# Speaker Recognition Using MFCC and Vector Quantization Model

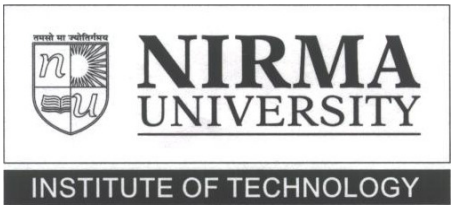
## A Major Project Report

*Submitted in Partial Fulfillment of the Requirements for the Degree of*

**Bachelor OF TECHNOLOGY**
**IN**
**ELECTRONICS & COMMUNICATION ENGINEERING**


By
DARSHAN MANDALIA (07BEC042)
PRAVIN GARETA (08BEC156)

Under the Guidance of
Prof. RACHNA SHARMA



**Department of Electrical Engineering**
**Electronics & Communication Engineering Program**
**Institute of Technology, NIRMA UNIVERSITY**
**AHMEDABAD 382 481**


**May 2011**

# *CERTIFICATE*

This is to certify that the Major Project Report entitled **"Speaker Recognition Using MFCC and Vector quantization model"** submitted by **Pravin Gareta** (Roll No. **08BEC156)** and **Darshan Mandalia** (Roll No. **07BEC042**) as the partial fulfillment of the requirements for the award of the degree of *Bachelor of Technology* in Electronics & Communication Engineering, Institute of Technology, **Nirma University** is the record of work carried out by them under my supervision and guidance. The work submitted in our opinion has reached a level required for being accepted for the examination.

**Date: 17/05/2011**

**Prof. RACHNA SHARMA**
**Project Guide**

**Prof. A. S. Ranade**
**HOD (Electrical Engineering)**
**Nirma University, Ahmedabad**

# Acknowledgement

In order to achieve better performance, we should have to learn our outside environment. There are lots of forces which will acting upon us to get the better result, but for that we have to change our attitude to see them. There are lots of problems we facing, but due to it we don't stop. It is not a good human being ,if he is stop. Yes, we are sometimes frustrated due to the problems we can't solve it. But the next day we act on the problem with same efficiency and strength we have. Moreover, We are highly thankful to our parents and teachers were there all the times, backing us up and with their undue support has been the pushing drive for us to complete project within time.

We would deeply like to express our sincere gratitude towards project guide Prof. Rachna Sharma and faculty members of our panel who have guided until the completion of the project. We also extend our thanks towards our Head of the Department Prof. A.S.Ranade and all the staff, Department of Electronics and Communication Engineering, the Institute of Technology, Nirma University for their assistance during the project whether it was a technical help or concerned with providing facilities for internet, implementing and simulating the ideas of the project. Their excessive support has been the source of motivation to perform our best regarding the project.

We would like to express our gratitude towards our parents who have been there not only in this project but through all our entire life. If their helping hand, moral as well as financial support, had not been there we wouldn't have been able to finish in such a proficient way. We are grateful for their aid and support.

**Pravin Gareta (08BEC156)**

**Darshan Mandalia (07BEC042)**

# Abstract

Speech Recognition is the process of automatically recognizing a certain word spoken by a particular speaker based on individual information included in speech waves. This technique makes it possible to use the speaker‟s voice to verify his/her identity and provide controlled access to services like voice based biometrics, database access services, voice based dialing, voice mail and remote access to computers.

Signal processing front end for extracting the feature set is an important stage in any speech recognition system. The optimum feature set is still not yet decided though the vast efforts of researchers. There are many types of features, which are derived differently and have good impact on the recognition rate. This project presents one of the techniques to extract the feature set from a speech signal, which can be used in speech recognition systems.

The key is to convert the speech waveform to some type of parametric representation (at a considerably lower information rate) for further analysis and processing. This is often referred as the *signal-processing front end.* A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as *Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC),* and others. MFCC is perhaps the best known and most popular, and these will be used in this project. MFCCs are based on the known variation of the human ear‟s critical bandwidths with frequency filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. However, another key characteristic of speech is quasi-stationarity, i.e. it is short time stationary which is studied and analyzed using short time, frequency domain analysis.

To achieve this, we have first made a comparative study of the MFCC approach. The voice based biometric system is based on isolated or single word recognition. A particular speaker utters the password once in the training session so as to train and store the features of the access word. Later in the testing session the user utters the password again in order to achieve recognition if there is a match. The feature vectors unique to that speaker are obtained in the training phase and this is made use of later on to grant authentication to the same speaker who once again utters the same word in the testing phase. At this stage an intruder can also test the system to test the inherent security feature by uttering the same word .

# Index

# List of Figures

# Chapter 1

# Introduction

## 1.1 Introduction

Speech is the most natural way to communicate for humans. While this has been true since the dawn of civilization, the invention and widespread use of the telephone, audio-phonic storage media, radio, and television has given even further importance to speech communication and speech processing [2]. The advances in digital signal processing technology has led the use of speech processing in many different application areas like speech compression, enhancement, synthesis, and recognition [4]. In this thesis, the issue of speech recognition is studied and a speech recognition system is developed for Isolated word using Vector quantization model.

The concept of a machine than can recognize the human voice has long been an accepted feature in Science Fiction. From „Star Trek" to George Orwell"s „1984" - "Actually he was not used to writing by hand. Apart from very short notes, it was usual to dictate everything into the speakwriter." - It has been commonly assumed that one day it will be possible to converse naturally with an advanced computer-based system. Indeed in his book „The Road Ahead", Bill Gates (co-founder of Microsoft Corp.) hails Automatic Speaker Recognition (ASR) as one of the most important innovations for future computer operating systems.

From a technological perspective it is possible to distinguish between two broad types of ASR: direct voice input" (DVI) and „large vocabulary continuous speech recognition" (LVCSR). DVI devices are primarily aimed at voice command-and-control, whereas LVCSR systems are used for form filling or voice-based document creation. In both cases the underlying technology is more or less the same. DVI systems are typically configured for small to medium sized vocabularies (up to several thousand words) and might employ word or phrase spotting techniques. Also, DVI systems are usually required to respond immediately to a voice command. LVCSR systems involve vocabularies of perhaps hundreds of thousands of words, and are typically configured to transcribe continuous speech. Also, LVCSR need not be performed in real-time - for example, at least one vendor has offered a telephone-based dictation service in which the transcribed document is e-mailed back to the user.

From an application viewpoint, the benefits of using ASR derive from providing an extra communication channel in hands-busy eyes-busy human-machine interaction (HMI), or simply from the fact that talking can be faster than typing.

## 1.2 Motivation

The motivation for ASR is simple; it is man's principle means of communication and is, therefore, a convenient and desirable mode of communication with machines. Speech communication has evolved to be efficient and robust and it is clear that the route to computer based speech recognition is the modeling of the human system. Unfortunately from pattern recognition point of view, human recognizes speech through a very complex interaction between many levels of processing; using syntactic and semantic information as well very powerful low level pattern classification and processing. Powerful classification algorithms and sophisticated front ends are, in the final analysis, not enough; many other forms of knowledge, e.g. linguistic, semantic and pragmatic, must be built into the recognizer. Nor, even at a lower level of sophistication, is it sufficient merely to generate "a good" representation of speech (i.e. a good set of features to be used in a pattern classifier); the classifier itself must have a considerable degree of sophistication. It is the case, however, it do not effectively discriminate between classes and, further, that the better the features the easier is the classification task.

Automatic speech recognition is therefore an engineering compromise between the ideal, i.e. a complete model of the human, and the practical, i.e. the tools that science and technology provide and that costs allow.

At the highest level, all speaker recognition systems contain two main modules (refer to Fig 1.1): feature extraction and feature matching. Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing extracted features from his/her voice input with the ones from a set of known speakers. We will discuss each module in detail in later sections.

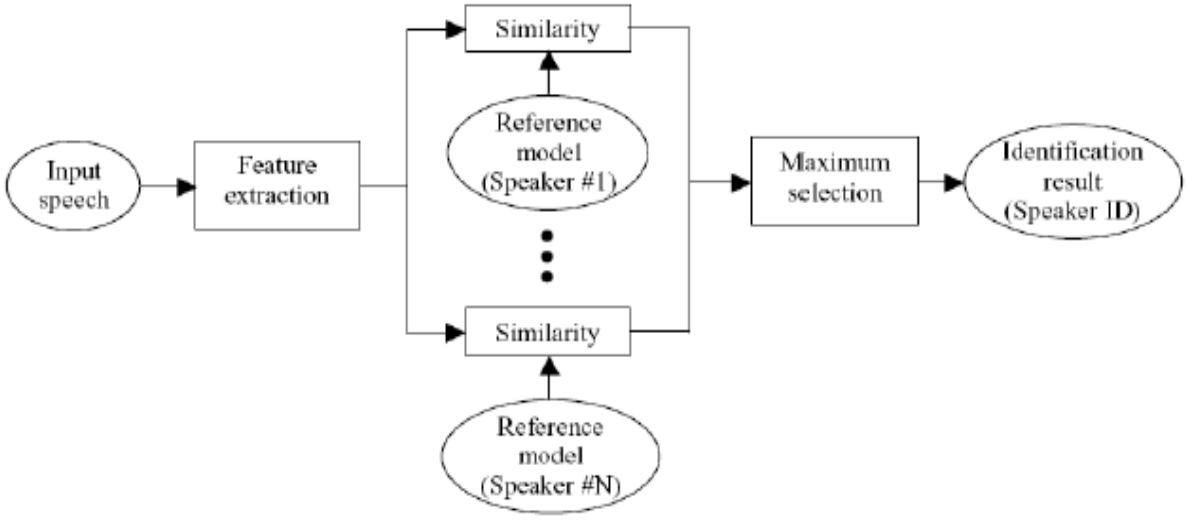



**Fig. 1.1. Speaker Identification Training**

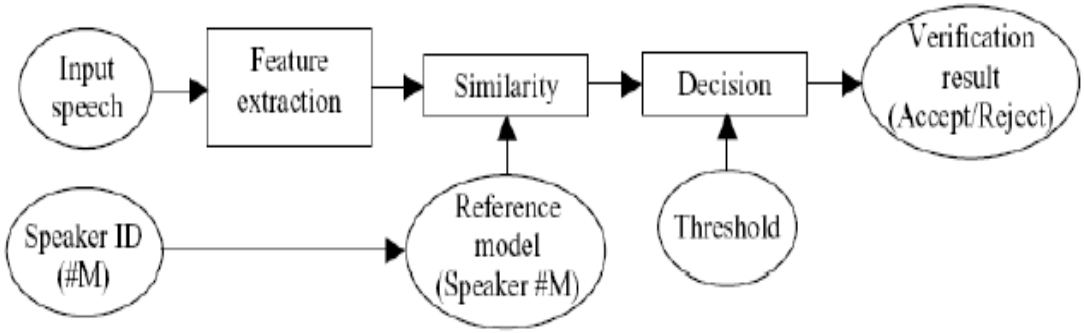



**Fig. 1.2. Speaker Identification Testing**

All Recognition systems have to serve two different phases. The first one is referred to the enrollment sessions or training phase while the second one is referred to as the operation sessions or testing phase. In the training phase, each registered speaker has to provide samples of their speech so that the system can build or train a reference model for that speaker. In case of speaker verification systems, in addition, a speaker-specific threshold is also computed from the training

samples. During the testing (operational) phase (see Figure 1.2), the input speech is matched with stored reference model(s) and recognition decision is made.

Speech recognition is a difficult task and it is still an active research area. Automatic speech recognition works based on the premise that a person's speech exhibits characteristics that are unique to the speaker. However this task has been challenged by the highly variant of input speech signals. The principle source of variance is the speaker himself. Speech signals in training and testing sessions can be greatly different due to many facts such as people voice change with time, health conditions (e.g. the speaker has a cold), speaking rates, etc. There are also other factors, beyond speaker variability, that present a challenge to speech recognition technology.

Examples of these are acoustical noise and variations in recording environments (e.g. speaker uses different telephone handsets). The challenge would be make the system "**Robust**".

So what characterizes a "**Robust System**"? When people use an automatic speech recognition (ASR) system in real environment, they always hope it can achieve as good recognition performance as human's ears do which can constantly adapt to the environment characteristics such as the speaker, the background noise and the transmission channels. Unfortunately, at present, the capacities of adapting to unknown conditions on machines are greatly poorer than that of ours. In fact, the performance of speech recognition systems trained with clean speech may degrade significantly in the real world because of the mismatch between the training and testing environments. If the recognition accuracy does not degrade very much under mismatch conditions, the system is called "**Robust**".

## 1.3 Objective

The objective of the project is to Design a Speaker recognition model using MFCC extraction technique and also with Vector quantization model.

## Outline of thesis

This thesis is organized as follows.

In chapter 2, Introduction about basic speech signal.

In chapter 3, Introduction about automatic speaker recognition system is discussed.

In chapter 4, Feature extraction method is discussed.

In chapter 5, Algorithm used in this thesis is discussed.

In Chapter 6, GUI of all three methods will be given.

In chapter 7, Source code, conclusion, application of these project and Scope for future work will be discussed.

# Chapter 2

## Basic Acoustics and Speech Signal

As relevant background to the field of speech recognition, this chapter intends to discuss how the speech signal is produced and perceived by human beings. This is an essential subject that has to be considered before one can pursue and decide which approach to use for speech recognition.

## 2.1 The Speech Signal

Human communication is to be seen as a comprehensive diagram of the process from speech production to speech perception between the talker and listener, See Figure 2.1.

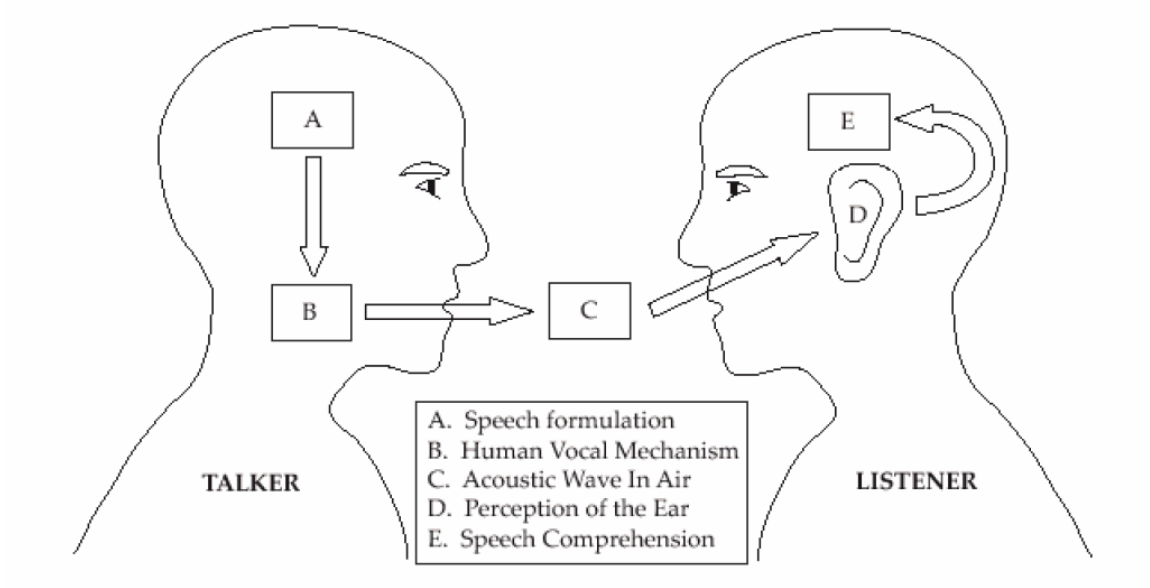



**Fig. 2.1. Schematic Diagram of the Speech Production/Perception Process**

Five different elements, A. Speech formulation, B. Human vocal mechanism, C. Acoustic air, D. Perception of the ear, E. Speech comprehension.

The first element (A. Speech formulation) is associated with the formulation of the speech signal in the talker's mind. This formulation is used by the human vocal mechanism (B. Human vocal mechanism) to produce the actual speech waveform. The waveform is transferred via the air (C.

Acoustic air) to the listener. During this transfer the acoustic wave can be affected by external sources, for example noise, resulting in a more complex waveform. When the wave reaches the listener's hearing system (the ears) the listener percepts the waveform (D. Perception of the ear) and the listener's mind (E. Speech comprehension) starts processing this waveform to comprehend its content so the listener understands what the talker is trying to tell him.
One issue with speech recognition is to "simulate" how the listener process the speech produced by the talker. There are several actions taking place in the listeners head and hearing system during the process of speech signals. The perception process can be seen as the inverse of the speech production process.

The basic theoretical unit for describing how to bring linguistic meaning to the formed speech, in the mind, is called phonemes. Phonemes can be grouped based on the properties of either the time waveform or frequency characteristics and classified in different sounds produced by the human vocal tract.
Speech is:

- Time-varying signal,
- Well-structured communication process,
- Depends on known physical movements,
- Composed of known, distinct units (phonemes),
- Is different for every speaker,
- May be fast, slow, or varying in speed,
- May have high pitch, low pitch, or be whispered,
- Has widely-varying types of environmental noise,
- May not have distinct boundaries between units (phonemes),
- Has an unlimited number of words.

## 2.2 Speech Production

To be able to understand how the production of speech is performed one need to know how the human's vocal mechanism is constructed, see Figure II.2 . The most important parts of the human vocal mechanism are the vocal tract together with nasal cavity, which begins at the velum. The velum is a trapdoor-like mechanism that is used to formulate nasal sounds when needed. When the velum is lowered, the nasal cavity is coupled together with the vocal tract to formulate the desired speech signal. The crosssectional area of the vocal tract is limited by the tongue, lips, jaw and velum and varies from 0-20 cm2.

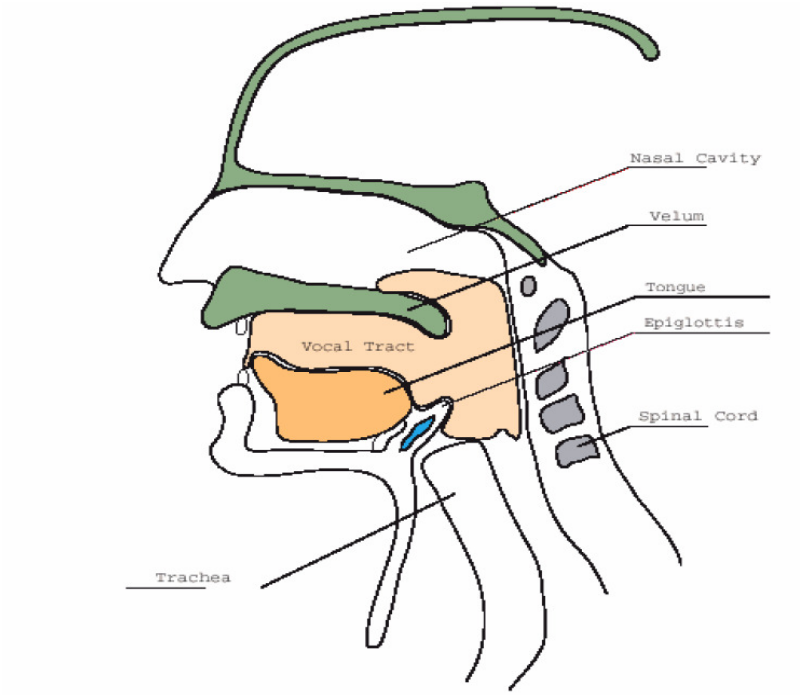



**Fig. 2.2. Human Vocal Mechanism**

## 2.3 Properties of Human Voice

One of the most important parameter of sound is its frequency. The sounds are discriminated from each other by the help of their frequencies. When the frequency of a sound increases, the sound gets high-pitched and irritating. When the frequency of a sound decreases, the sound gets deepen. Sound waves are the waves that occur from vibration of the materials. The highest value

of the frequency that a human can produce is about 10 kHz. And the lowest value is about 70 Hz. These are the maximum and minimum values. This frequency interval changes for every person. And the magnitude of a sound is expressed in decibel (dB). A normal human speech has a frequency interval of 100Hz - 3200Hz and its magnitude is in the range of 30 dB - 90 dB. A human ear can perceive sounds in the frequency range between 16 Hz and 20 kHz. And a frequency change of 0.5 % is the sensitivity of a human ear.

Speaker Characteristics,

- Due to the differences in vocal tract length, male, female, and children's speech are different.
- Regional accents are the differences in resonant frequencies, durations, and pitch.
- Individuals have resonant frequency patterns and duration patterns that are unique (allowing us to identify speaker).

➢ Training on data from one type of speaker automatically "learns" that group or person's characteristics, makes recognition of other speaker types much worse.

# Chapter 3

# Automatic Speech Recognition System(ASR)

## 3.1 Introduction

Speech processing is the study of speech signals and the processing methods of these signals. The signals are usually processed in a digital representation whereby speech processing can be seen as the interaction of digital signal processing and natural language processing. Natural language processing is a subfield of artificial intelligence and linguistics. It studies the problems of automated generation and understanding of natural human languages. Natural language generation systems convert information from computer databases into normal-sounding human language, and natural language understanding systems convert samples of human language into more formal representations that are easier for computer programs to manipulate.

## Speech coding:

It is the compression of speech (into a code) for transmission with speech codecs that use audio signal processing and speech processing techniques. The techniques used are similar to that in audio data compression and audio coding where knowledge in psychoacoustics is used to transmit only data that is relevant to the human auditory system. For example, in narrow band speech coding, only information in the frequency band of 400 Hz to 3500 Hz is transmitted but the reconstructed signal is still adequate for intelligibility.

However, speech coding differs from audio coding in that there is a lot more statistical information available about the properties of speech. In addition, some auditory information which is relevant in audio coding can be unnecessary in the speech coding context. In speech coding, the most important criterion is preservation of intelligibility and "pleasantness" of speech, with a constrained amount of transmitted data.

It should be emphasized that the intelligibility of speech includes, besides the actual literal content, also speaker identity, emotions, intonation, timbre etc. that are all important for perfect intelligibility. The more abstract concept of pleasantness of degraded speech is a different property than intelligibility, since it is possible that degraded speech is completely intelligible, but subjectively annoying to the listener.

## Speech synthesis:

Speech synthesis is the artificial production of human speech. A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech. Synthesized speech can also be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or diaphones provides the largest output range, but may lack clarity. For specific usage domains, the storage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output.

The quality of a speech synthesizer is judged by its similarity to the human voice, and by its ability to be understood. An intelligible text-to-speech program allows people with visual impairments or reading disabilities to listen to written works on a home computer. Many computer operating systems have included speech synthesizers since the early 1980s.

## Voice analysis:

Voice problems that require voice analysis most commonly originate from the vocal cords since it is the sound source and is thus most actively subject to tiring. However, analysis of the vocal cords is physically difficult. The location of the vocal cords effectively prohibits direct measurement of movement. Imaging methods such as x-rays or ultrasounds do not work because the vocal cords are surrounded by cartilage which distorts image quality. Movements in the vocal cords are rapid, fundamental frequencies are usually between 80 and 300 Hz, thus preventing usage of ordinary video. High-speed videos provide an option but in order to see the vocal cords the camera has to be positioned in the throat which makes speaking rather difficult.

Most important indirect methods are inverse filtering of sound recordings and electroglottographs (EGG). In inverse filtering methods, the speech sound is recorded outside the mouth and then filtered by a mathematical method to remove the effects of the vocal tract. This method produces an estimate of the waveform of the pressure pulse which again inversely

indicates the movements of the vocal cords. The other kind of inverse indication is the electroglottographs, which operates with electrodes attached to the subject's throat close to the vocal cords. Changes in conductivity of the throat indicate inversely how large a portion of the vocal cords are touching each other. It thus yields one-dimensional information of the contact area. Neither inverse filtering nor EGG is thus sufficient to completely describe the glottal movement and provide only indirect evidence of that movement.

## Speech recognition:

Speech recognition is the process by which a computer (or other type of machine) identifies spoken words. Basically, it means talking to your computer, and having it correctly recognize what you are saying. This is the key to any speech related application.

As shall be explained later, there are a number ways to do this but the basic principle is to somehow extract certain key features from the uttered speech and then treat those features as the key to recognizing the word when it is uttered again.

## 3.2 Speech Recognition Basics

### Utterance

An utterance is the vocalization (speaking) of a word or words that represent a single meaning to the computer. Utterances can be a single word, a few words, a sentence, or even multiple sentences.

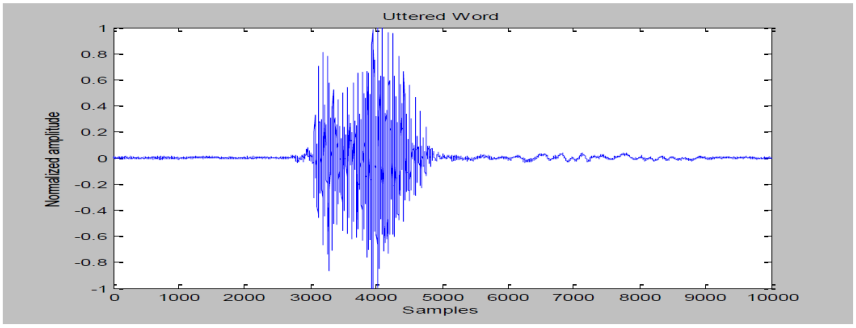



**Fig. 3.1. Utterance of "HELLO"**

## Speaker Dependence

Speaker dependent systems are designed around a specific speaker. They generally are more accurate for the correct speaker, but much less accurate for other speakers. They assume the speaker will speak in a consistent voice and tempo. Speaker independent systems are designed for a variety of speakers. Adaptive systems usually start as speaker independent systems and utilize training techniques to adapt to the speaker to increase their recognition accuracy.

## Vocabularies

Vocabularies (or dictionaries) are lists of words or utterances that can be recognized by the SR system. Generally, smaller vocabularies are easier for a computer to recognize, while larger vocabularies are more difficult. Unlike normal dictionaries, each entry doesn't have to be a single word. They can be as long as a sentence or two. Smaller vocabularies can have as few as 1 or 2 recognized utterances (e.g.” Wake Up"), while very large vocabularies can have a hundred thousand or more!

## Accuracy

The ability of a recognizer can be examined by measuring its accuracy - or how well it recognizes utterances. This includes not only correctly identifying an utterance but also identifying if the spoken utterance is not in its vocabulary. Good ASR systems have an accuracy of 98% or more! The acceptable accuracy of a system really depends on the application.

## Training

Some speech recognizers have the ability to adapt to a speaker. When the system has this ability, it may allow training to take place. An ASR system is trained by having the speaker repeat standard or common phrases and adjusting its comparison algorithms to match that particular speaker. Training a recognizer usually improves its accuracy.

Training can also be used by speakers that have difficulty speaking, or pronouncing certain words. As long as the speaker can consistently repeat an utterance, ASR systems with training should be able to adapt

## 3.3 Classification of ASR System

A speech recognition system can operate in many different conditions such as speaker dependent/independent, isolated/continuous speech recognition, for small/large vocabulary. Speech recognition systems can be separated in several different classes by describing what types of utterances they have the ability to recognize. These classes are based on the fact that one of the difficulties of ASR is the ability to determine when a speaker starts and finishes an utterance. Most packages can fit into more than one class, depending on which mode they're using.

### Isolated Words

Isolated word recognizers usually require each utterance to have quiet (lack of an audio signal) on BOTH sides of the sample window. It doesn't mean that it accepts single words, but does require a single utterance at a time. Often, these systems have "Listen/Not-Listen" states, where they require the speaker to wait between utterances (usually doing processing during the pauses). Isolated Utterance might be a better name for this class.

### Connected Words

Connect word systems (or more correctly 'connected utterances') are similar to Isolated words, but allow separate utterances to be 'run-together' with a minimal pause between them.

### 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 allow users to speak almost naturally, while the computer determines the content. Basically, it's computer dictation.

## Spontaneous Speech

There appears to be a variety of definitions for what spontaneous speech actually is. At a basic level, it can be thought of as speech that is natural sounding and not rehearsed. An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together, "ums" and "ahs", and even slight stutters.

## Speaker Dependence

ASR engines can be classified as speaker dependent and speaker independent. Speaker Dependent systems are trained with one speaker and recognition is done only for that speaker. Speaker Independent systems are trained with one set of speakers. This is obviously much more complex than speaker dependent recognition. A problem of intermediate complexity would be to train with a group of speakers and recognize speech of a speaker within that group. We could call this speaker group dependent recognition.

## 3.4 Why is Automatic Speaker Recognition hard?

There are a few problems in speech recognition that haven‟t yet been discovered. However there are a number of problems that have been identified over the past few decades most of which still remain unsolved. Some of the main problems in ASR are:

## Determining word boundaries

Speech is usually continuous in nature and word boundaries are not clearly defined. One of the common errors in continuous speech recognition is the missing out of a minuscule gap between words. This happens when the speaker is speaking at a high speed.

## Varying Accents

People from different parts of the world pronounce words differently. This leads to errors in ASR. However this is one problem that is not restricted to ASR but which plagues human listeners too.

## Large vocabularies

When the number of words in the database is large, similar sounding words tend to cause a high amount of error i.e. there is a good probability that one word is recognized as the other.

### Changing Room Acoustics

Noise is a major factor in ASR. In fact it is in noisy conditions or in changing room acoustic that the limitations of present day ASR engines become prominent.

## Temporal Variance

Different speakers speak at different speeds. Present day ASR engines just cannot adapt to that.

## 3.5 Speech Analyzer

Speech analysis, also referred to as front-end analysis or feature extraction, is the first step in an automatic speech recognition system. This process aims to extract acoustic features from the speech waveform. The output of front-end analysis is a compact, efficient set of parameters that represent the acoustic properties observed from input speech signals, for subsequent utilization by acoustic modeling.

There are three major types of front-end processing techniques, namely linear predictive coding (LPC), mel-frequency cepstral coefficients (MFCC), and perceptual linear prediction (PLP), where the latter two are most commonly used in state-of-the-art ASR systems.

## Linear predictive coding

LPC starts with the assumption that a speech signal is produced by a buzzer at the end of a tube (voiced sounds), with occasional added hissing and popping sounds. Although apparently crude, this model is actually a close approximation to the reality of speech production. The glottis (the space between the vocal cords) produces the buzz, which is characterized by its intensity (loudness) and frequency (pitch). The vocal tract (the throat and mouth) forms the tube, which is characterized by its resonances, which are called formants. Hisses and pops are generated by the action of the tongue, lips and throat during sibilants and plosives.

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 after the subtraction of the filtered modeled signal is called the residue.The numbers which describe the intensity and frequency of the buzz, the formants, and the residue signal, can be stored or transmitted somewhere else. LPC synthesizes the speech signal by reversing the process: use the buzz parameters and the residue to create a source signal, use the formants to create a filter (which represents the tube), and run the source through the filter, resulting in speech.

Because speech signals vary with time, this process is done on short chunks of the speech signal, which are called frames; generally 30 to 50 frames per second give intelligible speech with good compression.

## Mel Frequency Cepstrum Coefficients

These are derived from a type of cepstral representation of the audio clip (a "spectrum-of-a-spectrum"). The difference between the cepstrum and the Mel-frequency cepstrum is that in the MFC, the frequency bands are positioned logarithmically (on the mel scale) which approximates the human auditory system's response more closely than the linearly-spaced frequency bands obtained directly from the FFT or DCT. This can allow for better processing of data, for example, in audio compression. However, unlike the sonogram, MFCCs lack an outer ear model and, hence, cannot represent perceived loudness accurately.

MFCCs are commonly derived as follows:

1. Take the Fourier transform of (a windowed excerpt of) a signal
2. Map the log amplitudes of the spectrum obtained above onto the Mel scale, using triangular overlapping windows.
3. Take the Discrete Cosine Transform of the list of Mel log-amplitudes, as if it were a signal.
4. The MFCCs are the amplitudes of the resulting spectrum.

## Perceptual Linear Prediction

Perceptual linear prediction, similar to LPC analysis, is based on the short-term spectrum of speech. In contrast to pure linear predictive analysis of speech, perceptual linear prediction (PLP) modifies the short-term spectrum of the speech by several psychophysically based transformations.

This technique uses three concepts from the psychophysics of hearing to derive an estimate of the auditory spectrum:

(1) The critical-band spectral resolution,

(2) The equal-loudness curve, and

(3) The intensity-loudness power law.

The auditory spectrum is then approximated by an autoregressive all-pole model. In comparison with conventional linear predictive (LP) analysis, PLP analysis is more consistent with human hearing.

## 3.6 Speech Classifier

The problem of ASR belongs to a much broader topic in scientific and engineering so called *pattern recognition*. The goal of pattern recognition is to classify objects of interest into one of a number of categories or classes. The objects of interest are generically called *patterns* and in our case are sequences of acoustic vectors that are extracted from an input speech using the techniques described in the previous section. The classes here refer to individual speakers. Since the classification procedure in our case is applied on extracted features, it can be also referred to as *feature matching*.

The state-of-the-art in feature matching techniques used in speaker recognition includes Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ).

## Dynamic Time Warping

Dynamic time warping is an algorithm for measuring similarity between two sequences which may vary in time or speed. For instance, similarities in walking patterns would be detected, even

if in one video the person was walking slowly and if in another they were walking more quickly, or even if there were accelerations and decelerations during the course of one observation. DTW has been applied to video, audio, and graphics -indeed, any data which can be turned into a linear representation can be analyzed with DTW.

A well known application has been automatic speech recognition, to cope with different speaking speeds. In general, it is a method that allows a computer to find an optimal match between two given sequences (e.g. time series) with certain restrictions, i.e. the sequences are "warped" non-linearly to match each other. This sequence alignment method is often used in the context of hidden Markov models.

## Hidden Markov Model

The basic principle here is to characterize words into probabilistic models wherein the various phonemes which contribute to the word represent the states of the HMM while the transition probabilities would be the probability of the next phoneme being uttered (ideally 1.0). Models for the words which are part of the vocabulary are created in the training phase.

Now, in the recognition phase when the user utters a word it is split up into phonemes as done before and it‟s HMM is created. After the utterance of a particular phoneme, the most probable phoneme to follow is found from the models which had been created by comparing it with the newly formed model. This chain from one phoneme to another continues and finally at some point we have the most probable word out of the stored words which the user would have uttered and thus recognition is brought about in a finite vocabulary system. Such a probabilistic system would be more efficient than just cepstral analysis as these is some amount of flexibility in terms of how the words are uttered by the users.

## Vector Quantization

VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a centroid. The collection of all codewords is called a codebook.

Fig 3.2 shows a conceptual diagram to illustrate this recognition process. In the figure, only two speakers and two dimensions of the acoustic space are shown. The circles refer to the acoustic vectors from the speaker 1 while the triangles are from the speaker 2. In the training phase, a speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors. The result codewords (centroids) are shown in Figure by black circles and black triangles for speaker 1 and 2, respectively. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of an unknown voice is "vector-quantized" using each trained codebook and the total VQ distortion is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified.

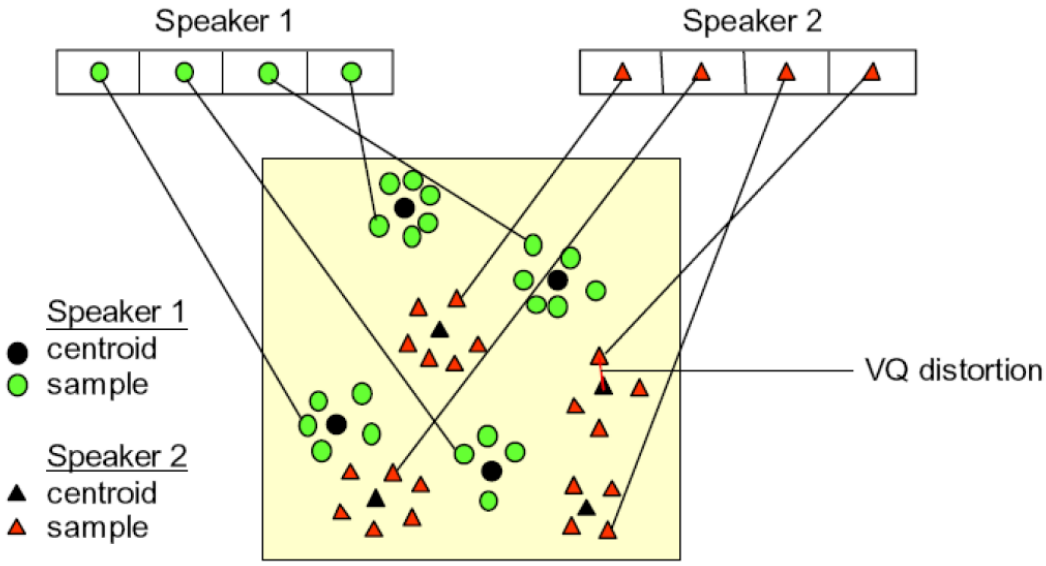



**Fig. 3.2. Conceptual diagram illustrating vector quantization codebook formation**.

One speaker can be discriminated from another based of the location of centroids.

# Chapter 4

# Feature Extraction

## 4.1. Processing

Obtaining the acoustic characteristics of the speech signal is referred to as Feature Extraction. Feature Extraction is used in both training and recognition phases.

It comprise of the following steps:

1. Frame Blocking
2. Windowing
3. FFT (Fast Fourier Transform)
4. Mel-Frequency Wrapping
5. Cepstrum (Mel Frequency Cepstral Coefficients)

## Feature Extraction

This stage is often referred as speech processing front end. The main goal of Feature Extraction is to simplify recognition by summarizing the vast amount of speech data without losing the acoustic properties that defines the speech [12]. The schematic diagram of the steps is depicted in Figure 4.1.

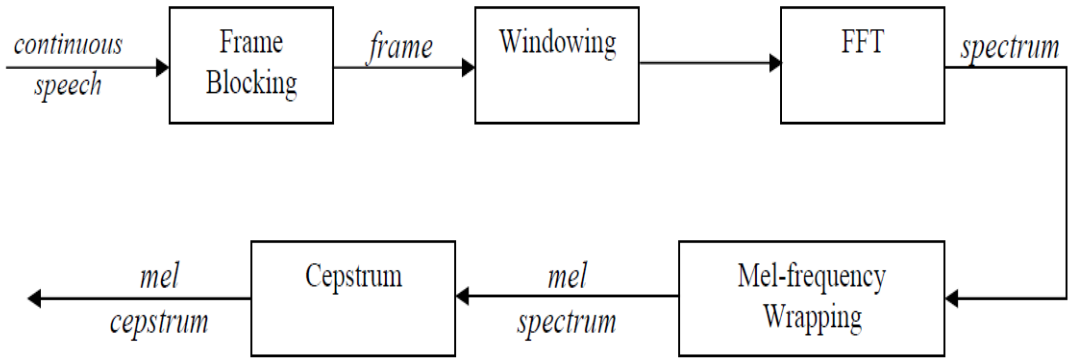



**Fig. 4.1. Feature Extraction Steps**.

### 4.1.1 Frame Blocking

Investigations show that speech signal characteristics stays stationary in a sufficiently short period of time interval (It is called quasi-stationary). For this reason, speech signals are processed in short time intervals. It is divided into frames with sizes generally between 30 and 100 milliseconds. Each frame overlaps its previous frame by a predefined size. The goal of the overlapping scheme is to smooth the transition from frame to frame [12].

### 4.1.2 Windowing

The second step is to window all frames. This is done in order to eliminate discontinuities at the edges of the frames. If the windowing function is defined as w(n), $0 < n < N-1$ where N is the number of samples in each frame, the resulting signal will be; $y(n) = x(n)w(n)$. Generally hamming windows are used [12].

### 4.1.3 Fast Fourier Transform

The next step is to take Fast Fourier Transform of each frame. This transformation is a fast way of Discrete Fourier Transform and it changes the domain from time to frequency [12].

### 4.1.4 Mel Frequency Warping

The human ear perceives the frequencies non-linearly. Researches show that the scaling is linear up to 1 kHz and logarithmic above that. The Mel-Scale (Melody Scale) filter bank which characterizes the human ear perceiveness of frequency. It is used as a band pass filtering for this stage of identification. The signals for each frame is passed through Mel-Scale band pass filter to mimic the human ear [17][12][18].

As mentioned above, psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, *f*, measured in Hz, a subjective pitch is measured on a scale called the „mel" scale. The *mel-frequency* scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore we can use the following approximate formula to compute the mels for a given frequency *f* in Hz:

$$mel(f) = 2595 * \log_{10}(1 + f / 700)$$

One approach to simulating the subjective spectrum is to use a filter bank, one filter for each desired mel-frequency component. That filter bank has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel-frequency interval. The modified spectrum of *S(□)* thus consists of the output power of these filters when *S(□)* is the input. The number of mel cepstral coefficients, *K*, is typically chosen as 20.

Note that this filter bank is applied in the frequency domain; therefore it simply amounts to taking those triangle-shape windows in the Fig 4.2 on the spectrum. A useful way of thinking about this mel-warped filter bank is to view each filter as a histogram bin (where bins have overlap) in the frequency domain. A useful and efficient way of implementing this is to consider these triangular filters in the Mel scale where they would in effect be equally spaced filters.

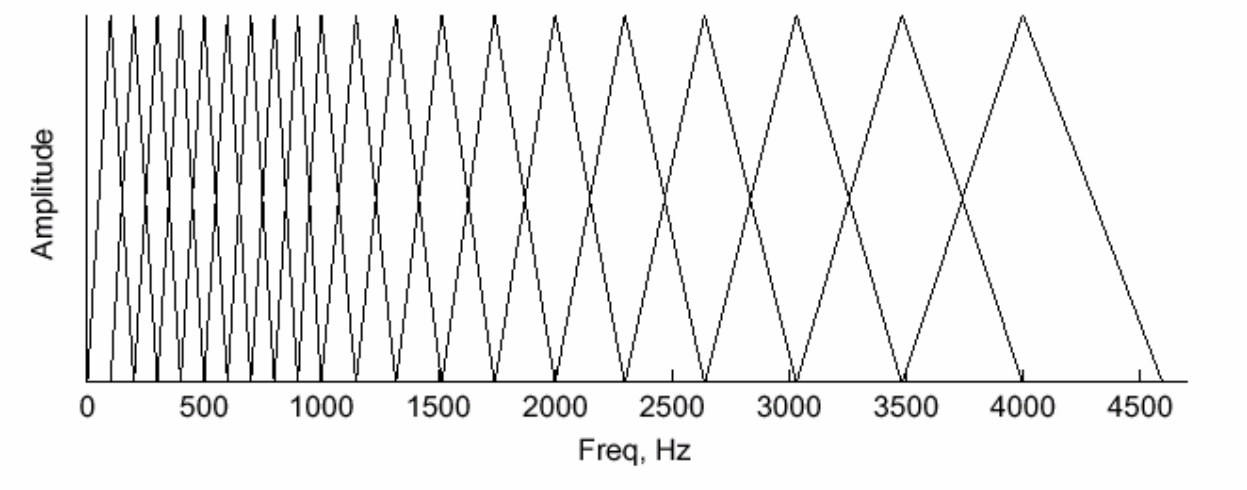



**Fig. 4.2. Filter Bank in Mel frequency scale**

### 4.1.5 Cepstrum

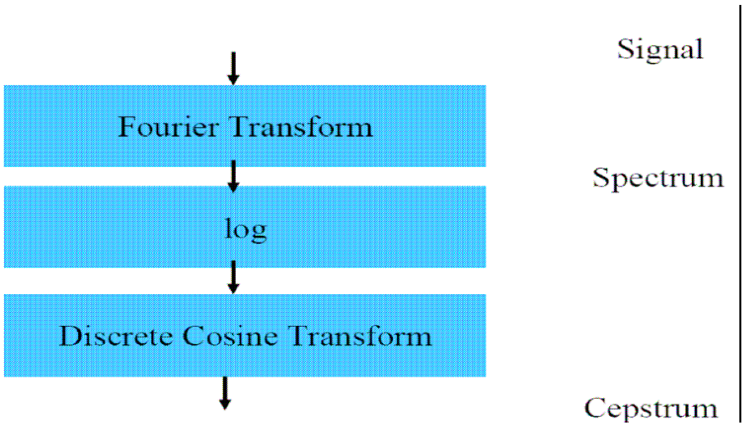
Cepstrum name was derived from the spectrum by reversing the first four letters of spectrum. We can say cepstrum is the Fourier Transformer of the log with unwrapped phase of the Fourier Transformer.

- Mathematically we can say Cepstrum of signal = FT(log(FT(the signal))+j2IIm)

Where m is the interger required to properly unwrap the angle or imaginary part of the complex log function.

- Algorithmically we can say – Signal - FT - log - phase unwrapping - FT - Cepstrum.

For defining the real values real cepstrum uses the logarithm function. While for defining the complex values whereas the complex cepstrum uses the complex logarithm function. The real cepstrum uses the information of the magnitude of the spectrum. where as complex cepstrum holds information about both magnitude and phase of the initial spectrum, which allows the reconstruction of the signal. We can calculate the cepstrum by many ways. Some of them need a phase-warping algorithm, others do not. Figure below shows the pipeline from signal to




Cepstrum. As we discussed in the Framing and Windowing section that speech signal is composed of quickly varying part e(n) excitation sequence convolved with slowly varying part (n) vocal system impulse response.

$$s(n) = e(n) * \theta(n)$$

Once we convolved the quickly varying part and slowly varying part it makes difficult to separate the two parts, cepstrum is introduced to separate this two parts. The equation for the cepstrum is given below:

$$c_s(n) = \mathfrak{F}^{-1}\left\{ log\left|\mathfrak{F}\{s(n)\}\right| \right\}$$

$\mathfrak{F}$ is the Discrete Time Fourier Transformer and $\mathfrak{F}^{-1}$ is the Inverse Discrete Time Fourier Transformer. By moving the signal from time domain to frequency domain convolution becomes the multiplication.

$$S(\omega) = E(\omega)\Theta(\omega)$$

The multiplication becomes the addition by taking the logarithm of the spectral magnitude

$$log|S(\omega)| = log|E(\omega)\Theta(\omega)| = log|E(\omega)| + log|\Theta(\omega)| = C_e(\omega) + C_\theta(\omega)$$

The Inverse Fourier Transform work individually on the two components as it is a linear

$$c_s(n) = \mathfrak{F}^{-1}\{C_e(\omega) + C_\theta(\omega)\} = \mathfrak{F}^{-1}\{C_e(\omega)\} + \mathfrak{F}^{-1}\{C_\theta(\omega)\} = c_e(n) + c_\theta(n)$$

The domain of the signal *cs*(*n*) is called the quefrency-domain.

### 4.1.6 Mel Frequency Cepstrum Coefficient

In this project we are using Mel Frequency Cepstral Coefficient. Mel frequency Cepstral Coefficients are coefficients that represent audio based on perception. This coefficient has a great success in speaker recognition application. It is derived from the Fourier Transform of the audio clip. In this technique the frequency bands are positioned logarithmically, whereas in the Fourier Transform the frequency bands are not positioned logarithmically. As the frequency bands are positioned logarithmically in MFCC, it approximates the human system response more closely than any other system. These coefficients allow better processing of data.

In the Mel Frequency Cepstral Coefficients the calculation of the Mel Cepstrum is same as the real Cepstrum except the Mel Cepstrum's frequency scale is warped to keep up a correspondence to the Mel scale.

The Mel scale was projected by Stevens, Volkmann and Newman in 1937. The Mel scale is mainly based on the study of observing the pitch or frequency perceived by the human. The scale is divided into the units mel. In this test the listener or test person started out hearing a frequency of 1000 Hz, and labelled it 1000 Mel for reference. Then the listeners were asked to change the frequency till it reaches to the frequency twice the reference frequency. Then this frequency labelled 2000 Mel. The same procedure repeated for the half the frequency, then this frequency labelled as 500 Mel, and so on. On this basis the normal frequency is mapped into the Mel frequency. The Mel scale is normally a linear mapping below 1000 Hz and logarithmically spaced above 1000 Hz. Figure below shows the example of normal frequency is mapped into the Mel frequency.

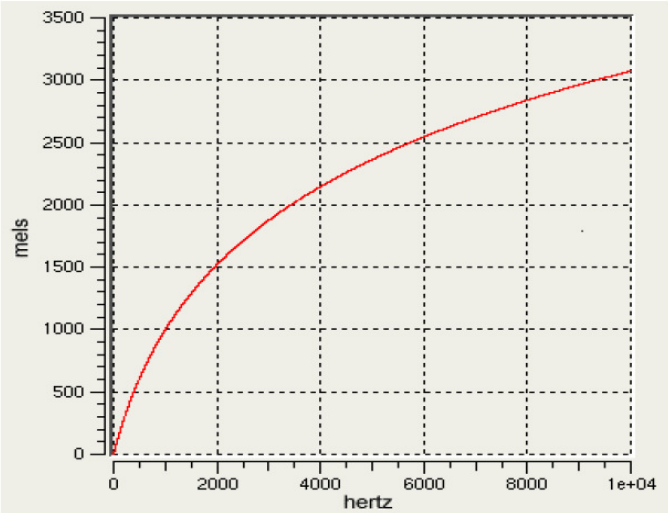



**Fig. 4.3.Mel frequency scale**

$$m = 1127.01048 \log_e(1 + f/700) \quad \text{.....(1)}$$

$$f = 700(e^{m/1127.01048} - 1) \quad \text{.....(2)}$$

The equation (1) above shows the mapping the normal frequency into the Mel frequency and equation (2) is the inverse, to get back the normal frequency.

# Chapter 5

# Algorithm

We have make ASR system using three methods namely:

(1) MFCC approach

(2) FFT approach

(3) Vector quantization

## 5.1. MFCC approach:

A block diagram of the structure of an MFCC processor is as shown in Fig 4.1.1. The speech input is typically recorded at a sampling rate above 10000 Hz. This sampling frequency was chosen to minimize the effects of *aliasing* in the analog-to-digital conversion. These sampled signals can capture all frequencies up to 5 kHz, which cover most energy of sounds that are generated by humans. The main purpose of the MFCC processor is to mimic the behavior of the human ears. In addition, rather than the speech waveforms themselves, MFCC‟s are shown to be less susceptible to mentioned variations.

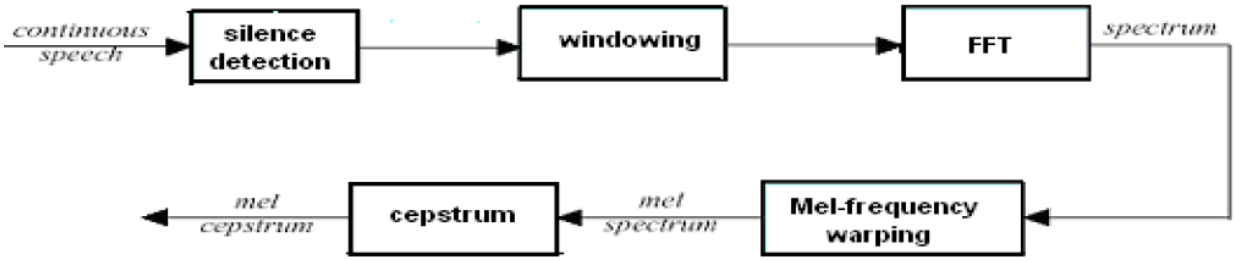



**Fig. 5.1.MFCC Approch**

We first stored the speech signal as a 10000 sample vector. It was observed from our experiment that the actual uttered speech eliminating the static portions came up to about 2500 samples, so, by using a simple threshold technique we carried out the silence detection to extract the actual uttered speech.

It is clear that what we wanted to achieve was a voice based biometric system capable of recognizing isolated words. As our experiments revealed almost all the isolated words were uttered within 2500 samples. But, when we passed this speech signal through a MFCC processor,

it spilt this up in the time domain by using overlapping windows each with about 250 samples. Thus when we convert this into the frequency domain we just have about 250 spectrum values under each window. This implied that converting it to the Mel scale would be redundant as the Mel scale is linear till 1000 Hz. So, we eliminated the block which did the Mel warping. We directly used the overlapping triangular windows in the frequency domain. We obtained the energy within each triangular window, followed by the DCT of their logarithms to achieve good compaction within a small number of coefficients as described by the MFCC approach.

This algorithm however, has a drawback. As explained earlier the key to this approach is using the energies within each triangular window, however, this may not be the best approach as was discovered. It was seen from the experiments that because of the prominence given to energy, this approach failed to recognize the same word uttered with different energy. Also, as this takes the summation of the energy within each triangular window it would essentially give the same value of energy irrespective of whether the spectrum peaks at one particular frequency and falls to lower values around it or whether it has an equal spread within the window. This is why we decided not to go ahead with the implementation of the MFCC approach.

The simulation was carried out in MATLAB. The various stages of the simulation have been represented in the form of the plots shown. The input continuous speech signal considered as an example for this project is the word “HELLO”.

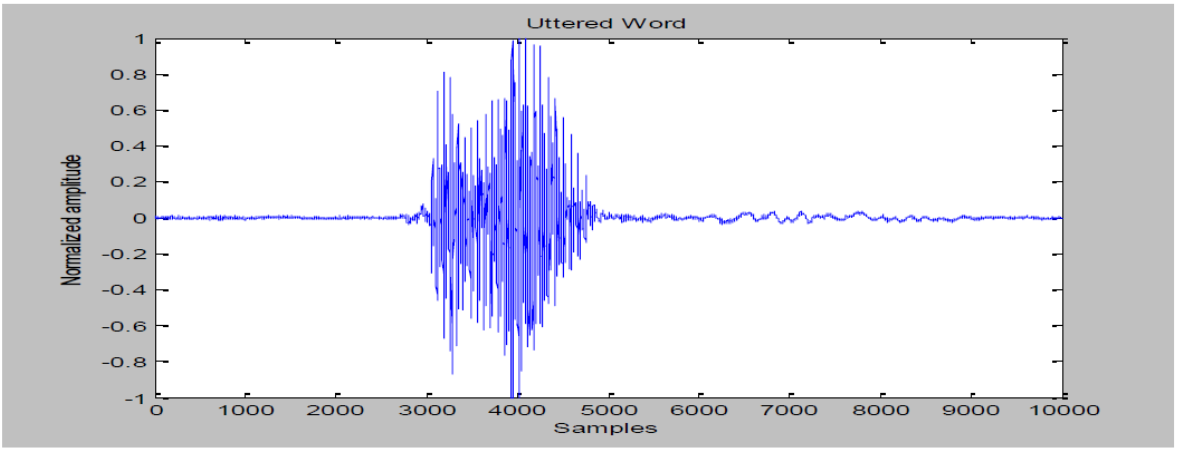



**Fig. 5.2. The word “Hello” taken for analysis**

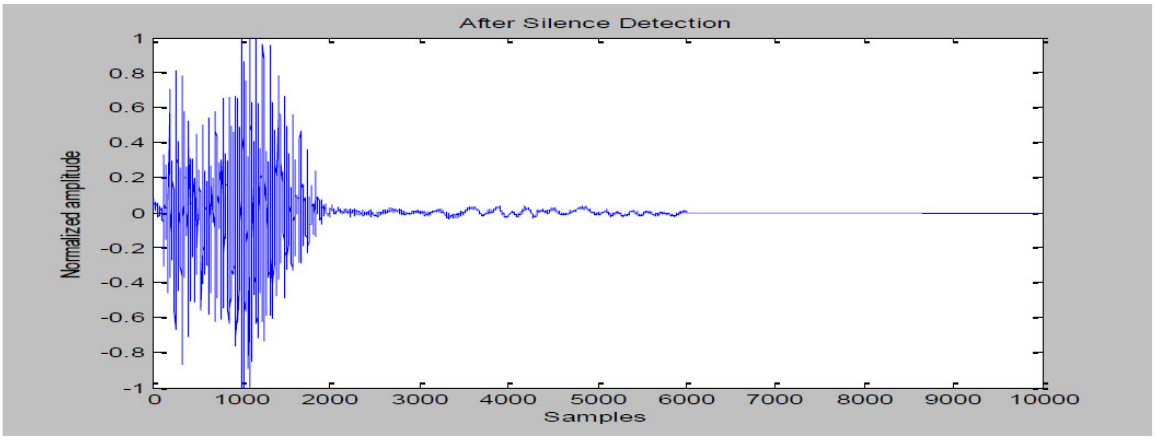



**Fig. 5.3. The word "Hello" after silence detection**

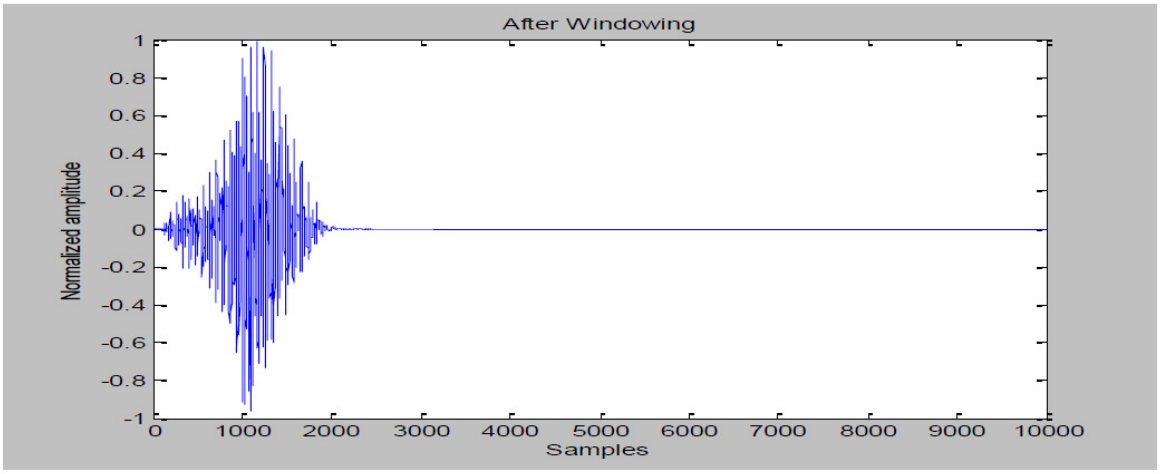



**Fig. 5.4. The word "Hello" after windowing using Hamming window**

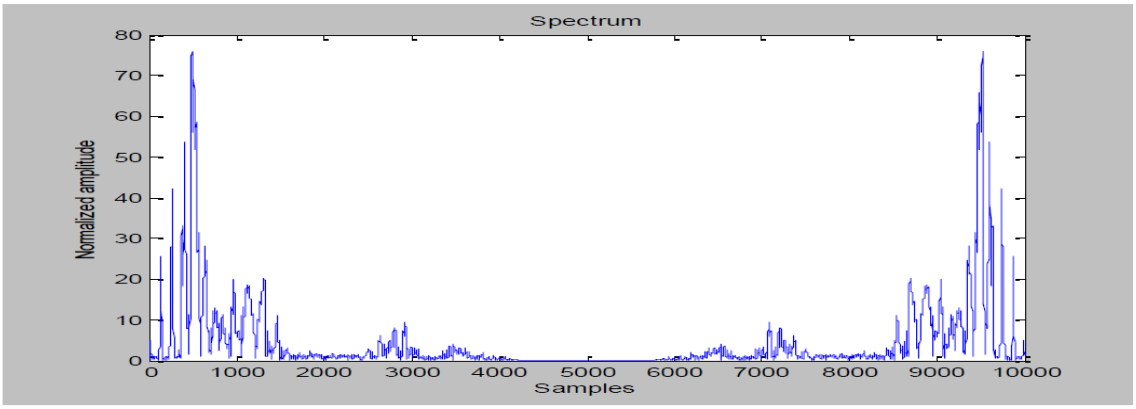



**Fig. 5.5. The word "Hello" after FFT**

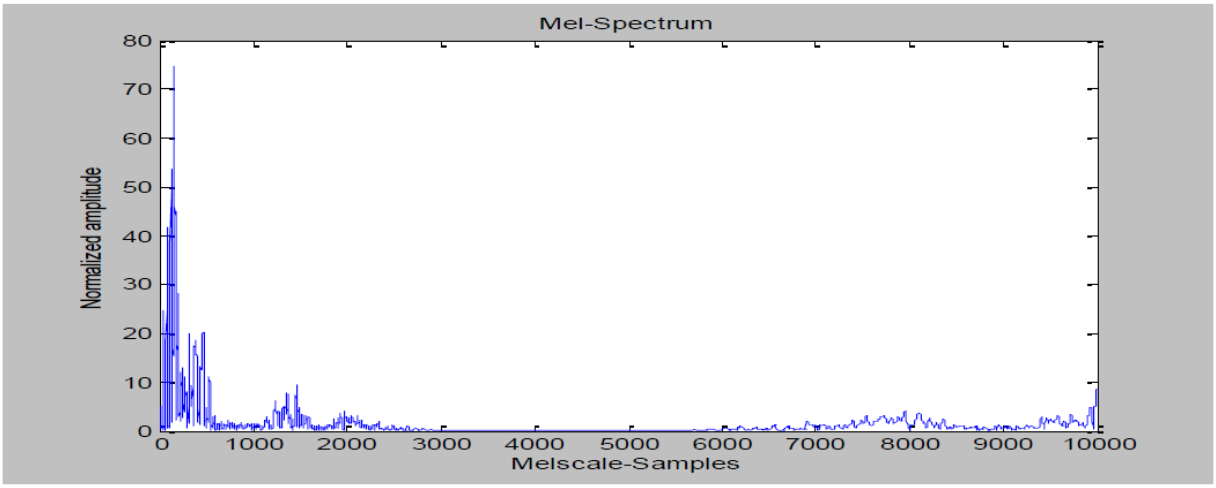
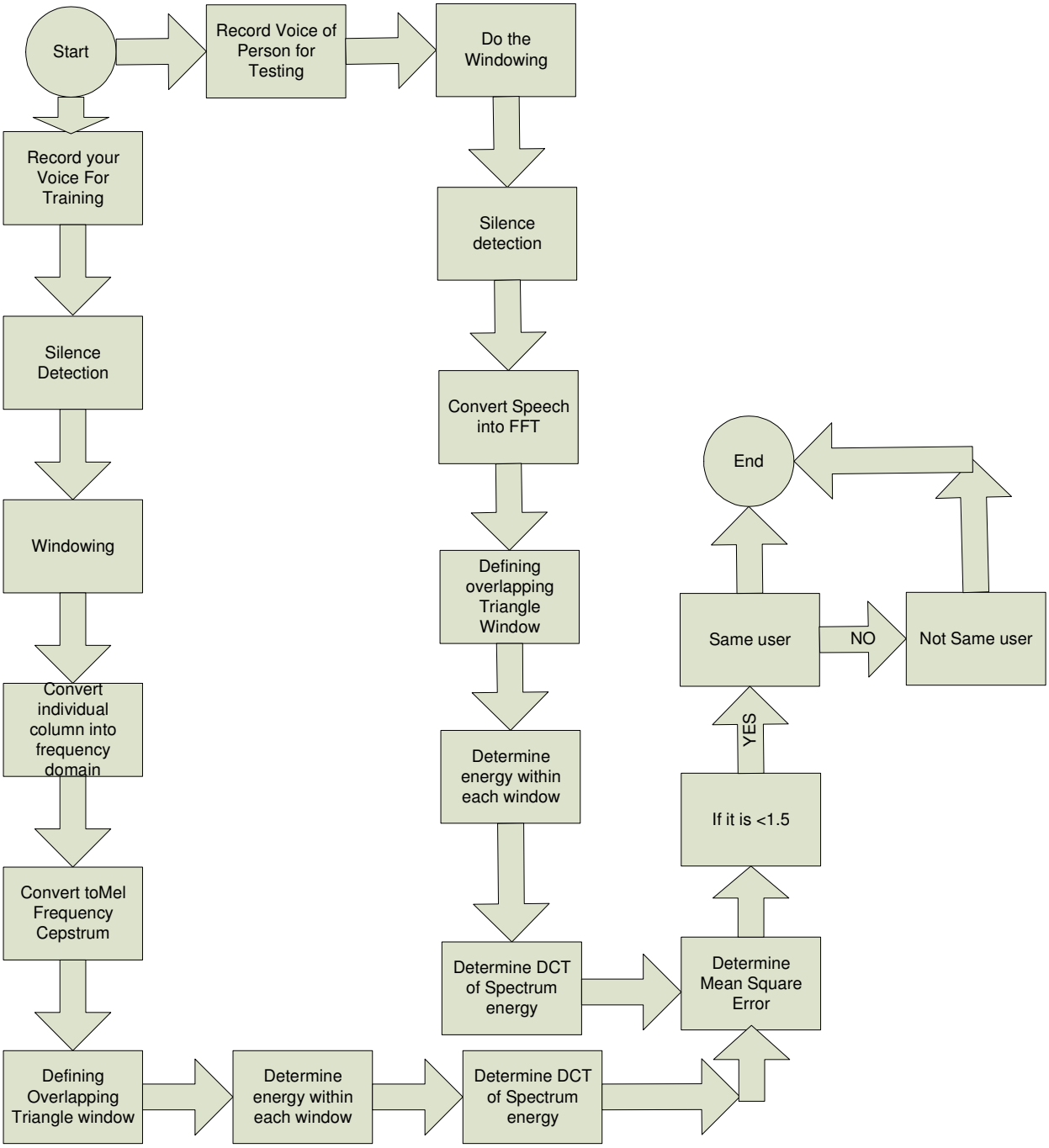
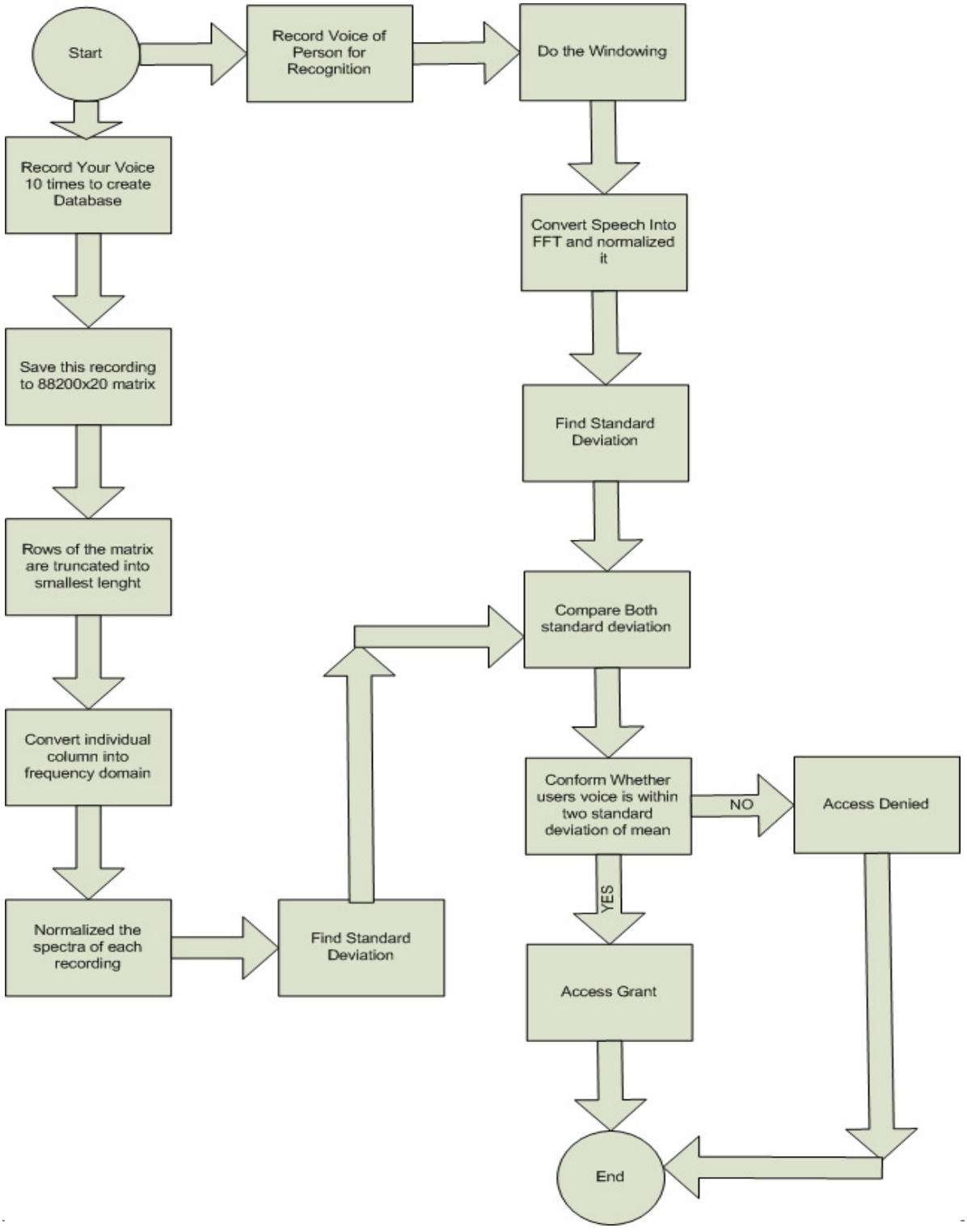



**Fig. 5.6. The word "Hello" after Mel-warping**

## 5.1. MFCC approach Algorithm:



## 5.2. FFT approach:



## 5.3. Using VQ:

A speaker recognition system must able to estimate probability distributions of the computed feature vectors. Storing every single vector that generate from the training mode is impossible, since these distributions are defined over a high-dimensional space. It is often easier to start by quantizing each feature vector to one of a relatively small number of template vectors, with a process called vector quantization. VQ is a process of taking a large set of feature vectors and producing a smaller set of measure vectors that represents the centroids of the distribution.

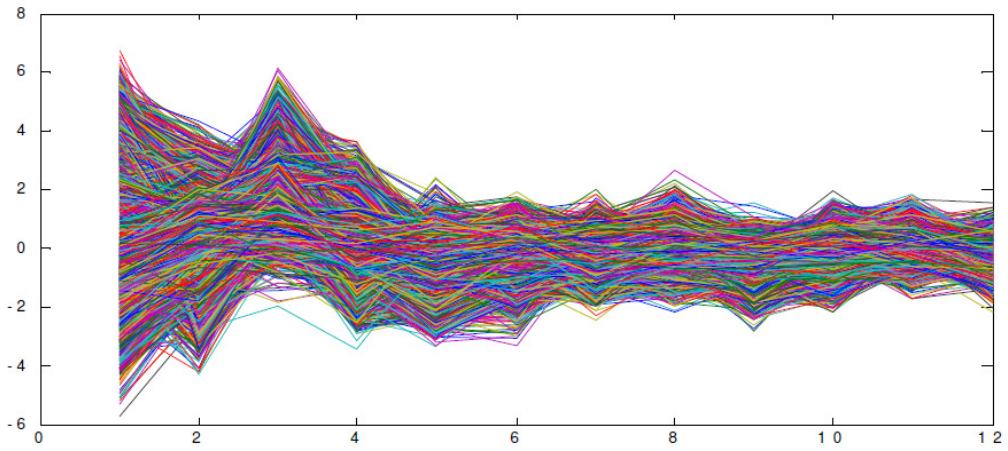



**Fig. 5.7. the vectors generated from training before VQ**

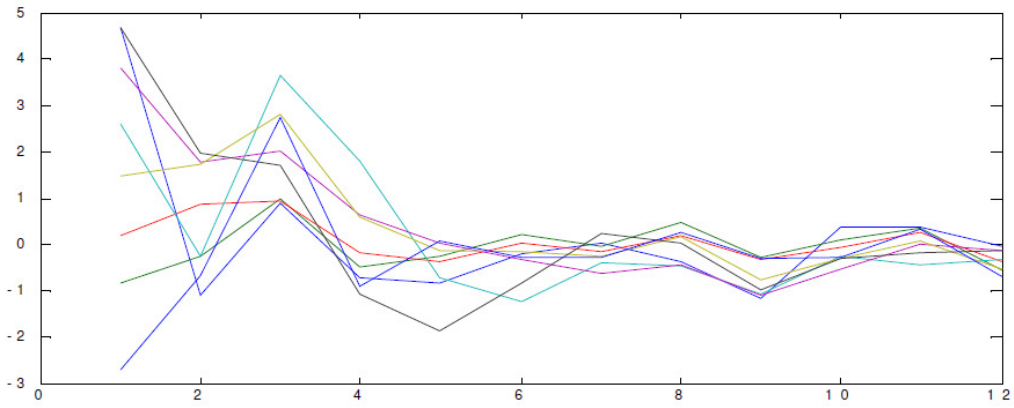



**Fig. 5.8. the representative feature vectors resulted after VQ**

The technique of VQ consists of extracting a small number of representative feature vectors as an efficient means of characterizing the speaker specific features. By means of VQ, storing every single vector that we generate from the training is impossible.

By using these training data features are clustered to form a codebook for each speaker. In the recognition stage, the data from the tested speaker is compared to the codebook of each speaker and measure the difference. These differences are then use to make the recognition decision.

The problem of speaker recognition belongs to a much broader topic in scientific and engineering so called *pattern recognition*. The goal of pattern recognition is to classify objects of interest into one of a number of categories or classes. The objects of interest are generically called *patterns* and in our case are sequences of acoustic vectors that are extracted from an input speech using the techniques described in the previous section. The classes here refer to individual speakers. Since the classification procedure in our case is applied on extracted features, it can be also referred to as *feature matching*.

Furthermore, if there exists some set of patterns that the individual classes of which are already known, then one has a problem in *supervised pattern recognition*. This is exactly our case since during the training session, we label each input speech with the ID of the speaker (S1 to S8). These patterns comprise the *training set* and are used to derive a classification algorithm. The remaining patterns are then used to test the classification algorithm; these patterns are collectively referred to as the *test set*. If the correct classes of the individual patterns in the test set are also known, then one can evaluate the performance of the algorithm.

The state-of-the-art in feature matching techniques used in speaker recognition include Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ). In this project, the VQ approach will be used, due to ease of implementation and high accuracy. VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a *cluster* and can be represented by its center called a *codeword*. The collection of all codewords is called a *codebook*.

Figure 5.7 shows a conceptual diagram to illustrate this recognition process. In the figure, only two speakers and two dimensions of the acoustic space are shown. The circles refer to the acoustic vectors from the speaker 1 while the triangles are from the speaker 2. In the training

phase, a speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors. The result codewords (centroids) are shown in Figure 5 by black circles and black triangles forspeaker 1 and 2, respectively. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of an unknown voice is "vector-quantized" using each trained codebook and the *total VQ distortion* is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified.

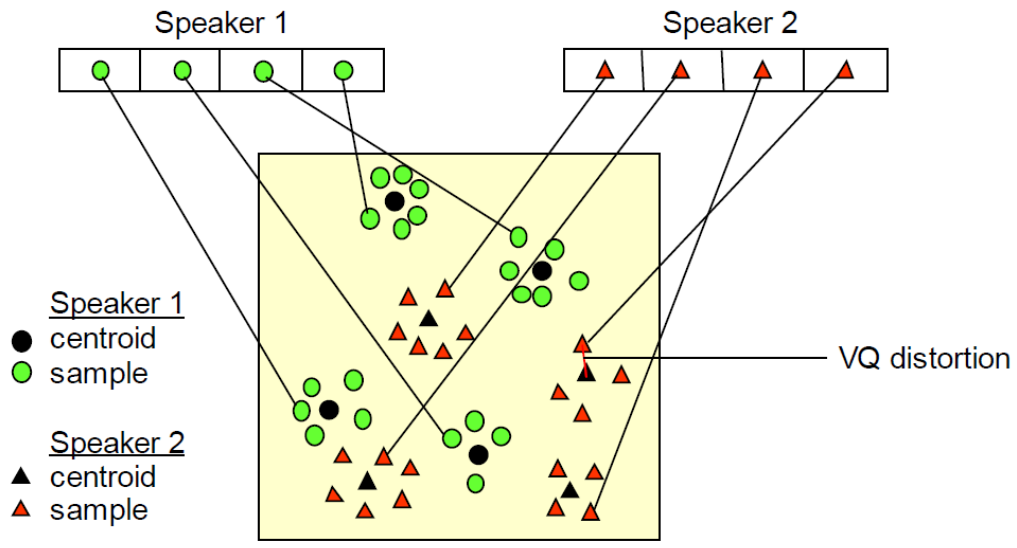



**Fig. 5.9. Conceptual diagram illustrating vector quantization codebook formation. One speaker can be discriminated from another based of the location of centroids.**

### 5.3.1 Clustering the training vector

After the enrolment session, the acoustic vectors extracted from input speech of a speaker provide a set of training vectors. As described above, the next important step is to build a speaker-specific VQ codebook for this speaker using those training vectors. There is a well-know algorithm, namely LBG algorithm [Linde, Buzo and Gray, 1980], for clustering a set of $L$ training vectors into a set of $M$ codebook vectors. The algorithm is formally implemented by the following recursive procedure:

1. Design a 1-vector codebook; this is the centroid of the entire set of training

vectors (hence, no iteration is required here).

2. Double the size of the codebook by splitting each current codebook **y***n* according

$$y_n^+ = y_n(1+\varepsilon)$$
$$y_n^- = y_n(1-\varepsilon)$$

to the rule where *n* varies from 1 to the current size of the codebook, and e is a splittingparameter (we choose e =0.01).

3. Nearest-Neighbor Search: for each training vector, find the codeword in the Current codebook that is closest (in terms of similarity measurement), and assign that vector to the corresponding cell (associated with the closest codeword).

4. Centroid Update: update the codeword in each cell using the centroid of the training vectors assigned to that cell.

5. Iteration 1: repeat steps 3 and 4 until the average distance falls below a preset Threshold

6. Iteration 2: repeat steps 2, 3 and 4 until a codebook size of *M* is designed

Intuitively, the LBG algorithm designs an *M*-vector codebook in stages. It starts first by designing a 1-vector codebook, then uses a splitting technique on the codewords to initialize the search for a 2-vector codebook, and continues the splitting process until the desired *M*-vector codebook is obtained. Figure 5.10 shows, in a flow diagram, the detailed steps of the LBG algorithm. "*Cluster vectors*" is the nearest-neighbor search procedure which assigns each training vector to a cluster associated with the closest codeword. "*Find centroids*" is the centroid update procedure. "*Compute D (distortion)*" sums the distances of all training vectors in the nearest-neighbor search so as to determine whether the procedure has converged.

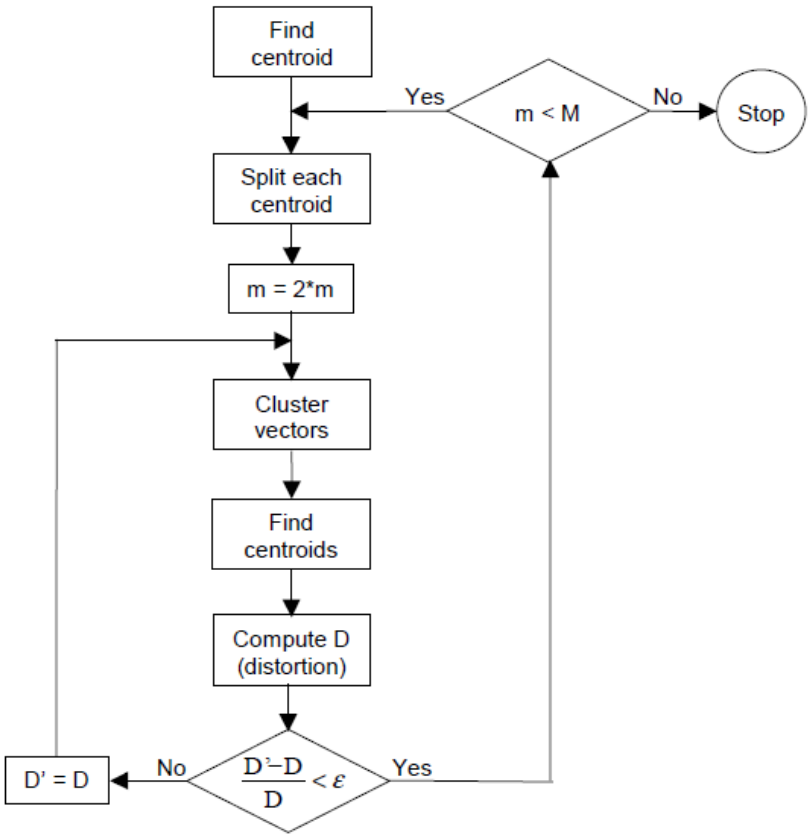



**Fig. 5.10. Flow diagram of the LBG algorithm (Adapted from Rabiner and Juang, 1993)**

# Chapter 6

# Sample training and Recognition Session with Screenshot

## 6.1. Main Menu

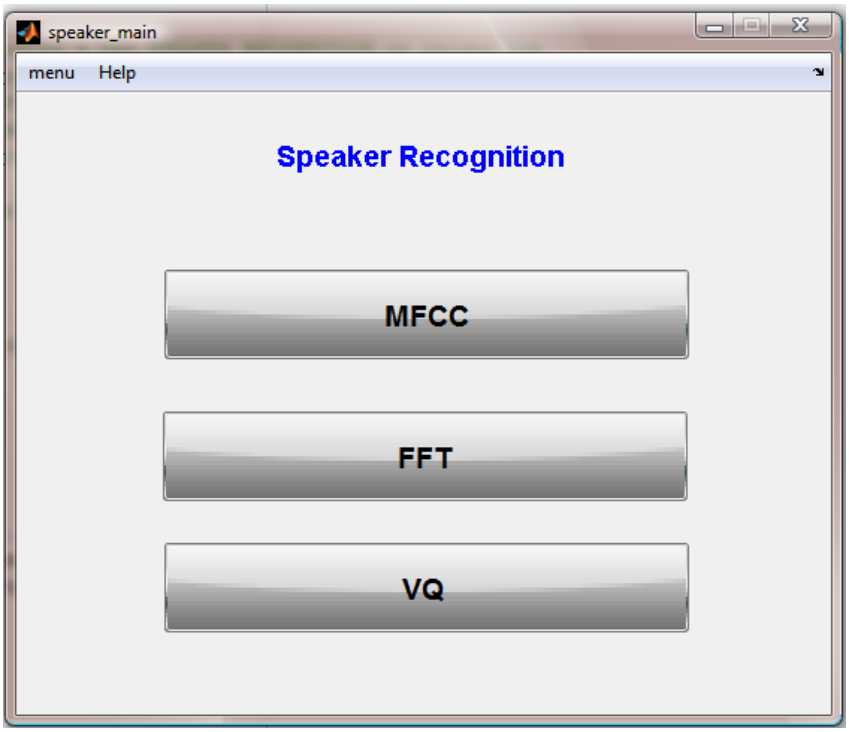



**Fig. 6.1. Main Menu of the Speech Recognition Application**

This is the main menu of our code form this menu you can select any method from which you want to recognize the person. So, we go one by one method first we look about mfcc method.

## 6.2 GUI of MFCC method :

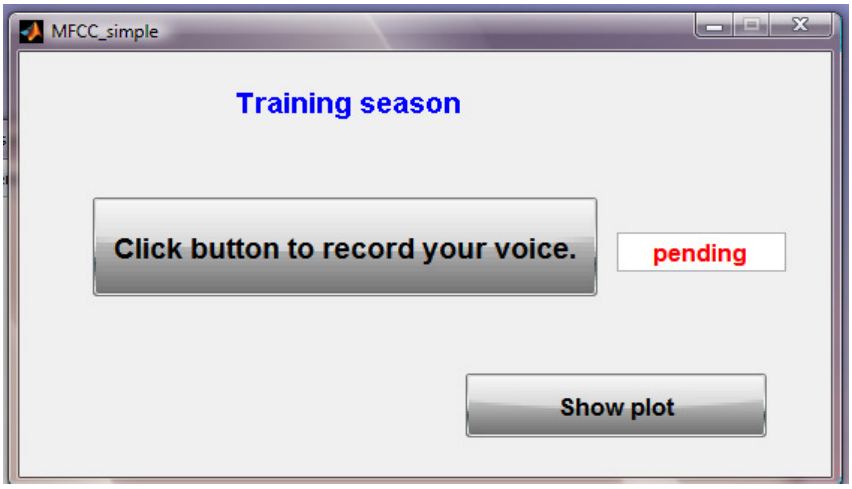



**Fig. 6.2. Training Menu in MFCC approch**

Here you just click the button and record your voice and also see your speech plot by clicking show plot button.

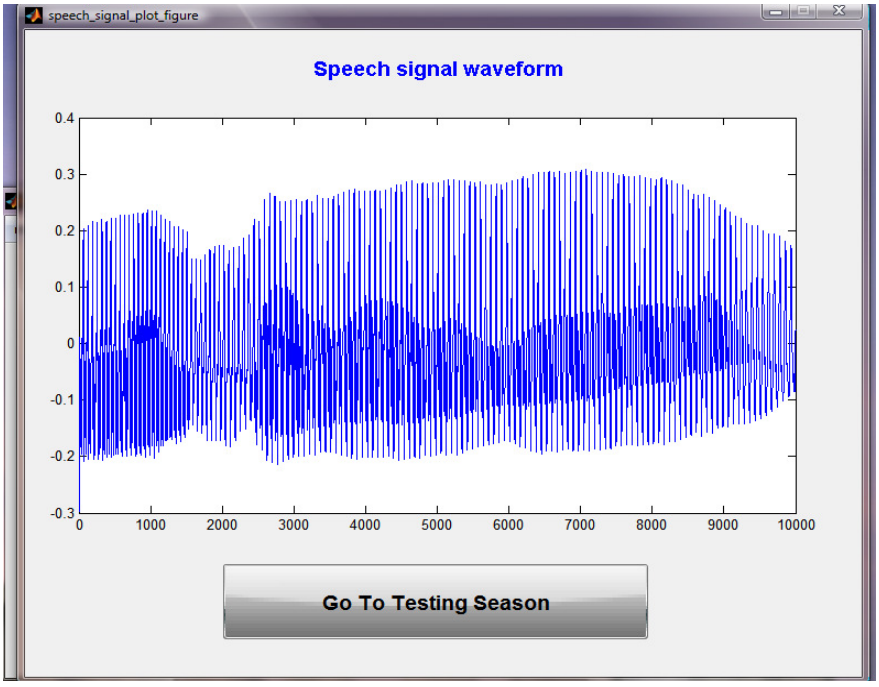



**Fig. 6.3. Waveform of Training Session**

After clicking the go to testing session you can now check whether the speaker is identified or not.

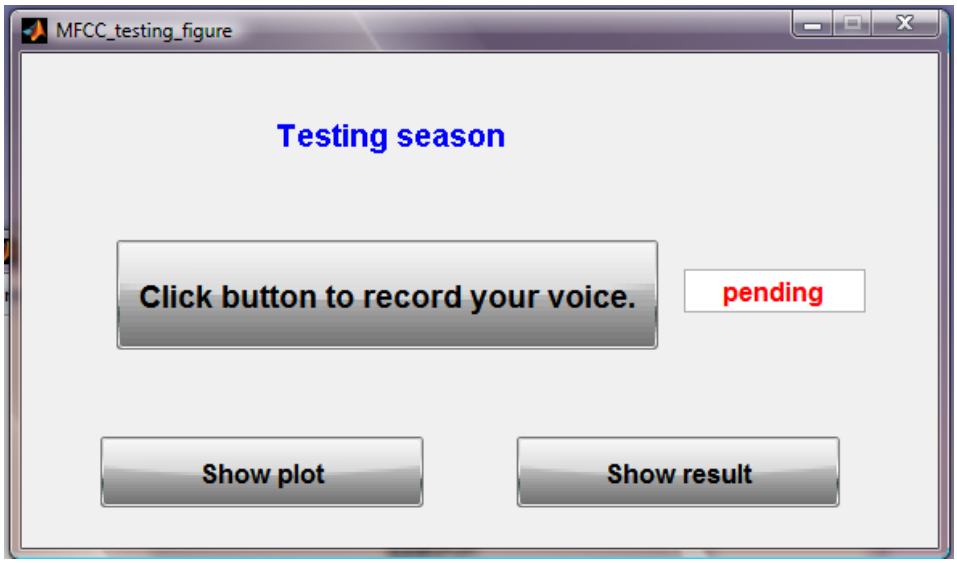



**Fig. 6.4. Testing Session GUI**

If the speaker is identified then,

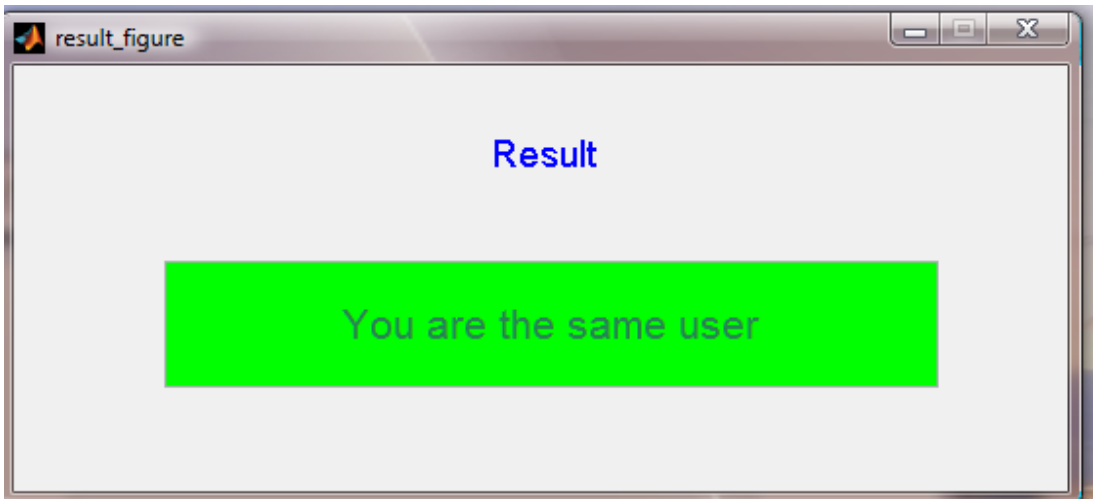



**Fig. 6.5. Final Result(1)**

If  it is not identified then,

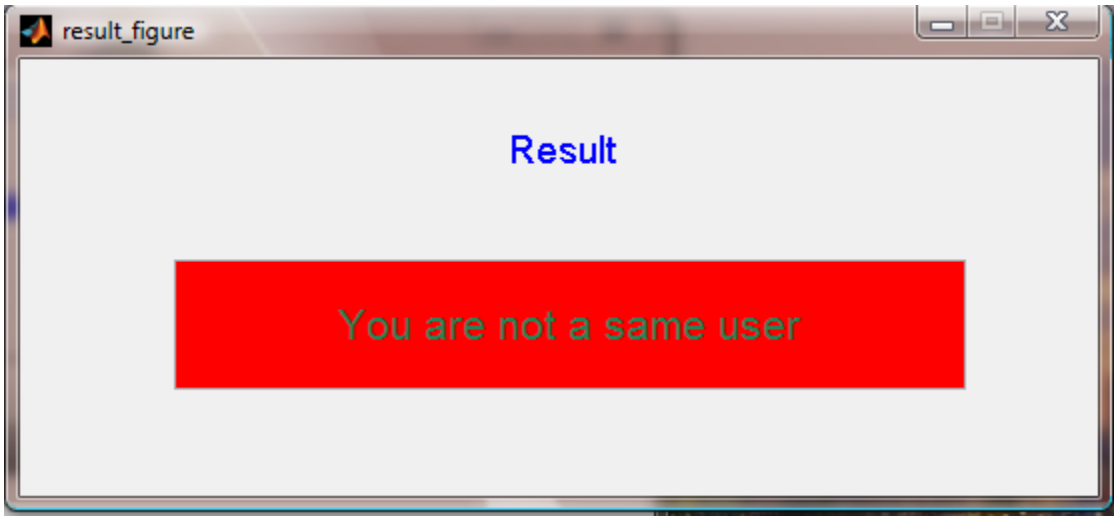



**Fig. 6.6. Final Result(2)**

## 6.3 GUI of FFT method :

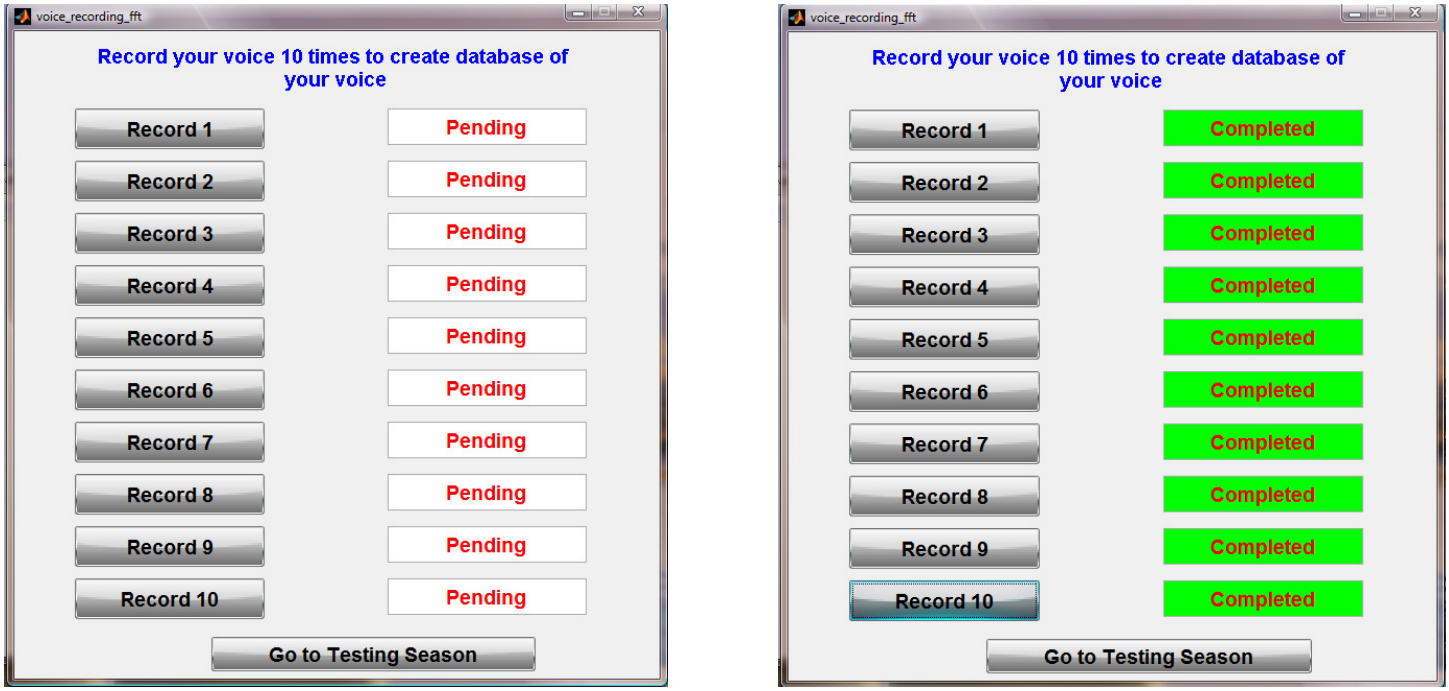



**Fig. 6.7. Create Database GUI**

Here First of all you have to create the database of the user for that we are going to record user voice 10 times. Then we going for the testing phase.

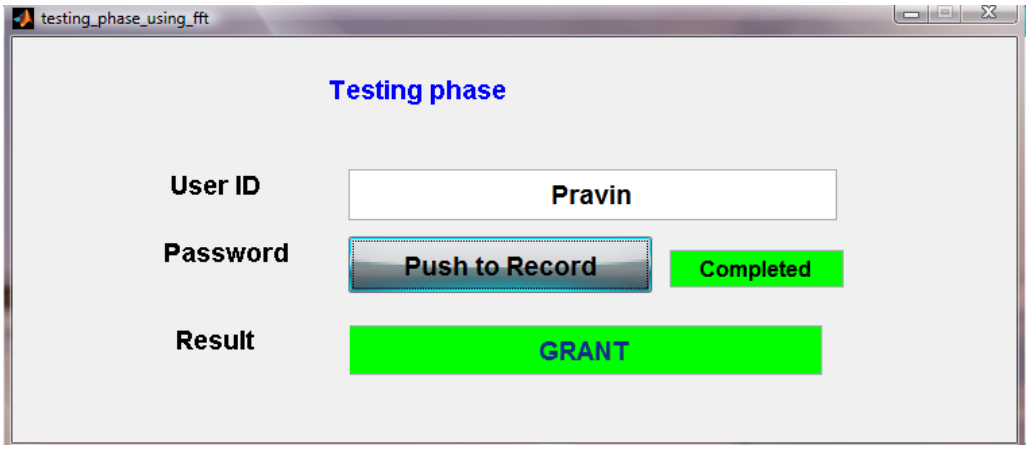



**Fig. 6.8. User authentication GUI**

## 6.3 GUI of VQ method :

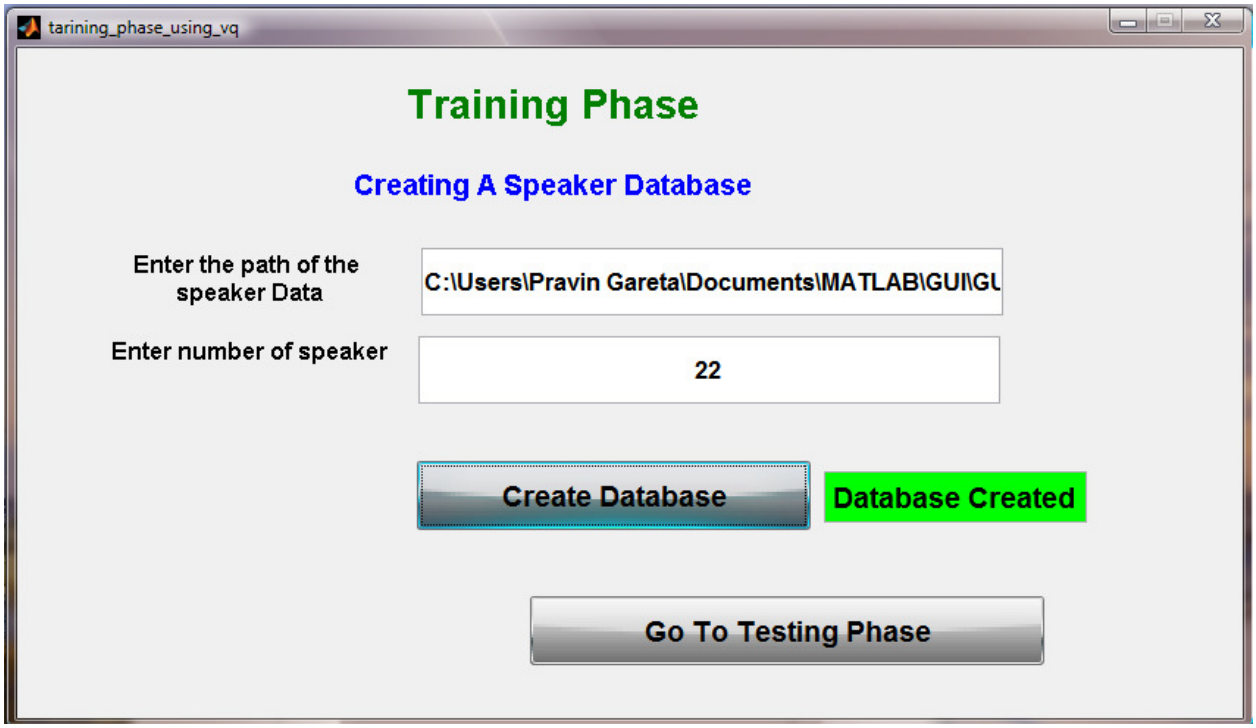



**Fig. 6.9. GUI of Database creation**

Here first we have to decide number of speaker and their training paths in the computer we have to enter this paths then codebook will be created now after we have to go for the testing phase in the testing phase again we have to store the voice of user for authentication in the computer and give the path and VQ will matches the speaker to speaker from training to the testing phase.

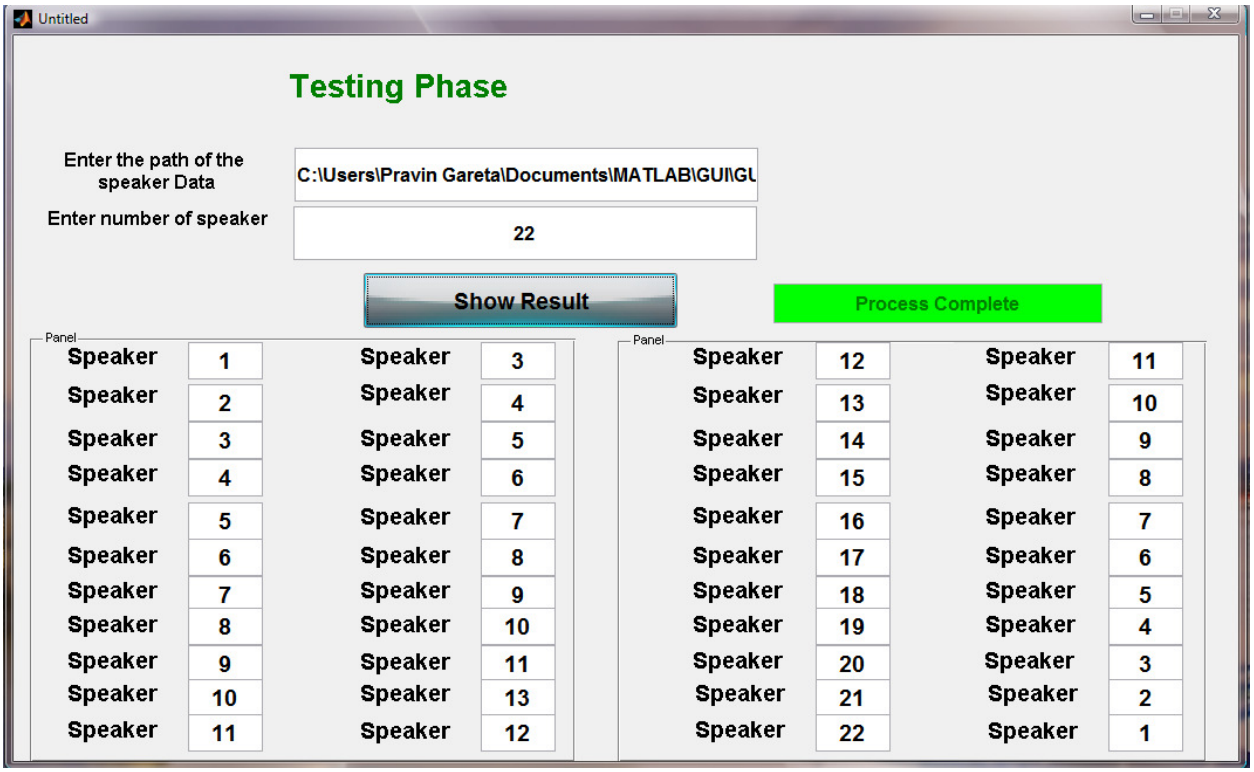



**Fig. 6.10. GUI of User matching**

## Conclusion

The goal of this project was to create a speaker recognition system, and apply it to a speech of an unknown speaker. By investigating the extracted features of the unknown speech and then compare them to the stored extracted features for each different speaker in order to identify the unknown speaker.

The feature extraction is done by using MFCC (Mel Frequency Cepstral Coefficients). The speaker was modeled using Vector Quantization (VQ). A VQ codebook is generated by clustering the training feature vectors of each speaker and then stored in the speaker database. In this method, the LBG algorithm is used to do the clustering. In the recognition stage, a distortion measure which based on the minimizing the Euclidean distance was used when matching an unknown speaker with the speaker database.

During this project, we have found out that the VQ based clustering approach
provides us with the faster speaker identification process than only mfcc approach or FFT approach.

## Applications

After nearly sixty years of research, speech recognition technology has reached a relatively high level. However, most state-of-the-art ASR systems run on desktop with powerful microprocessors, ample memory and an ever-present power supply. In these years, with the rapid evolvement of hardware and software technologies, ASR has become more and more expedient as an alternative human-to-machine interface that is needed for the following application areas:

- Stand-alone consumer devices such as wrist watch, toys and hands-free mobile phone in car where people are unable to use other interfaces or big input platforms like keyboards are not available.

- Single purpose command and control system such as voice dialing for cellular, home, and office phones where multi-function computers (PCs) are redundant.

Some of the applications of speaker verification systems are:

- Time and Attendance Systems
- Access Control Systems
- Telephone-Banking/Broking
- Biometric Login to telephone aided shopping systems
- Information and Reservation Services
- Security control for confidential information
- Forensic purposes

Voice based Telephone dialing is one of the applications we simulated. The key focus of this application is to aid the physically challenged in executing a mundane task like telephone dialing. Here the user initially trains the system by uttering the digits from 0 to 9. Once the system has been trained, the system can recognize the digits uttered by the user who trained the system. This system can also add some inherent security as the system based on cepstral approach is speaker dependent. The algorithm is run on a particular speaker and the MFCC coefficients determined. Now the algorithm is applied to a different speaker and the mismatch was clearly observed. Thus the inherent security provided by the system was confirmed.

Presently systems have also been designed which incorporate Speech and Speaker Recognition. Typically a user has two levels of check. She/he has to initially speak the right password to gain access to a system. The system not only verifies if the correct password has been said but also focused on the authenticity of the speaker. The ultimate goal is do have a system which does a Speech, Iris, Fingerprint Recognition to implement access control.

## Scope for future work

This project focused on “Isolated Word Recognition”. But we feel the idea can be extended to “Continuous Word Recognition” and ultimately create a Language Independent Recognition System based on algorithms which make these systems robust. The use of Statistical Models like HMMs, GMMs or learning models like Neural Networks and other associated aspects of Artificial Intelligence can also be incorporated in this direction to improve upon the present project. This would make the system much tolerant to variations like accent and extraneous conditions like noise and associated residues and hence make it less error prone. Some other aspects which can be looked into are:

- The detection used in this work is only based on the frame energy in MFCC which is not good for a noisy environment with low SNR. The error rate of determining the beginning and ending of speech segments will greatly increase which directly influence the recognition performance at the pattern recognition part. So, we should try to use some effective way to do detection. One of these methods could be to use the statistical way to find a distribution which can separate the noise and speech from each other.

- The size of the training data i.e. the code book can be increased in VQ as it is clearly proven that the greater the size of the training data, the greater the recognition accuracy. This training data could incorporate aspects like the different ways via the accents in which a word can be spoken, the same words spoken by male/female speakers and the word being spoken under different conditions say under conditions in which the speaker may have a sore throat etc.

- Our VQ code takes a very long time for the recognition of the actual voice averagely half and hour we have found this can be decreased using some other algorithm. Here we used LBG algorithm. But someone can try k-means algorithm also. This field is very vast and research is also done for that purpose.

## References


[1] Lawrence Rabiner, Biing-Hwang Juang – „*Fundamentals of Speech Recognition'*

[2] Wei Han, Cheong-Fat Chan, Chiu-Sing Choy and Kong-Pang Pun – „*An Efficient MFCC Extraction Method in Speech Recognition'*, Department of Electronic Engineering, The Chinese University of Hong Kong, Hong, IEEE – ISCAS, 2006

[3] Leigh D. Alsteris and Kuldip K. Paliwal – „*ASR on Speech Reconstructed from Short- time Fourier Phase Spectra',* School of Microelectronic Engineering Griffth University, Brisbane, Australia, ICLSP - 2004

[4] Waleed H. Abdulla – „*Auditory Based Feature Vectors for Speech Recognition Systems'*, Electrical & Electronic Engineering Department, The University of Auckland

[5] Pradeep Kumar P and Preeti Rao – „*A Study of Frequency-Scale Warping for Speaker Recognition'*, Dept of Electrical Engineering, IIT- Bombay, National Conference on Communications, NCC 2004, IISc Bangalore, Jan 30 -Feb 1, 2004

[6] Beth Logan – „*Mel Frequency Cepstral Coefficients for Music Modeling'*, Cambridge Research Laboratory, Compaq Computer Corporation

[7] Keller, E.: "Fundamentals of Speech Synthesis and Speech Recognition", John Wiley & Sons, New York, USA, (1994).

[8] Markowitz, J.A.: "Using Speech Recognition", Prentice Hall, (1996).

[9] Yılmaz, C.: "A Large Vocabulary Speech Recognition System for Turkish", MS Thesis, Bilkent University, Institute of Engineering and Science, Ankara, Turkey, (1999).

[10] Mengüsoglu, E.: "Rule Based Design and Implementation of a Speech Recognition System for Turkish Language", MS Thesis, Hacettepe University, Inst. for Graduate Studies in Pure and Applied Sciences, Ankara, Turkey, (1999).

[11] Zegers, P.: "Speech Recognition Using Neural Networks", MS Thesis, University of Arizona, Department of Electrical Engineering in the Graduate College, Arizona, USA, (1998).

[12] Woszczyna, M.: "JANUS 93: Towards Spontaneous Speech Translation", IEEE

Proceedings Conference on Neural Networks, (1994).
[13] Somervuo, P.: “Speech Recognition using context vectors and multiple feature streams”, MS Thesis, (1996).
[14] Nilsson, M.; Ejnarsson, M.: “Speech Recognition Using HMM: Performance Evaluation in Noisy Environments”, MS Thesis, Blekinge Institute of Technology, Department of Telecommunications and Signal Processing, (2002).
[15] Hakkani-Tur, D.; Oflazer, K.; Tur, G.:. “Statistical Morphological Disambiguation for Agglutinative Languages”, Technical Report, Bilkent University, (2000).
[16] Ursin, M.: “Triphone Clustering in Continuous Speech Recognition”, MS Thesis, Helsinki University of Technology, Department of Computer Science, (2002).
[17] www.dspguide.com/zipped.htm: “The Scientist and Engineer's Guide to Digital Signal Processing” (Access date: March 2005).
[18] Brookes, M.: “VOICEBOX: a MATLAB toolbox for speech processing”, www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, (2003).
[19] Davis, S.; Mermelstein, P.: “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 4 (1980).
[20] Skowronski, M.D.: “Biologically Inspired Noise-Robust Speech Recognition for Both Man and Machine”, PhD Thesis, The Graduate School of the University of Florida, (2004).
[21] MATLAB Product Help: “MATLAB Compiler: Introducing the MATLAB Compiler: Why Compile M-Files?”, (2001).