT-MFCC | 14 | 640 | 72.03 | CDB _exAll_14mfcc.m |
| **Handset 1** | DWT-MFCC | 16 | 640 | 75.00 | CDB _exAll_16mfcc.m |
| **Handset 1** | DWT-MFCC | 18 | 640 | 77.34 | CDB _exAll_18mfcc.m |
| **Handset 1** | DWT-MFCC | 20 | 640 | 80.00 | CDB _exAll_20mfcc.m |
| **Handset 1** | DWT-MFCC | 22 | 640 | 80.16 | CDB _exAll_22mfcc.m |
| **Handset 1** | DWT-MFCC | 24 | 640 | 81.09 | CDB _exAll_24mfcc.m |
| **Handset 1** | DWT-MFCC | 26 | 640 | 80.78 | CDB _exAll_26mfcc.m |
| **Handset 1** | DWT-MFCC | 28 | 640 | 84.22 | CDB _exAll_28mfcc.m |
| **Handset 1** | DWT-MFCC | 30 | 640 | 83.91 | CDB _exAll_30mfcc.m |
| **Handset 1** | DWT-MFCC | 32 | 640 | 83.13 | CDB _exAll_32mfcc.m |
| **Handset 1** | DWT-MFCC | 34 | 640 | 84.38 | CDB _exAll_34mfcc.m |
| **Handset 1** | DWT-MFCC | 36 | 640 | 85.16 | CDB _exAll_36mfcc.m |
| **Handset 1** | DWT-MFCC | 38 | 640 | 85.94 | CDB _exAll_38mfcc.m |
| **Handset 1** | DWT-MFCC | 40 | 640 | 87.19 | CDB _exAll_40mfcc.m |

**Table 5.27 of Number of MFCCs on increased Telephonic (H2) Data**

| Database | Technique | No. of MFCC | Samples | Accuracy | File Name |
|---|---|---|---|---|---|
| **Handset 2** | DWT-MFCC | 10 | 640 | 57.81 | CDB_exAll_10mfcc.m |
| **Handset 2** | DWT-MFCC | 12 | 640 | 64.69 | CDB_exAll_12mfcc.m |
| **Handset 2** | DWT-MFCC | 14 | 640 | 69.84 | CDB _exAll_14mfcc.m |
| **Handset 2** | DWT-MFCC | 16 | 640 | 73.59 | CDB _exAll_16mfcc.m |
| **Handset 2** | DWT-MFCC | 18 | 640 | 77.97 | CDB _exAll_18mfcc.m |
| **Handset 2** | DWT-MFCC | 20 | 640 | 81.09 | CDB _exAll_20mfcc.m |
| **Handset 2** | DWT-MFCC | 22 | 640 | 83.13 | CDB _exAll_22mfcc.m |
| **Handset 2** | DWT-MFCC | 24 | 640 | 85.31 | CDB _exAll_24mfcc.m |
| **Handset 2** | DWT-MFCC | 26 | 640 | 84.53 | CDB _exAll_26mfcc.m |
| **Handset 2** | DWT-MFCC | 28 | 640 | 84.84 | CDB _exAll_28mfcc.m |
| **Handset 2** | DWT-MFCC | 30 | 640 | 85.16 | CDB _exAll_30mfcc.m |
| **Handset 2** | DWT-MFCC | 32 | 640 | 86.41 | CDB _exAll_32mfcc.m |
| **Handset 2** | DWT-MFCC | 34 | 640 | 85.31 | CDB _exAll_34mfcc.m |
| **Handset 2** | DWT-MFCC | 36 | 640 | 84.22 | CDB _exAll_36mfcc.m |
| **Handset 2** | DWT-MFCC | 38 | 640 | 86.88 | CDB _exAll_38mfcc.m |
| **Handset 2** | DWT-MFCC | 40 | 640 | 85.78 | CDB _exAll_40mfcc.m |

**Table 5.28 Effect of Wavelet Type on Best Technique (Non Telephonic)**

| Database | Wavelet Type | No. of MFCCs | Accuracy | File name |
|---|---|---|---|---|
| **Non Telephonic** | Db1 | 38 | 95.47 | exAllb_db1_NT.m |
| **Non Telephonic** | Db2 | 38 | 95.94 | exAllb_db2_NT.m |
| **Non Telephonic** | Db3 | 38 | 95.47 | exAllb_db3_NT.m |
| **Non Telephonic** | Haar | 38 | 95.94 | exAllb_har_NT.m |
| **Non Telephonic** | Sym7 | 38 | 96.25 | exAllb_sym_NT.m |
| **Non Telephonic** | Dsicrete Meyer | 38 | 96.09 | exAllb_dm_NT.m |

**Table 5.29 Effect of Wavelet Type on Best Technique (Telephonic-H1)**

| Database | Wavelet Type | No. of MFCCs | Accuracy | Filename |
|---|---|---|---|---|
| **Handset 1** | Db1 | 38 | 86.41 | exAllb_db1_h1.m |
| **Handset 1** | Db2 | 38 | 85.78 | exAllb_db2_h1.m |
| **Handset 1** | Db3 | 38 | 84.69 | exAllb_db3_h1.m |
| **Handset 1** | Haar | 38 | 85.47 | exAllb_har_h1.m |
| **Handset 1** | Sym7 | 38 | 85.94 | exAllb_sym_h1.m |
| **Handset 1** | Discrete Meyer | 38 | 85.47 | exAllb_dm_h1.m |

**Table 5.30 Effect of Wavelet Type on Best Technique (Telephonic-H2)**

| Database | Wavelet Type | No. of MFCCs | Accuracy | Filename |
|---|---|---|---|---|
| **Handset 2** | Db1 | 38 | 87.03 | exAllb_db1_h2.m |
| **Handset 2** | Db2 | 38 | 85.78 | exAllb_db2_h2.m |
| **Handset 2** | Db3 | 38 | 87.19 | exAllb_db3_h2.m |
| **Handset 2** | Haar | 38 | 85.78 | exAllb_har_h2.m |
| **Handset 2** | Sym7 | 38 | 86.88 | exAllb_sym_h2.m |
| **Handset 2** | Discrete Meyer | 38 | 86.88 | exAllb_dm_h2.m |

**Table 5.31 Effect of Decomposition Levels on Best Technique (Non-Telephonic)**

| Database | Technique | Dec. Levels | Accuracy | Filename |
|---|---|---|---|---|
| **Non Telephonic** | DWT(sym)-MFCC[38] | 1 | 96.25 | exAllbD1_NT.m |
| **Non Telephonic** | DWT(sym)-MFCC[38] | 2 | 89.84 | exAllbD2_NT.m |
| **Non Telephonic** | DWT(sym)-MFCC[38] | 3 | 87.66 | exAllbD3_NT.m |

**Table 5.32 Effect of Decomposition Level on Best Technique (Telephonic-H1)**

| Database | Technique | Dec. Levels | Accuracy | Filename |
|---|---|---|---|---|
| **Handset-1** | DWT(sym)-MFCC[38] | 1 | 85.94 | exAllbD1_h1.m |
| **Handset-1** | DWT(sym)-MFCC[38] | 2 | 82.50 | exAllbD2_h1.m |
| **Handset-1** | DWT(sym)-MFCC[38] | 3 | 76.25 | exAllbD3_h1.m |

**5.33 Effect of Decomposition Level on Best Technique (Telephonic-H2)**

| Database | Technique | Dec. Levels | Accuracy | Filename |
|---|---|---|---|---|
| **Handset-2** | DWT(sym)-MFCC[38] | 1 | 86.88 | exAllbD1_h2.m |
| **Handset-2** | DWT(sym)-MFCC[38] | 2 | 86.56 | exAllbD2_h2.m |
| **Handset-2** | DWT(sym)-MFCC[38] | 3 | 81.09 | exAllbD3_h2.m |

### 5.3.11 Appendix-B: Graphical User Interface

For a computer human interaction, there exists a high demand of graphical user interface in order to communicate. These graphical user interfaces take input from user, perform requested task according to the input and produce the desired output. . This section presents the design and working of graphical user interface of TAURUS and its working for speaker recognition system's software.

**Error! Reference source not found.** shows the first window of graphical user interface for speaker recognition system. It has two phases

1. Enrollment
2. Identification

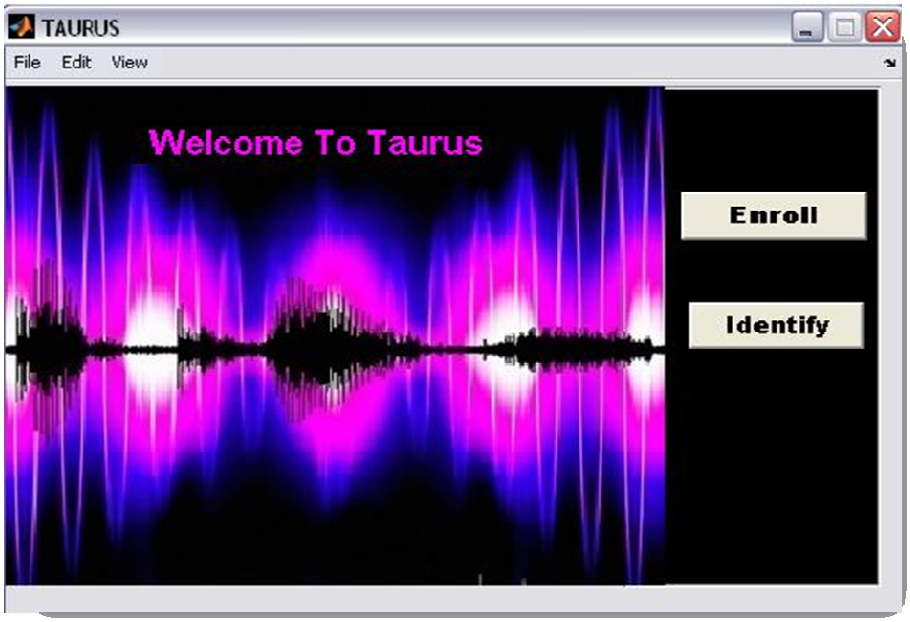



**Figure B. 1 Main Interface of System**

The Enroll button opens another window for Enrollment and similarly clicking identify button opens window for identification.

The menu bar contains File, View, and Edit options. File menu gives users an option to open the desired window, save and print the results. View helps the users see current database. Similarly Edit option lets the user copy the results or database information.

## Enrollment

On clicking enrollment menu, a new window opens in which new speakers can be enrolled. **Error! Reference source not found.** shows enrollment phase of GUI.

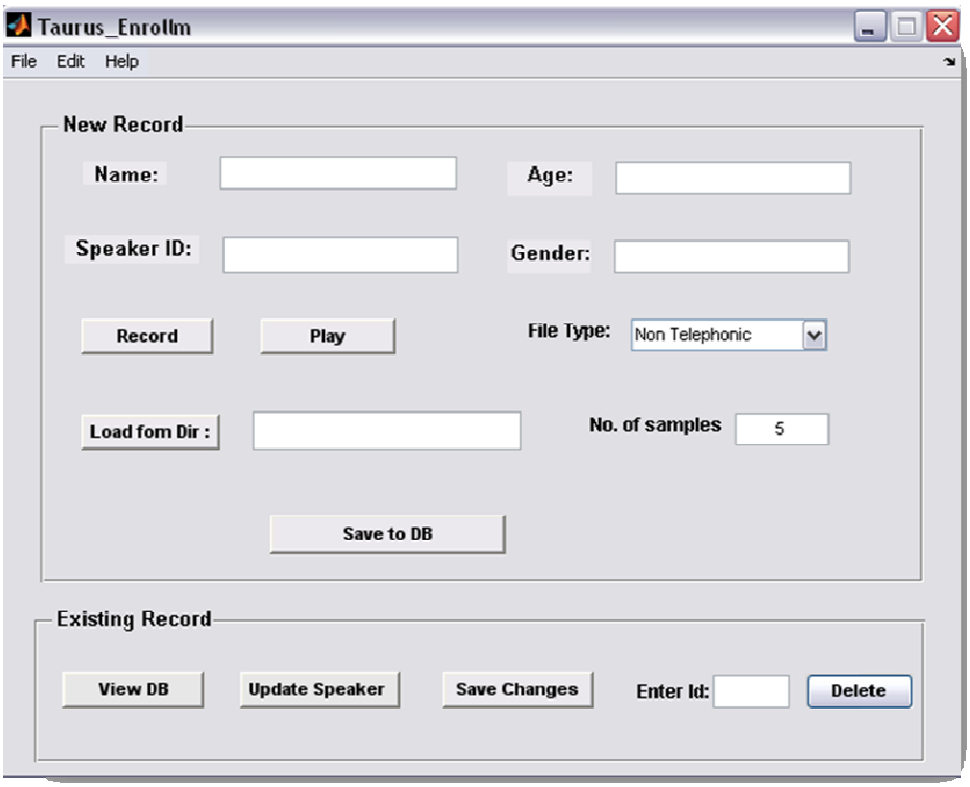



**Figure B. 2 Enrollment Phase**

The Enrollment window lets the user enroll a new speaker as well as to make changes to previous speakers.

**New Record:**

In this section a new speaker is being enrolled in by specifying the related information and files:

- Name of the speaker
- Age of person
- Gender
- ID of the person. (specified by the system administrator)

After entering the required information, voice sample of the speaker is required to enroll him in. The voice sample can either be recorded or saved on the spot or it can be loaded from pre stored samples. No. of samples define on how many speech files

does the new subject needs to be trained  Default number is 5. This can be seen in Figure B.3

Another thing that needs to be selected is where we need to enroll the new speaker, either in telephonic or non telephonic database. This can be selected via drop down menu. Lastly, when “Saved to DB” button is clicked, the model and information regarding the speaker is stored at the back end.

After the model is being saved, a confirmation message is displayed on the window stating “Saved to Model”.

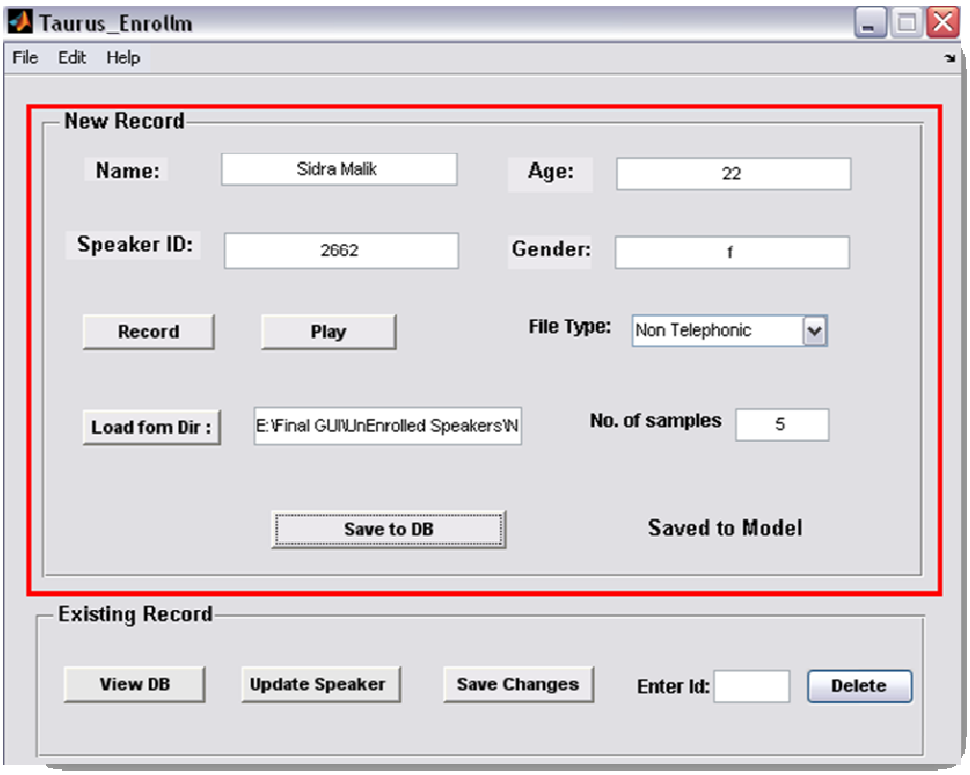



**Figure B. 3 New Record of Enrollment phase**

### Existing Record:

This section allows the user to

- View the speakers in current database
- Update the speaker’s information
- Delete a speaker

In view and edit section user can view the information of existing records and he can also update them. Moreover, on clicking save changes, the record is permanently saved to the database. Figure B.4 shows the existing record.

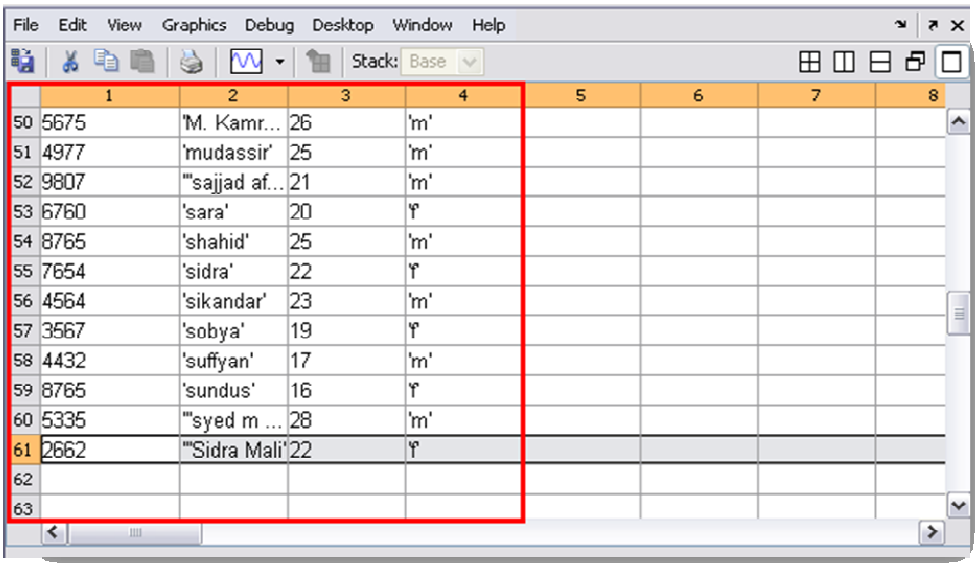



**Figure B. 4 List of Existing Speakers in Database**

On Deletion of particular user, ID of that speaker is given and it is deleted permanently as shown in Figure B.5.

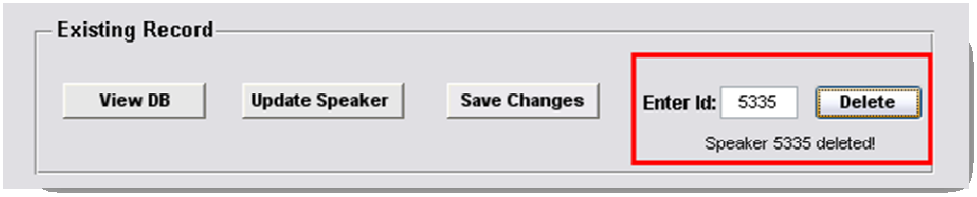



**Figure B. 5 Deletion of Existing Records**

## Identification

In this phase, user can provide a testing file either by recording a sample at that time via microphone or from a telephonic call (that is connected to the system) or the file can be taken as a stored file saved in a system.

If a new recording is made in this, time of recording must be specified. After specifying the file, the drop down menu should be used to select which identification

mode this task needs to be performed on i.e. telephonic or non telephonic. ID of person is specified.

By specifying all the above, lastly “Identify” button is clicked. On clicking this button , if the specified ID matches with the voice, “ identified successfully” is displayed on Status. This can be seen in Figure B.6.

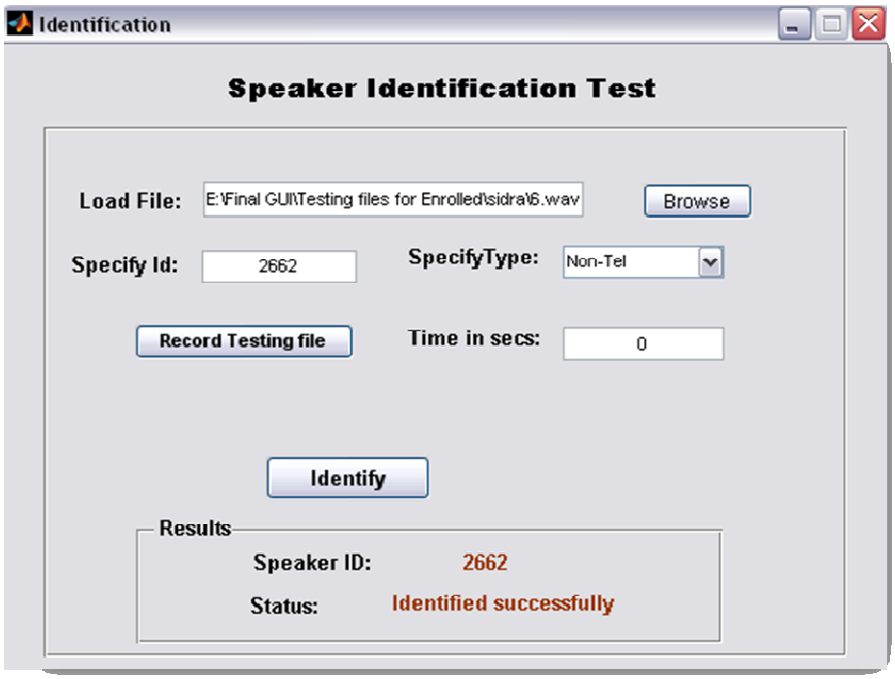



**Figure B. 6 Speaker Identification Phase**

If the speaker tries to intrude with random ID, and speaker’s voice is not matched to any of the speakers, “Enter correct ID” is displayed on screen. This is shown in Figure B.7

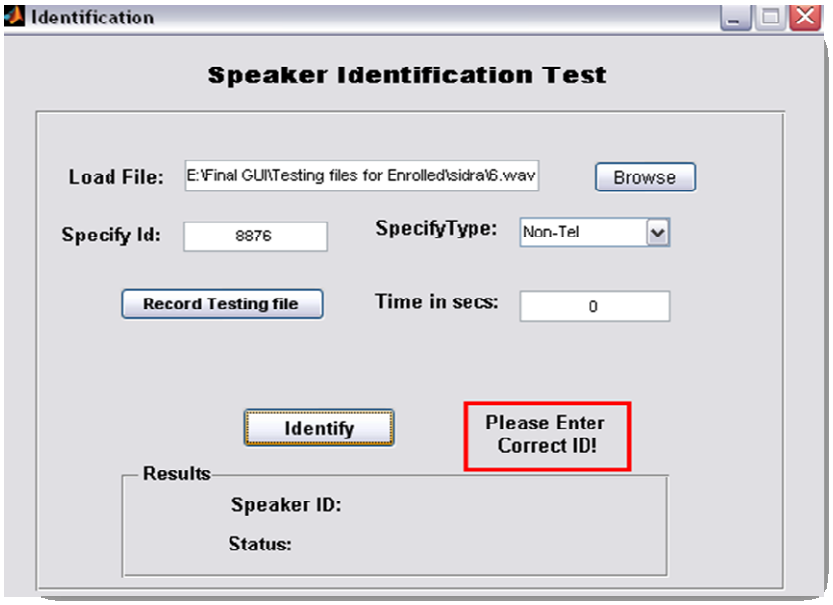



**Figure B. 7 Entering incorrect ID**

Similarly if the speaker tries to intrude with another ID that exists in database, but the voice of intruder doesn't match the ID's voice, system identifies it as an unsuccessful attempt as shown in Figure B.8.

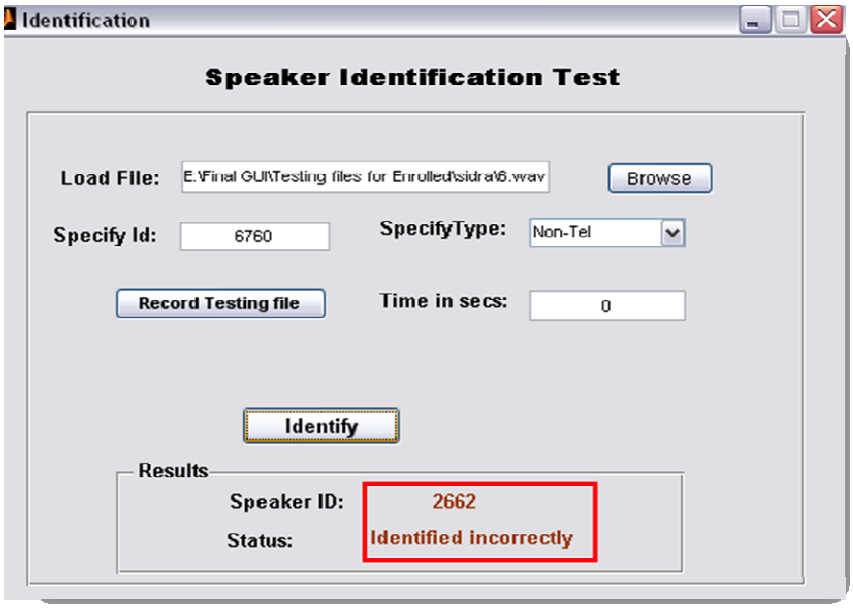



**Figure B. 8 Entering another user's ID**

# Chapter 6
# Conclusion and Future Work

This thesis attempted to study various techniques regarding automatic speaker recognition and to implement one of them to develop a system that is efficient and robust to non telephonic and telephonic data.

The various approaches have been studied as well as implemented by the end of this project. Moreover the best of these techniques has been selected to be used for developing a system.

## 6.1 Conclusion

In this approach, we have presented wavelet based MFCCs as efficient feature set for speaker recognition task. The approach was motivated by multi resolution property of wavelets in denoising speech signal and MFCCs were used to mimic the behavior of human ear which emphasizes lower frequencies. Feature set for an individual speaker is constructed using approximations from wavelet decomposition and Mel-coefficients.

Recogntion rate achieved is quite good i.e. 96.25% for non telephonic and 86.77% for telephonic speech. Moreover for the proposed technique, we analyzed to select best parameters and for PIEAS Speech Database, 38 MFCCs based on wavelet type Symlets 7 and decomposition level 1 have proven best. The performance of proposed method is comparable to approaches for feature extraction based on wavelet transform, LPCs and LPCCs as well as to speaker modeling techniques like VQ.

## 6.2 Future Directions

A possible future direction would be extending this classification to open-set and using another classification approach than minimum distortion. Also, effect of using wavelet packet transform can also be analyzed. Most important, this method can be improved to make it efficient for mismatched conditions, i.e. where the system is trained on non telephonic data and tested on telephonic data.

# References


[1] D. A. Reynolds, "An Overview of Automatic Speaker Recognition Technology", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2002)

[2] L. Feng and L. K. Hansen, "A New Database for Speaker Recognition", Informatics and Mathematical Modeling, Technical University of Denmark, (2001)

[3] S. Mallat, "A Wavelet Tour of Signal Processing", Wikipedia Organization.
URL: http://en.wikipedia.org/wiki/Discrete_Wavelet_Transform

[4] W.Khaldi, W.Fakhar,W.Hamdy, "ASR In Noisy Environments using WaveletTransform", *IEEE proc. Signal Processing*, (2002)

[5] W.Chen, C. Hsieh and E. Lai, "Multiband Approach to Robust Text-Independent Speaker Identification", *Computational Linguistics and Chinese Language Processing*, Vol. 9, No. 2 , pp. 63-76, (2004).

[6] W. Chen, C. Hsieh and E.Lai, "Robust speech features based on wavenet transform with application to speaker identification", *IEEE Proceedings online no. 20020121*, (2002)

[7] P.K Tomi, "*Spectral Features for Automatic Text-Independent Speaker Recognition*", Finland, December 21, (2003)

[8] X. Huang., A. Acero, and W. Hon. "*Spoken Language Processing: a Guide to Theory, Algorithm, and System Development*", Prentice-Hall, New Jersey, (2001)

[9] J. Makhoul, "Linear prediction: a tutorial review", *Proceedings of the IEEE* Vol-64, pp 561–580, (1975).

[10] M. Phythian, J. Ingram, and S. Sridharan, "Effects of speech coding on text-dependent speaker recognition", *In Proceedings of IEEE Region 10 Annual Conference on Speech and Image Technologies for Computing and Telecommunications (TENCON'97)* pp. 137–140, (1997).

[11] L. Besacier, S. Grassi, A. Dufaux, M. Ansorge, and F. Pellandini, "GSM speech coding and speaker recognition". *In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2000)* (Istanbul, Turkey, 2000), pp. 1085–1088.

[12] N. D. Minh, "An Automatic Speaker Recognition System." Digital Signal Processing Mini-Project. 14 June 2005.
URL : http://lcavwww.epfl.ch/~minhdo/asr_project/asr_project.html

[13] M. K. Hasan, Software Practice Organization
URL:http://www.softwarepractice.org/wiki/Team_D_Speaker_Recognition.

[15] E. Karpov, "Real-Time Speaker Identification" , *IEEE press*, ( 2003)

[16] P. Kabal, "*Time Windows for Linear Prediction of Speech*", Department of Electrical & Computer Engineering, McGill University, Canada, (2005)

[17] Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, 2000.

[18] K. Teknomo, "Kmeans Tutorial",
URL: http://people.revoledu.com/kardi/tutorial/kMean/index.html

[19] H. Gish and M. Schmidt, "*Text Independent Speaker Identification*, IEEE Signal Processing Magazine, Vol. 11, pp. 18-32, (1994)

[20] D. Reynolds, T. Quatieri, Dunn, "Speaker verfication using adapted gaussian mixture model". *Digital Signal Processing* 10 (1), 19–41.Petrovska, (1998)

[21] P. Renevey and A. Drygajl, "Entropy Based Voice Activity Detection in Very Noisy Conditions" Swiss Center for Electronics and Microtechnology, Switzerland, (2001)

[22] V. K. Prasad, T. Nagarajan, H. A. Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions", *Speech Communication 42*, pp. 429–446 (2004)

[23] W. Bing-Fei and K. Wang "Voice Activity Detection Based on Auto-Correlation Function Using Wavelet Transform and Teager Energy Operator"

[24] V.I. Galonoov, S.N.Gramantiski and others, "VQ and GMM combination for text independent speaker recognition on Telephone channel"

[25] N. Bagge, C. Donica, "ELEC 301: Final Project Text Independent Speaker Recognition", *ELEC 301,* Signals and Systems Group Projects 2001.

[26] D. A. Reynolds, R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Trans Speech Audio Processing*, vol. 3, no. 1, pp. 72-83,(1995).

[27] Chapter of Speech Enhancement, Computer Science department. URL : http://www.cs.tut.fi/sgn/arg/8003051/ehostus_en.pdf

[28] M. Westral, "*The use of Cepstral Means in Conversational Speech Recognition*", University of Karlsruhe, Germany., (2001)

[29] W.B. Kleijn and K.K. Paliwal, eds., "Speech Coding and Synthesis", Elsvier, the NetherLands, (1995)

[30] J. Veth & L. Boves "Comparison of channel Normalization Techniques for Automatic Speech Recognition Over the Phone", *Proceedings of IEEE*, pp. 2332-2334, (2002)

[31] S. Lupembe, D. Mashao, "A combination of Speaker- and Channel-Normalization for recognition of Telephone and GSM speech" *SATNAC2004*, pp. 141-144 (2004)

[32] M. Hallstein, T.Svendsen and E. Harborg2, "Experiments on Cepstral Mean Subtraction (CMS) and RASTA- filtering Applied to SAMPA Phoneme Recognition", *COST249-meeting (Continuous Speech Recognition Over the Telephone) in Nancy, France*, (1995).

[33] P. Mokhtari, "*An Acoustic-Phonetic and Articulatory Study of Speech Speaker Dichotomy"*. PhD thesis, School of Computer Science, University of New SouthWales, Canberra, Australia, 1998

[34] F. Harris "Use of windows for harmonic analysis with the discrete fourier transform". *Proceedings of the IEEE 66*, 1 pp. 51–84 (1978)

[35] Rodr´ iguez-Li˜ nares, L., Garc´ ia-Mateo, C., and Alba-Castro, J." On combining classifiers for speaker authentication", *Pattern Recognition* 36, pp. 347–359, (2003).

[36] S. Farah, " Speaker Recognition System", BS(CIS) 2003-2007, Pakistan Institute of Engineering and Applied Sciences, Islamabad, Pakistan.

[37] A. Jain, "Introduction to Biometrics", Michigan State University East Lansing, MI.

[38] J. A. Markowitz, “Using *Speech Recognition*”, Prentice Hall Publications, Upper Saddle River, New Jersey, 292 (1996).

[39] T. F. Quatieri, “Discrete Time Speech Signal Processing”, Pearson Education (Singapore), 780, (2004)

## *About Author*

Sidra Malik was born in Bahawalpur, Pakistan on 26th September, 1986. She did her matriculation in year 2003 and FSc. in year 2005 from Dominican Convent School, Bahawalpur. Later, she joined Pakistan Institute of Engineering and Applied Sciences (PIEAS) in year 2005 under the degree program for BS in Computer and Information Sciences which was completed in year 2009.

Contact Information:
Email: sidraa.malik@gmail.com