A

DISSERTATION REPORT

On

Speaker Identification through natural and whisper speech signal

By

AMRITA SINGH
2015PEB5115

Under the supervision of

Dr. AMIT JOSHI

Assistant Professor, ECE Department
MNIT, Jaipur

Submitted in partial fulfillment of the requirements of the degree of
MASTER OF TECHNOLOGY
to the

DEPARTMENT OF ELECTRONICS AND COMMUNICATION
ENGINEERING

MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY, JAIPUR

JULY- 2017

Department of Electronics and Communication Engineering

Malaviya National Institute of Technology, Jaipur

Certificate

This is to certify that this Dissertation report entitled "**Speaker Identification through natural and whisper speech signal** " by **Amrita Singh** (**2015PEB5115**), is the work completed under my supervision and guidance, hence approved for submission in partial fulfillment for the award of degree of **Master Of Technology** in **EMBEDDED SYSTEMS** to the Department of Electronics and Communication Engineering, Malaviya National Institute of Technology, Jaipur in the academic session 2016-2017 for full time post graduation program of 2015-2017. The matter embodied in the thesis has not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Amit Joshi

Dept. of Electronics & Comm. Engg.

National Institute of Technology
Jaipur, India-302017

Declaration

I, hereby declare that the work which is being presented in this project entitled "Speaker identification through natural and whisper speech signal " in partial fulfillment of degree of Master of Technology in Embedded Systems is an authentic record of my work carried out under the supervision and guidance of Dr.Amit Mahesh Joshi in Department of Electronics and Communication, Malaviya National Institute of Technology, Jaipur. I am fully responsible for the matter embodied in this project in case of any discrepancy found in the project, and the project has not been submitted for the award of any other degree. I also confirm that I have consulted the published work of others, the source is attributed and I have acknowledged all main sources of help.

(AMRITA SINGH)

Acknowledgement

I am grateful to my supervisor Dr. Amit Mahesh Joshi for his constant guidance and encouragement and support to carry out this work. His excellent cooperation and suggestion provided me with an impetus to work and made the completion of work possible. He has been a great source of inspiration to me, all through. I am very grateful to him for guiding me how to conduct research and how to clearly & effectively present the work done.

I would like to express my deepest sense of gratitude and humble regards to our Head of Department Prof. K. K. Sharma for giving encouragement in my endeavors and providing necessary facility in the Department. I am very thankful to all faculty member**s** of ECE, MNIT for their valuable assistance and advice.

I would also like to thank my friends for their support in discussions which proved valuable for me. I am indebted to my parents and family for their constant support, love, encouragement and sacrifices made by them so that I could grow up in a learning environment.

Finally, I express my sincere thanks to all those who helped me directly or indirectly to successfully complete this work.

(AMRITA SINGH)

Abstract

The main aim of this project is to develop speech identification through natural and whisper speech system using Gaussian Mixture Model (GMM) with a good accuracy at all the possible frequency range of human voice. This thesis deals with study of automatically identifying whispering speakers because people generally lower their voice in whisper mode to being identifying. Recently research in the area of speech has been devoted to soundness due to microphone and channel effects. However changes in whispering speech due to vocal effort is notable challenges in maintaining system accuracy.

When we speak, we utter different words .It is possible because we can change the resonant modes of the vocal cavity and periodic excitation in vowels so the spectral structure in neutral and whisper speech is different. Therefore accuracy of speaker's speech trained mainly with Mel Frequency Cepstral Coefficient (MFCC) algorithm and Exponential Frequency Cepstral Coefficient (EFCC) algorithm.

Second part describes the testing Model used for modeling and recognizing the speaker. In testing model we are using K-means algorithm and Gaussian Mixture Model (GMM). All the training samples are clustered using K-means algorithm and Gaussian mixture Model. GMM containing mean, variance and weight are modeling parameters. Here Expectation Maximization algorithm is used for testing the samples and re-estimate the parameters. Finally GMM algorithm recognizes the speaker that exactly matches for given database. Here all the results are done by the MATLAB tool and Microsoft window 7 operating system.

Contents

List of figures

Chapter 1

Introduction

1.1 Introduction

Speaker Recognition sometimes referred to as speaker biometrics system which also includes verification, classification in this system any procedure is used knowledge of the identity of a speaker based on his/her voice. Speaker identification is divided into two parts closed -set speaker identification and open-set speaker identification system, In closed-set speaker identification, the speech of claimed speaker is compared with all the available speaker in the database and closest matched speaker ID is given as output. Closed-set identification there is no rejection even if we are comparing a 5-year old child with all stored adult males database it gives a result with closest matched speaker. So closed-set speaker identification is not much practical in use, but it has own applications.

Open-set speaker identification system is a combination of closed -set speaker identification and speaker verification. In this, the output of closed-set speaker identification is used in speaker verification system. If the claimed speaker is matched with a database of the speaker then it returns the ID of matched speaker which is verified with true speaker and if verification fails then claimed speaker would be rejected, and there is no valid identification. Open-set speaker identification is a much complex than closed-set speaker identification.

Here we are using closet-set speaker identification system because closed-set speaker identification is initial step of speaker recognition which is used in speaker verification as first step to improve the security. Adaptation and identification of speaker have more application than speaker verification because speaker verification is limited only security. In speaker verification we can be used as adaptive user interference like an example would be a software program which would identify the audio of a speaker so that the interaction environment may be customized for that individual. In this case, there is no great loss by making a mistake. In fact, some match needs to be returned just to be able to pick a customization profile. If the speaker does not exist in the database, then there is generally no difference in what profile is used, unless profiles hold personal information in which case rejection or diversion to a different profile will become necessary.

1.2 Motivation

Speech recognition (speaker identification) has many numbers of real world applications like Financial Applications, Forensic and Legal Applications, Access Control (Security) Application, Teleconferencing Applications. The Speech recognition is a very large unexplored area because speech recognition is crucial for many applications where the voice information is preferred over text information.

Speech can be used in the real time application like if you have customers who want fast access to information then speech identification is required less time to providing a information in less time. In many circumstances, customers do not need or want to speak to a live operator. For example, if they have little time or they only require basic information then speech identification can be used to reduce the waiting times and provide customers with the information they want and in many situation that previously required human physical work involvement, such as recognizing simple spoken words to perform something like turning on fans or closing and opening of a gates. Speech can be used for security application also like as a password of personal computers using voice and word recognition, so accuracy in speech recognition is very important task. To increase in accuracy in recognition rate, more techniques have been used such as dynamic time warping, Artificial Neural Network, and Gaussian Mixture models have been used. In the present time Technology becomes very advanced. So identification of more complex speech is possible. Although current technology is still far away from identification of human speech to 100%.Thus much more research and development is required in this area. Therefore we explore in this field to give more development in a real-time-based speech application.

1.3 Background

The need for speaker recognition is verification method in telephone and internet transactions. Working with speech response based health and banking system and we can also use voice as audio signatures for digital document and in online education systems.

Design a machine which works on human speech has intrigued scientists and engineers for centuries. In 1930s, when Homer Dudley of Bell Laboratories design a system for speech synthesis and analysis the problem of speech identification for reorganization only for small set of sample of sounds and quiet room fixed high –quality mic, careful reading and application specific language is required but based on many advance research now speech recognition is used in many application which requires a human machine interface.

Today there are so many algorithms developed for speech recognition for feature extraction like Linear predictive co-efficient (LPC), and Principle component analysis (PCA) and for feature matching like Vector quantization, Hidden Markova Model, Artificial Neural Network but these algorithms are not good for practical speech identification due to some features such as high redundancy, complexity, accuracy and less correlation between original speech of speaker and tested speech sample. The most important thing that these algorithms are not able tested the original speaker.

So, it is important to develop an efficient algorithm for speech identification System, especially for the real-time application where speech is easy to access. Previous algorithms work on the neutral speech not in whisper speech so; we are going to develop an algorithm in which we are work on neutral and whisper speech.

1.4 User

The main idea behind this project is identification or recognition of speaker among many different speakers. The training phase is done with Mel Frequency Cepstral Co-efficient and testing phase is done with K-means Algorithm, LBG Algorithm using Vector Quantization and Gaussian Mixture Model (GMM) in which GMM gives more accuracy than other.

For good identification of speaker speech, we use Mel Frequency Cepstral Co-efficient which gives good results in between frequency range of human voice in a whisper and neutral speech. The overall this proposed model is to increase the rate of identification of speakers in any circumstances like increased number of the database for speakers or increased number of utterances and especially to recognize sounds consisting nasal consonants, reverberate sounds.

1.5 Problem Overview

Speech Recognition is used mostly in a biometric-based system which gives stem in speech processing area. There are many more activities in which speech processing is used. Speech recognition is a multidisciplinary problem regarding pattern recognition techniques and domain related problem like Acoustic and phonetics are necessary. In speech area there are many problems in which we work with them and these are like-

1. First, Body language -When a speaker is communicate with speech he is not only speech but also gives a body language like a postures, eye movement and hand waving which is completely is missed by speech Recognition.
2. Second, Noise - Speech is produced in an environment of sounds like- a clock ticking, a computer humming, another person is speaking in the background etc. These are the noise in speech signal which unwanted information. In speech identification these noises degrade the accuracy of the speech identification process.
3. Third, Spoken language- spoken language is important issue is in speech because when we speak we utter by with disfluences in words e.g. speech is filled with hesitations, repetitions, changes of subject in the middle of an utterance, slips of the tongue etc and this is most in whisper speech then normal speech. A human listener is usually differentiating the disfluences but in speech identification it is not identify by the system.

There are many problems in this field such as a speaking style of person, speed of speech. In this thesis we define some problem on speech but not all of them. Speech Identification is difficult task because it has large work search space and variability.

1.6 Objective

The objective of the project is to develop an efficient algorithm for the speech recognition in speech based applications.

- This thesis report considers an objective of speech identification technology, its application and software development.
- Speech identification has been one of the key technologies in speech processing area during the past decades. In speech identification, the speech degraded by acoustic noise and resonance of frequency is still considered a very challenging problem.
- Here we aim to improve the speech identification robustness in noisy conditions by designing pre-processing algorithms.

- Finally, the report concludes at the different potentials uses of the application and further improvements and considerations.

1.7 Overview of the report

The overview of this report is as follows

Chapter 2: In this chapter, we describe the nature of speech signal, phonetics, types of speech Voiced and Unvoiced decision making. In this, we also describe database which we are used in this project.

Chapter 3: It describes the method of feature extraction of the speech signal. We prefer to concentrate on a fundamental frequency of pitch in the proposed algorithms like, Mel-frequency cepstral coefficients (MFCC) and Exponential Frequency Cepstral Co-efficient (EFCC). Methods of MFCC like pre- emphasize, framing, windowing that is used during feature extraction are discussed. Results of different feature extraction methods are compared.

Chapter 4: Here we discuss K-means Algorithm and Gaussian Mixture Model and basics like-Gaussian distribution function and Gaussian mixture model.

Chapter 5: In chapter 5 we discussed about implementation steps of algorithm which we are proposed for training and testing of speech signal.

Chapter 6: In this chapter we discussed the results obtained. Here we compare the different differences in feature extraction methods and feature matching methods used. The recognition rate of the speaker is compared.

Chapter 7: Chapter 7 shows the conclusion and the future work in speech processing area.

Chapter 2

Literature survey and background theory

2.1 Domain Clarification

Here, we are going to describe the analysis of the speech of human voice and the available solution to recognition of human voice and find the initial step to be followed in our proposed method. Firstly we describe the different techniques detect the human speech features. Then, we secondly analysis the previous year papers and then we will see the software support.

2.2 Speech Signal

The human voice contains a sequence of sound; basically the unit of human voice or speech signal is sound. The expression of feeling and thoughts by sound is speech. Speech is random signal because they occur naturally. There are many algorithms which are developed by researchers on speech signal. The idea developed by the researchers based on their research of voice signal that the fundamental of human voice is unique and frequency of human voice has different range. Human voice of frequency range is discussed below

2.2.1 Frequency range

The fact is that human ear hear the sound between the range of frequencies from 200 Hz to 2000 Hz with most accurate range to hear in the region between 300 Hz to 9 kHz and the major energy of signal is lies in between 300 Hz to 3400 Hz. Reducing the range of frequency band width 300 Hz to 3400 KHz can predominant reduce the speech clarity and speech identification, Over sound quality will be increased if we increase the bandwidth of the frequency, however to increase gains in sound quality we have to considered against increased frequency usage.

2.2.2 Classification of speech

The speech of human voice is classified into two parts, a voiced parts and an unvoiced part. Two Dimensional description of a sound is a waveform. Time and intensity is representation of a two dimension waveform. In sound waveform the vertical dimension is represented by the intensity and horizontal dimension with time. Representation of sound in time domain is called waveform as they are representation of changes in intensity over time. Speech signal is classified in four classes namely, consonant, vowels, semivowels diphthongs as shown in figure 2.1

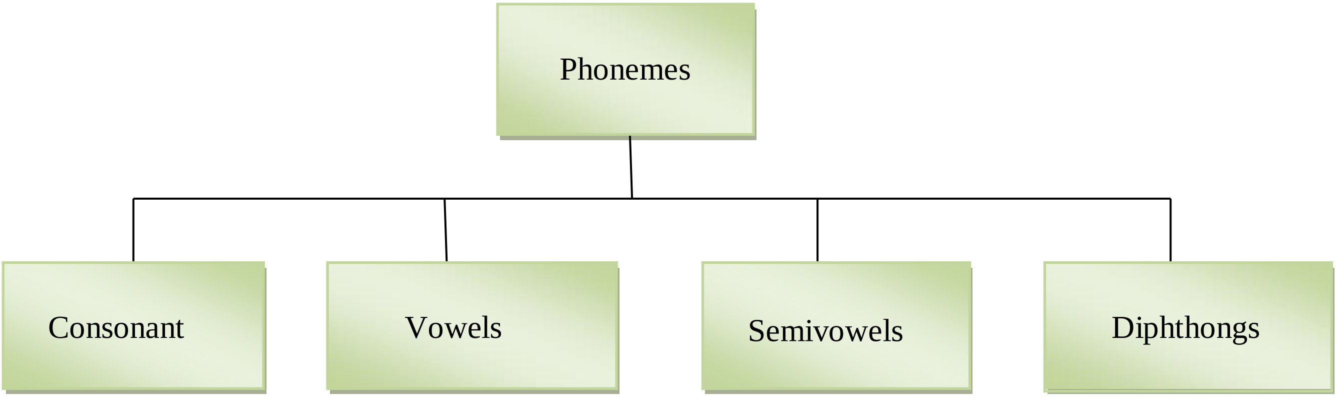


Figure 2.1 Phoneme Classifications

2.3 Literature review

Here we will go through some previous year's papers which are based on the speech identification using different algorithm. From last few years, in speech field, MFCC algorithm got much attention for feature extraction and GMM for classification. The researchers are developing a secure and robust algorithm for speech identification. MFCC algorithm works on the human hearing perception model which is based upon the natural selection. Retaining Speaker speech information by reducing data is the process of feature extraction techniques. To improve the accuracy it is not possible to meet all the criteria using single speech feature, so create a large number of speech feature and combine them to improve the accuracy. Speech Signal features are extracted like pitch and frequency to indentifying the emotion of human [1] To differentiate the speaker voice pitch is very useful because the range of the pitch of males is 60 to 120 Hz and female the average pitch is 120 to 200Hz while in children the average pitch is 200 to 400 Hz [2]. In Cepstral analysis method we extract the fundamental frequency which is based on cepstrum. The inverse discrete Fourier transform of a log magnitude of discrete Fourier signal is called cepstrum. Linear predictive coding (LPC) method is extracting the formant of the speech signal. A person voice of spectral peaks is called as formant. Human vocal tract have acoustic resonance is defined the formant and they are measured as amplitude peaks in the frequency spectrum of sound. LPC method is used in calculate the formant of the speech signal but. It extracts the feature of speech at lower order [1]. In 1976 B.S Atal used LPC method for speech feature. The speech parameter like the area function, auto-correlation and impulse response function and the predictor coefficient, the cepstrum function are derived from them. He used cepstrum function as input of speaker recognition and conclude that the cepstrum function is gives better results in speech recognition [3]. In 1994 Reynolds [4] find the different feature for speech, such as perceptual linear prediction cepstral coefficients (PLPCC),Linear predictive cepstral coefficients(LPCC), Linear frequency cepstral coefficients (LFCCs) and Mel frequency cepstral coefficients(MFCCs). He give the result form the experiment that LFCC and MFCC are best method than other feature extraction method because MFCC algorithm is divided the speech signal in to number of frames and coefficient So it is easy and fast to access[5].

2.4 Database

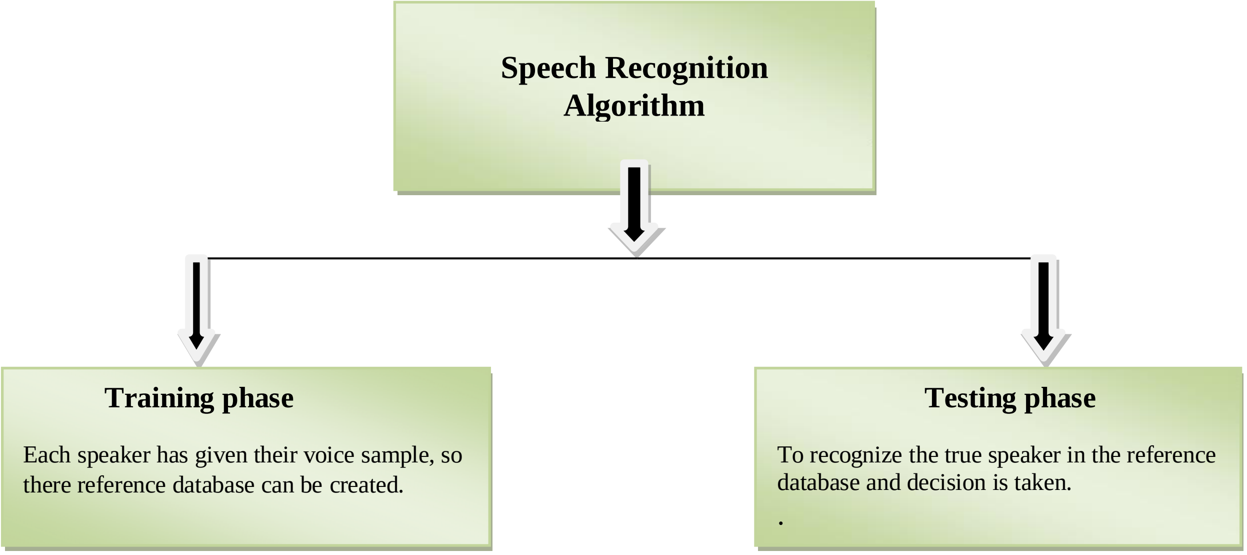
Database contains of two sentences each of ten second is phonetically balanced of five different speaker in both natural and whisper mode. They are recorded by normal laptop. These database spoken by five different speakers, among which two are male and three are female. These samples are created in ‘.wav’ format. This recorded utterance is taken as sampling frequency of 16000 Hz and 16-bits per sample. The ‘.wav’file is easy to access in MATLAB software tool in which speech is read in ‘.wav’ format.

Chapter 3

Feature Extraction of speech signal

In previous chapter, we have seen the various algorithms which were developed by researchers in the field of speech recognition but we see there are only few algorithms are robust in speaker identification in speech domain. After seeing this we realize that there will be need to develop an efficient algorithm which should be robust and accurate for speech recognition. After performing the literature survey, we realize that in above MFCC and LFCC gives best result in speech identification. But for both natural and whispering voice detection MFCC algorithm give more accurate result. So, here we are going to discuss the proposed idea to implement the algorithm.

The speech recognition algorithm is classified in two parts. The first part is training phase and the second part is testing phase as shown in figure 3.1[6].



Speech Recognition Algorithm

Training phase

Each speaker has given their voice sample, so there reference database can be created.

Testing phase

To recognize the true speaker in the reference database and decision is taken.

.

Figure3.1:- Process of speech identification

This chapter introduces the training phase to convert a speech signal in to significance feature vector and extracting the parameter of speech such as pitch of human voice and cepstral co-efficient which is useful in testing phase. MFCC is based on human hearing perception method with respect to Fundamental frequencies and cepstral Co-efficient therefore it is most accurate for speech recognition. We will explain the MFCC algorithm step by step in this chapter.

3.1 Feature Extraction in Training phase

The Feature extraction is a fundamental classification of speech signals and it is a best method to extract the feature of speaker. The accuracy is given by this phase is important for the next step because it reduce the model accuracy. This method is a mathematical representation of the speech signal. We have used MFCC algorithm for feature extraction.

3.2 Mel frequency cepstral co-efficient

Short time power spectrum of speech signal is defined as a Mel frequency of cepstrum.which is calculated as the linear cosine transform of the log power spectrum on a non-linear Mel scale of frequency. MFCC is used to calculate the two frequencies. First is linear frequency which is below 1000 Hz and second is logarithmic frequency which is above 1000 Hz. Mel frequency scale represent the important feature of the speech. The step for MFCC algorithm process is as follow-

- Record the speech signal word and utterance and pre-emphasis the signal.
- Divided the signal in to number of frames and applied in to hamming window function to make in continuation.
- Take the FFT of the window function and then squared magnitude of signal to find the power spectrum.
- Applied Magnitude spectrum in to Mel filter bank which is bank of triangular filter.
- Then Find the Discrete Cosine Transform of the signal to convert in time domain.

Flow Diagram for the Training phase using MFCC algorithm

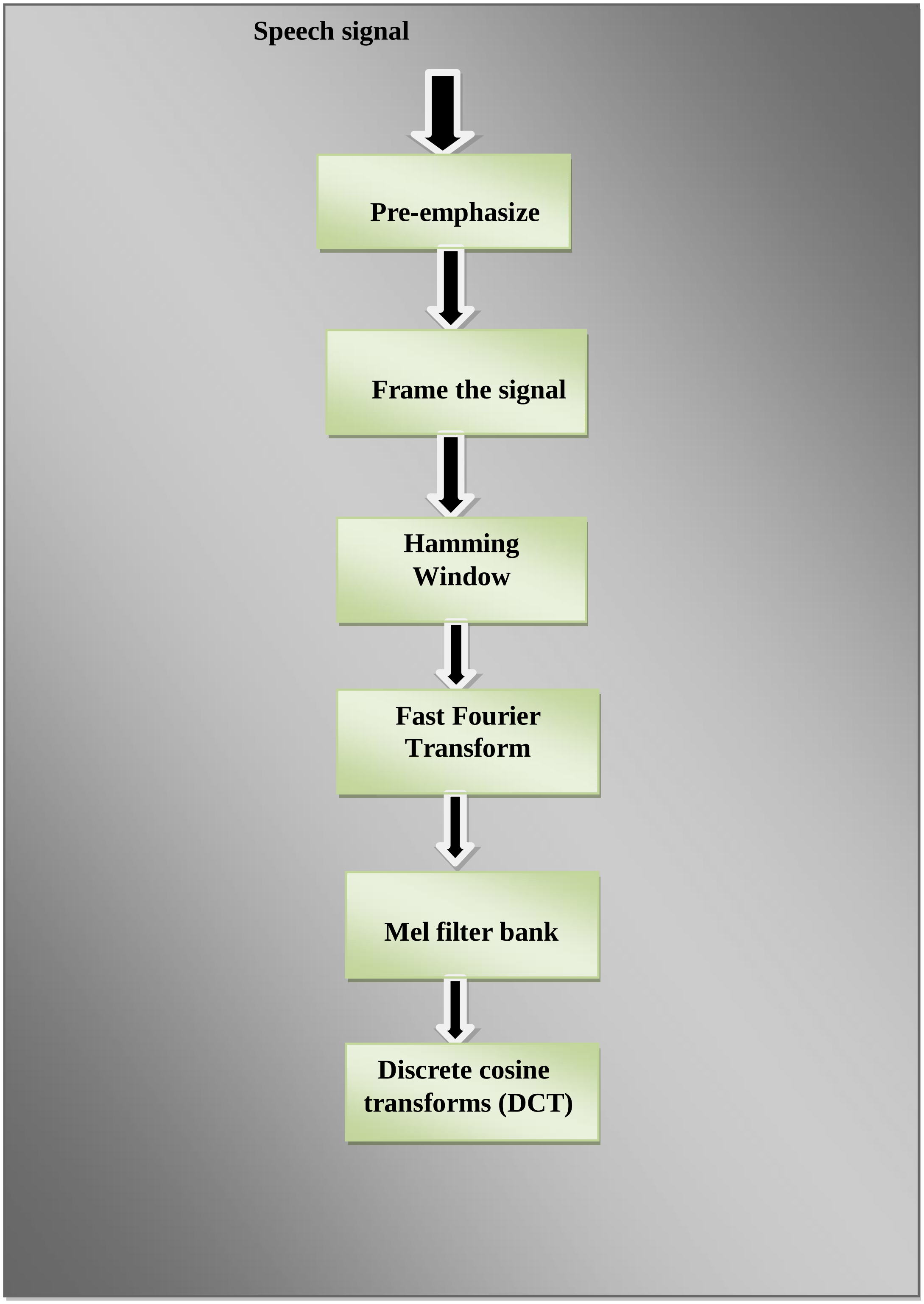


Figure 3.2 Block diagram of MFCC

3.2.1 Pre -emphasis

Human personality character is identifying by his/her vocal tract. Pre-emphasis is used to enhance the some feature of human vocal tract system and high frequency of the speech signal because during speech recording high frequency is lost. To perform the pre-emphasis high pass filter is used [7]. The Function of pre-emphasis is [8]-

$$X(n) = S(n) - k*S(n-1) \tag{3.1}$$

Where, k is pre-emphasis co-efficient its value is lies between 0.9 to 1

S (n) Original speech signal and

X (n) after pre-emphasis speech signal

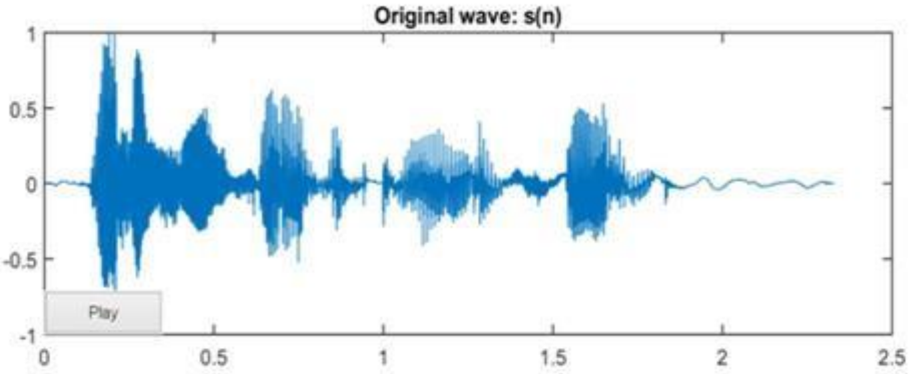


Figure 3.3: Signal before pre-emphasis

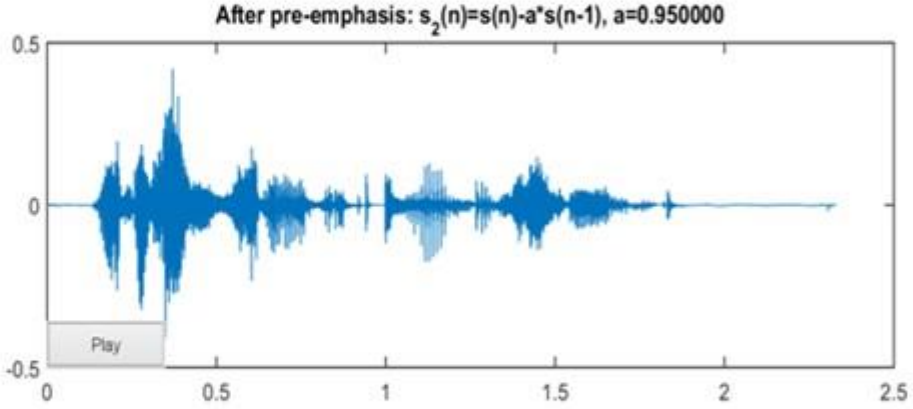
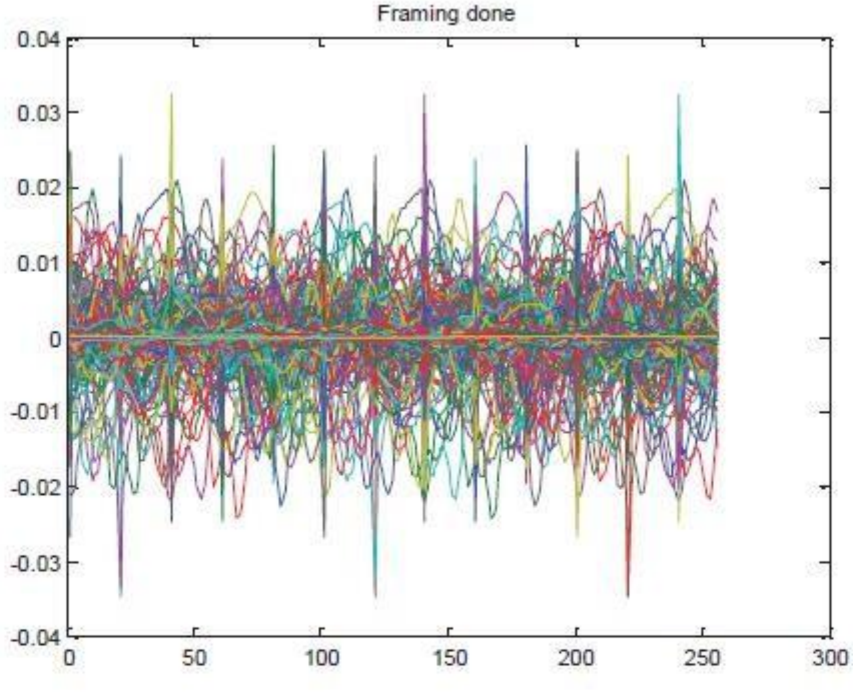


Figure 3.4: Pre-emphasis signal

3.2.2 Framing

Framing is useful in converting the speech signal from analog to digital. The speech samples after pre-emphasis segment in to 20-30ms frame. The standard frame size is 25ms. The length of the frame of 16 KHz frequency of speech signal is 400 samples. This is calculated by the multiply 25ms frame size with 16 KHz frequency. So for N samples the starting frame contain N sample and next one after the starting frame is M sample and overlaps it until the speech signal is end. M sample we usually chose is100 samples and for N is 256 (M<N).



.**Figure 3.5 framing of speech signal**

3.2.3 Hamming Windowing

Richard W.Hamming proposed the Hamming window function. After framing the speech signal we multiply the hamming window function with each frame to maintain the continuity of first frame to last frame. Hamming window is best in speech recognition area because in unvoiced signal fundamental period is not define so it is required the hamming window for each frame and in frequency domain its gives better result while in time domain we use rectangular window. Hamming window shape is like a cosine signal and is similar to Hann Window.

The Hamming window of each frame is like-

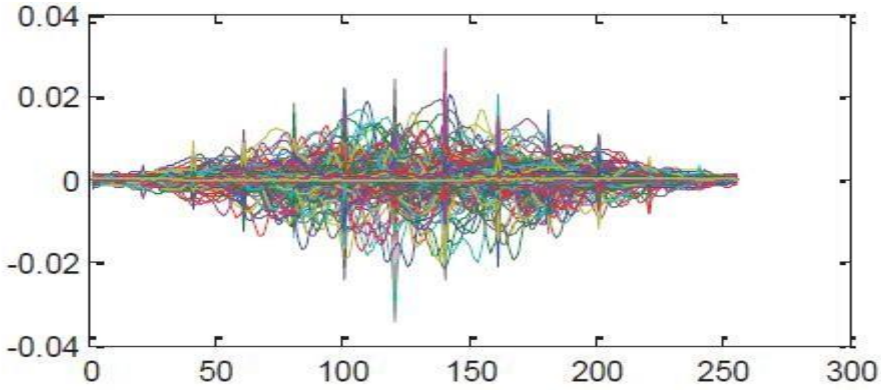


Figure3.6 Hamming Window of signal

The Hamming window function is

$$W\ (n) = \alpha - (1- \alpha)\cos\frac{(2\pi n)}{(N-1)} \qquad (3.2)$$

If W (n) is defined 0 to N-1 and α is a constant its value is 0.54

And n is defined between 0 to N-1

Where, N is number of samples in each frame

For different values of α hamming window function is like-

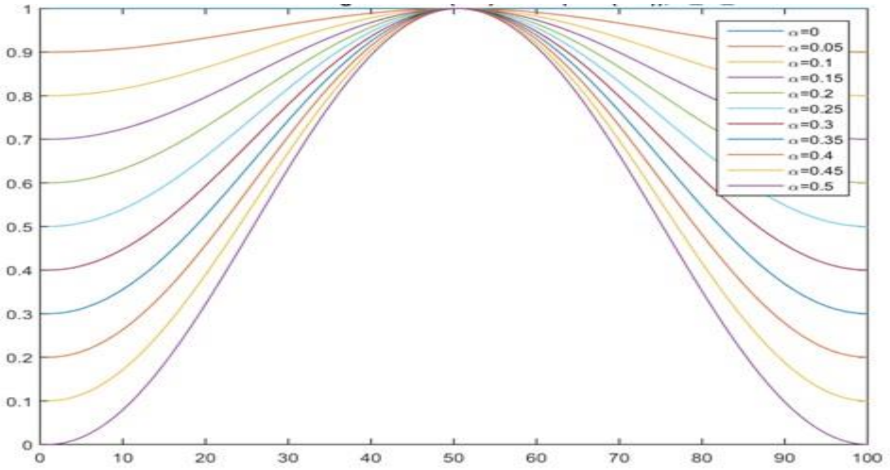


Figure 3.7:- Hamming window at different value of alpha

Resulted signal is

$$H\ (n) = S\ (n)*W\ (n) \qquad (3.3)$$

H (n) = resulted signal

S (n) = input signal

W (n) = Hamming window

Hamming window gives best performance than three other window rectangular, hann and triangular window [9]. If hamming window is not used after framing the signal then the first and last point of discontinuity makes peak wider and lesser of the frequency response so the use of hamming window make the peak sharper and clear in the frequency response[13].

3.2.4 Fourier Transform

The Fourier transform is used to convert time domain signal in to frequency domain is widely used in speech and signal processing area. In this each frame of time domain convert in frequency domain. Spectral analysis of different tone of speech signal shows the different energy distribution of speech signal. For magnitude spectrum of the each frame FFT is used.

In signal area we used fast Fourier transform as-

$$F\ (f) = \int_{-\propto}^{+\infty} (t)^{-j2\pi f t}\ dt \qquad (3.4)$$

So after Hamming window result we take FFT of H (n)

$$Y\ (w) = FFT\ [H\ (n)*\ X\ (n)] = H\ (w)\ *X\ (w) \qquad (3.5)$$

The Fourier Transform of signal is Y (w).

3.2.5 Triangular Filter Bank

In Triangular filter bank we multiply the K number of triangular band pass filter with power Spectrum of FFT signal. Typically the value of Triangular Filter Bank is taken 20. We calculate the power spectrum of FFT signal because human cochlea received sound in the form of vibration which vibrates at different frequency of sound. The location of the cochlea is particular telling the human brain nerve system that which frequency is present in speech signal. The cochlea is not differentiate the two closely spaced frequencies, so by this analysis we divide our periodic spectrum in to number of small group and sum of the entire group to find how much energy is present in various frequency regions. This is developed by the method of triangular filter bank. Triangular filter bank is divide in to two parts high frequency region and low frequency region and spacing or how much filter bank is wide is calculated by the Mel scale. The Mel frequency is represented on linear frequency as shown below the equation-

$$\text{Mel}(f) = 1125 * \ln(1+f/700) \qquad (3.6)$$

3.2.6 Mel frequency Scale

Above we calculate the filter bank energy then we have taken the logarithm of the filter bank energy because it is also based on human hearing perception system. We can not hear the loudness below 1000Hz frequency of linear scale, so we generally put 8 times energy to perceive the volume of a sound. The use Logarithm scale here because of channel normalization. Non-linear frequency of signal is heard by human’s ear, so Frequency of non-linear is represented by the logarithm scale above 1000 and below 1000 frequency is linear so it is represented on normal linear scale. Mel is basically taken the word melody.

The below diagram shows the relationship between the linear and Mel frequency.

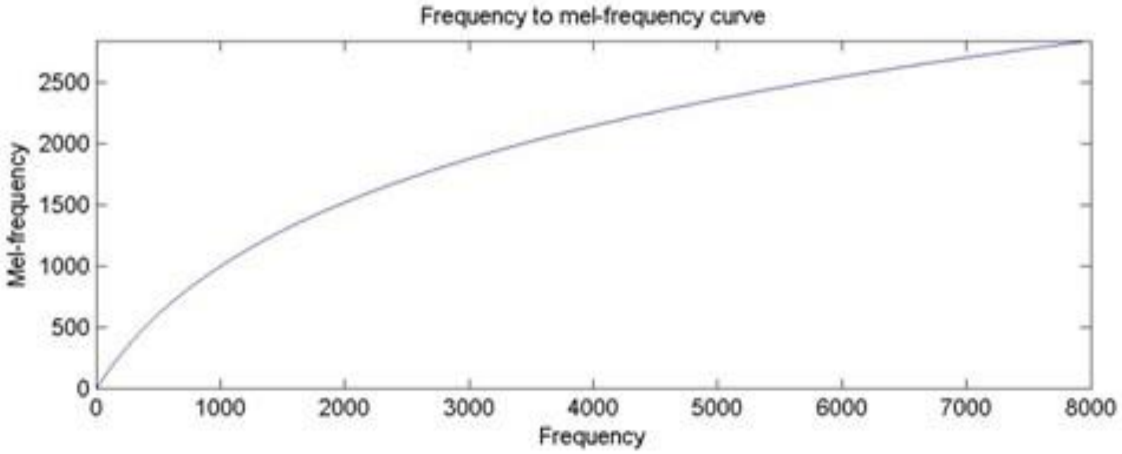


Figure 3.8: Mel frequency on linear scale

There are two criteria in which we use triangular filter bank as shown below –

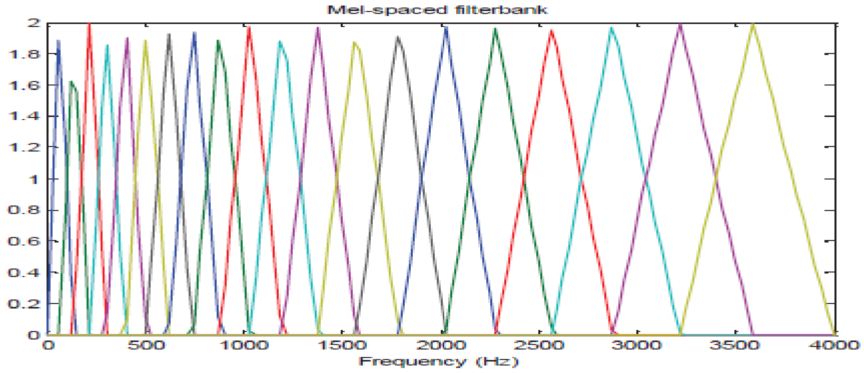


Figure 3.9: Mel filter bank

1. Harmonics of speech signal is in the form of envelop of the spectrum, so get the harmonics flattened we smooth the amplitude spectrum by the triangular filter.

2. Feature size of speech signal is reduced.

3.2.7 Discrete cosine transforms (DCT)

The last step is a discrete cosine transform. DCT is used to remove the noise in the speech signal. It is used for discrete the noise and speech energy as shown in this paper [11] that the DCT used DFT to unify the speech energy. We find the DCT of log energy filter bank. This is done because of two main reasons first is filtterbank is overleaping and second is that the energy of filter bank is correlate with each other. The Decorrelation is done by the DCT and we take only 12 out of 26 filter bank because the fast change in filterbank reduces the accuracy in the speech recognition. The calculation of DCT term is given by this function-

$$C_{m} = \sum_{k=-1} Ncos\ [\ *\ (K-0.5)*{}^{\pi}]_{*}\underset{N}{\_}E_{k} \qquad (3.8)$$

Where MFCC feature vector is represented by the C_m and

N represent number of triangular filters represents order of cepstral coefficients. The extracted feature vector of the speech as shown below

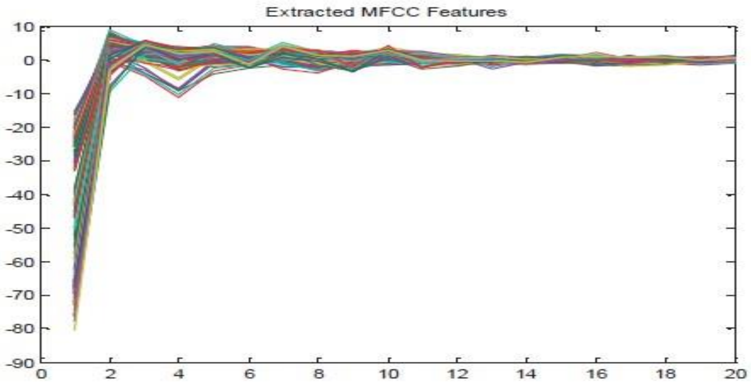


Figure 3.10: MFCC feature vector

3.4 Exponential frequency cepstral co-efficient (EFCC)

Unvoiced consonant and Voiced Consonant and other phonemes have different spectral energy in speech analysis. Extracting the Feature vector of speech signal by different scale it is best extract the feature vector and other phonemes one scale. From above analysis we saw that MFCC is emphasis the low frequency component than high frequency component, So MFFC is not useful in voiced and unvoiced consonant. Thus, EFCC is used for high frequency component and other phonemes. Extracting the Feature vector by exponential frequency scale is same as proposed the method in speech under stress [12]. The exponential frequency function is defined as-

$$Y = C * \left(10^{\left(\frac{f}{k}\right)} - 1\right), \quad 0 \leq f \leq 8000 \text{ Hz} \qquad (3.10)$$

Where C and k is calculated by these two equations:

$$C * \left(10^{\left(\frac{8000}{k}\right)} - 1\right) = 2595 * \log\left(1 + \frac{8000}{700}\right) \text{ ------------} \qquad 1$$

$$\{c, k\} = \min \{|\ (10^{4000/k} - 1) - \frac{4000}{k} * \ln 10 * C * 10^{\frac{4000}{k}}|\} \text{ ------------} \qquad 2$$

Put f = 4000 chosen the value from method [13] and solve the equation we get,

c = 6375 and k = 50000, so the exponential scale function is like-

$$y = 6375 * (10^{\left(\frac{f}{50000}\right)} - 1 \qquad (3.11)$$

The flow diagram of EFCC is as-

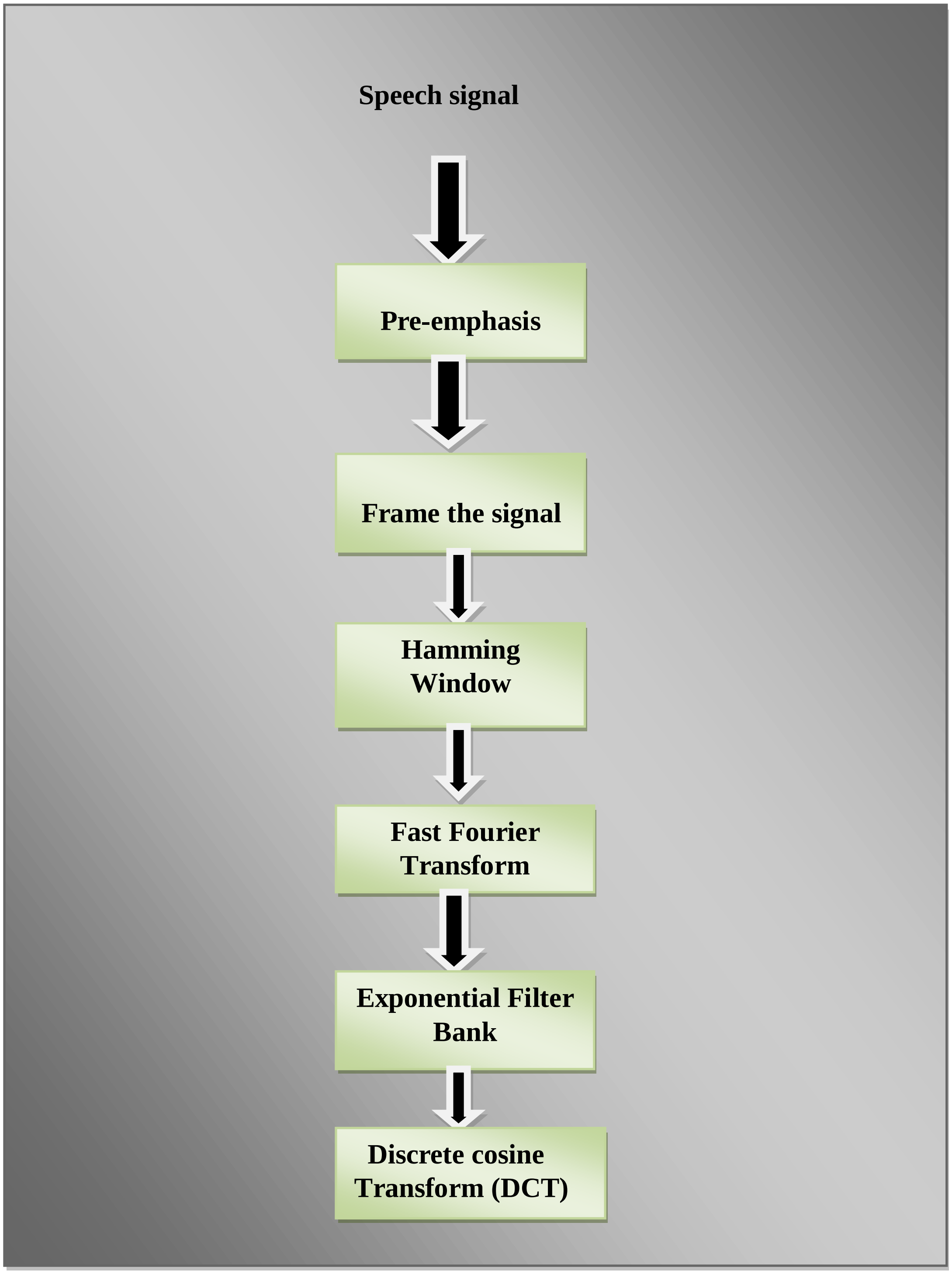


Figure 3.11 Block diagram EFCC

In the above Figure EFCC is used exponential filter bank rather than Mel filter bank. Unvoiced consonants and phonemes are processed by EFCC to enhance the whisper speech signal.

3.5 Overview

In this chapter, we have defined the proposed work in which we define each stage of the algorithm. In section 3.1, we discussed introduction in which we describe that what the method in speech identification process is. In section 3.2, we discussed about the MFCC algorithm and flow diagram of the algorithm, in which we describe what method for Feature extraction and how we are going to deal them. In section 3.3 we discussed about the EFCC algorithm for whisper speech and we discussed about the flow diagram of the algorithm in which we describe the different stages of the algorithm. Now in next chapter we are going to discuss about the classification of speech signal in details.

CHAPTER 4

Feature Classification of speech signal

Classification is building a unique model for every speaker in the database. Here we are using two methods of classification.

- Stochastic Model
- Template Model

1. A stochastic model is used in where random data, event, and uncertainty are present. It cannot predict the state with full accuracy, and its non-deterministic behavior is anatomization by probability theory. A stochastic model is opposite of deterministic model whereas the deterministic problem only has one solution while the stochastic model is complicated. In stochastic model based there are many algorithms like GMM, HMM, and ANN.
2. A template is a collection of features vector which can be defined as the repetition of data and continuation of frames. In the template based model, we cannot predict the prior information, and it is not a probabilistic model. In this, each frame is compared with the training frame and finds the minimum distance, so it’s weak in performance as compared to HMM and GMM. Template based model is Vector Quantization (VQ) and Dynamic time warping (DTW).

4.1 Stochastic Model

4.1.1 The Gaussian distribution

The term Gaussian is derived by a German mathematic scientist in 1809 to describe the air curve in planetary orbit. It is the distribution of statically mathematical model. It is normally referred to a normal distribution or bell curve. The Gaussian distribution formed when Multiple sources of variability and additively. The distribution is often a reasonable assumption and well behaves, Mathematical model. It gives central role because of the important mathematical theorem. The Gaussian distribution function is like that-

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \tag{4.1}$$

Where σ is the Standard distribution and μ is the mean

The Gaussian distribution is a probability distribution function curve. It can be proven (Central limit theorem) that if there are enough random uncorrelated things going on, you will always get Gaussian distribution function and most of the statics assume the Gaussian. The Gaussian curve is a basic model in our algorithm which is based on Gaussian Mixture Model. The Gaussian Curve is-

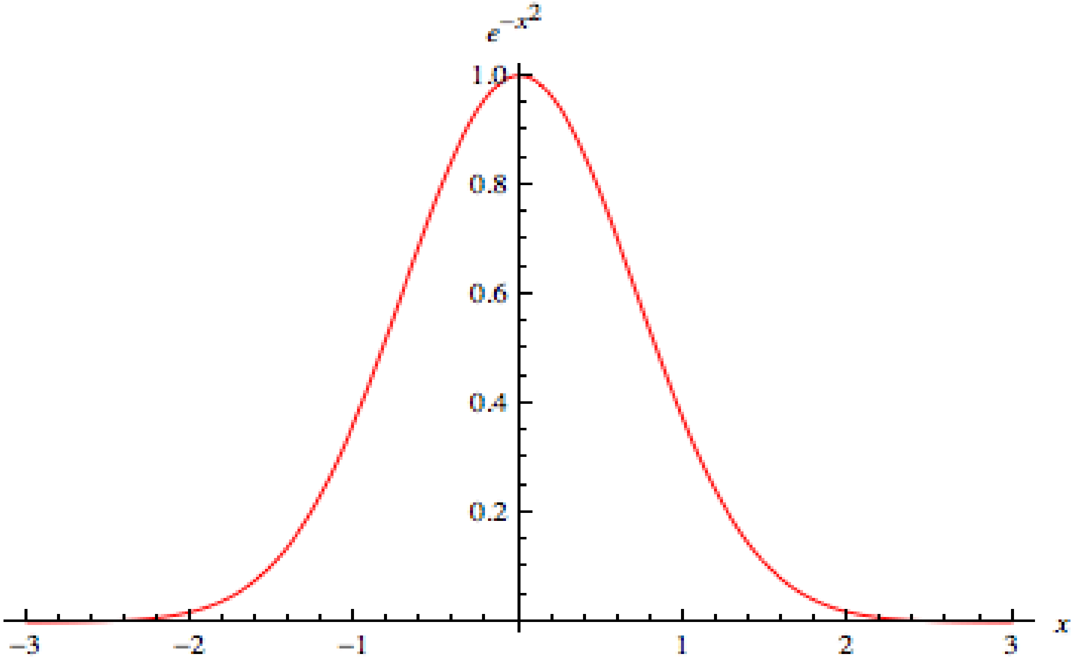


Figure 4.1: Gaussian Curve

4.1.2 Gaussian Mixture Model

Gaussian Mixture Model (GMM) is a probabilistic model which is used to define the probability density function of a random variable regarding weighted sum of its component.

In which each component is Gaussian density function. Like that-

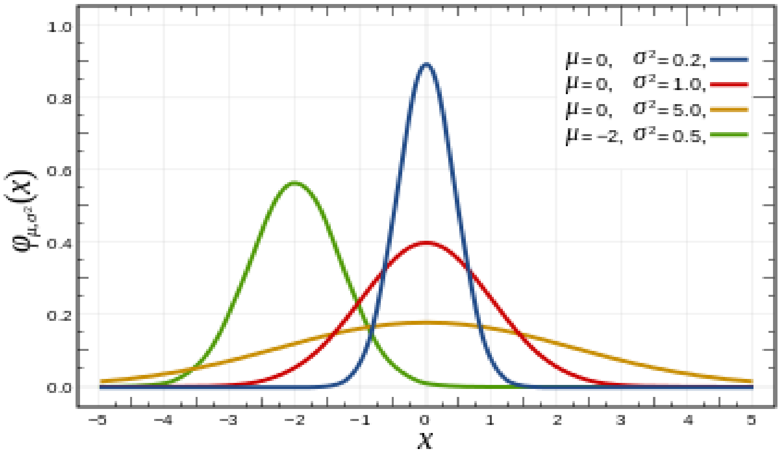


Figure 4.2: Gaussian Mixture Model

The Gaussian mixture model is used for clustering the data in this we use hard clustering method. A Gaussian mixture model (GMM) is used where normal distribution is failed it is a mixture varies probability density of normal distribution.

The Gaussian distribution for univariate is-

$$\mathcal{N}(\mathbf{x} \mid \underbrace{\boldsymbol{\mu}}_{\text{mean}}, \underbrace{\boldsymbol{\sigma}}_{\text{variance}}) = \frac{\mathbf{1}}{\sqrt{\mathbf{2\pi\sigma^2}}}\, \mathbf{e}^{-\frac{(\mathbf{x}-\boldsymbol{\mu})^2}{2\boldsymbol{\sigma}^2}} \tag{4.2}$$

Where x is a data point and we have calculated mean and variance for single speaker and for multi-speaker we have calculated the mean, covariance by multi-variant Gaussian distribution-

$$\mathcal{N}(\mathbf{x} \mid \underbrace{\boldsymbol{\mu}}_{\text{mean}}, \underbrace{\Sigma}_{\text{covariance}}) = \frac{\mathbf{1}}{(\mathbf{2\pi}|\Sigma|)^{1/2}}\, \mathbf{exp}\left\{-\frac{\mathbf{1}}{\mathbf{2}}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\, \Sigma^{-1}\, (\mathbf{x}-\boldsymbol{\mu})\right\} \tag{4.3}$$

We estimate the multivariate parameter of mean and covariance by the Expectation Maximization Algorithm.

4.2 Expectation Maximization Algorithm

There are two types of clustering method soft clustering and hard clustering. In soft clustering, data cluster may overlap but in hard clustering data do not overlap either data belong that cluster

Or not. Here we are using soft clustering in Mixture model because soft clustering data can belong more than one cluster it's created like a hierarchical model. Each cluster basically corresponds to a probability distribution, and each part contains a parameter like mean, covariance and a weighted average of the probability distribution are calculated by the Expectation Maximization Algorithm (EM). EM algorithm automatically finds the parameter of Gaussian distribution like we have a bunch of data points and these data points come from two sources and each source is coming from Gaussian then we estimate the mean and variance.

EM algorithm is working on Maximum Likelihood (ML) Method. In ML method we first took the log of multi-variant Gaussian distribution like that-

$$\ln p(\mathbf{x}\,|\,\boldsymbol{\mu},\Sigma) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{T}\,\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) \tag{4.4}$$

And then maximize it by taking the derivative of the distribution and equate to zero

$$\begin{array}{cc}
\dfrac{\partial \ln p(\mathbf{x}\,|\,\mu,\Sigma)}{\partial\mu} = 0 & \dfrac{\partial \ln p(\mathbf{x}\,|\,\mu,\Sigma)}{\partial\Sigma} = 0 \\
\Downarrow & \Downarrow \\
\mu_{\mathrm{ML}} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N}\mathbf{x}_n & \Sigma_{\mathrm{ML}} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N}(\mathbf{x}_n-\mu_{\mathrm{ML}})(\mathbf{x}_n-\mu_{\mathrm{ML}})^{T}
\end{array}$$

Here N is the number of sample and data point. We use this calculated value of mean and variance for probability distribution of data

$$\mathbf{p}(\mathbf{x}) = \underset{\text{Number of Gaussians}}{\underbrace{\sum_{k=1}^{K}}}\ \underset{\text{Mixing coefficient: weightage for each Gaussian dist.}}{\underbrace{\pi_k}}\,\mathcal{N}(\mathbf{x}\,|\,\mu_k,\Sigma_k)$$

Where π_k is the probability of prior weight of the k_{th} Gaussian and summation of the probability of any event is equal to one and for normalization lies between 0 to 1.

$$0 \le \pi_k \le 1, \sum_{k=1}^{K}\pi_k = 1$$

And now we calculate the likelihood by this equation-

$$\ln p(\mathbf{X} \mid \mu, \Sigma, \pi) = \sum_{n=1}^{N} \ln p(\mathbf{x}_n) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathrm{N}(\mathbf{x}_n \mid \mu_k, \Sigma_k) \right\} \tag{4.5}$$

Here ML is not converged the data points, and no closed form of solution is obtained, so we use further step by the EM algorithm –

For a given data points we can evaluate the posterior probabilities of the given data by Bayes rule so Firstly we calculate the mean, covariance and mixing co-efficient and evaluate the given data points initial value of log likelihood . Now E-step is responsible for calculating the current values of the parameter -

$$\gamma_j(\mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x} \mid \mu_j, \Sigma_j)} \tag{4.6}$$

Where $y_j(k)$ is the latent variable and M-step re-estimate the values of current variable until it converges -

$$\mu_j = \frac{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)\mathbf{x}_n}{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)} \qquad \Sigma_j = \frac{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)(\mathbf{x}_n - \mu_j)(\mathbf{x}_n - \mu_j)^T}{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)} \qquad \pi_j = \frac{1}{N} \sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)$$

And then evaluate the log likelihood

$$\ln p(\mathbf{X} \mid \mu, \Sigma, \pi) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathrm{N}(\mathbf{x}_n \mid \mu_k, \Sigma_k) \right\} \tag{4.7}$$

And if it does not converge then it calculates the E-step, M-step, and likelihood.

4.3 Template Model

4.3.1 Introduction

The Template based Vector Quantization (VQ) is the technique is used in a speech application, an image application, speech identification and speech synthesis is a most successful algorithm. Vector quantization is a classical quantization technique from signal processing that allows the modeling of probability density functions by the distribution of prototype of the vectors. It was originally used for data compression. It works by dividing a large set of points into a group having approximate the same number of point's closet to them. Each group is a represented by a centroid point, as in K-means and some other clustering algorithm the density matching property of vector quantization is a powerful, especially for identifying the density of large and high dimension data. Since data are represented by the index of their closet centroid, commonly occurring data have low error and rate data high error. So VQ is suitable for lossy data compression. It can also be used for lossy data correction and density estimation. A simple training algorithm for vector quantization is picked a sample point at random move the nearest quantization vector centroid towards the sample point by a small fraction of distance is a repeat. The vector quantization based algorithm is LBG using K-means algorithm. Here we are using vector quantization based k-means algorithm as a feature classification.

4.3.2 K-means Algorithm

There are many ways to classify the data like semi-parametric, parametric and non-parametric approaches. In parametric data is found by the known distribution and in semi-parametric the data is calculated by a mixture of Gaussian while in non-parametric we have nothing to predict the data. K-means algorithm work on parametric approaches.

Macqueen in 1967 is proposed the K-means Algorithm. This Algorithm is based on the iterative procedure for defining the cluster here K represents the number of clusters. K-means clustering is classify the data based on the minimum distance of the centroid. We partition the whole space into k times. Its aim is to find K mean vectors (µ1, . . . , µK) the mean will be given the k cluster centroids. It is randomly chosen initial cluster centers from the data set.

- The input is a set of a data sample of speech vector $x_1 \ldots\ldots x_n$.

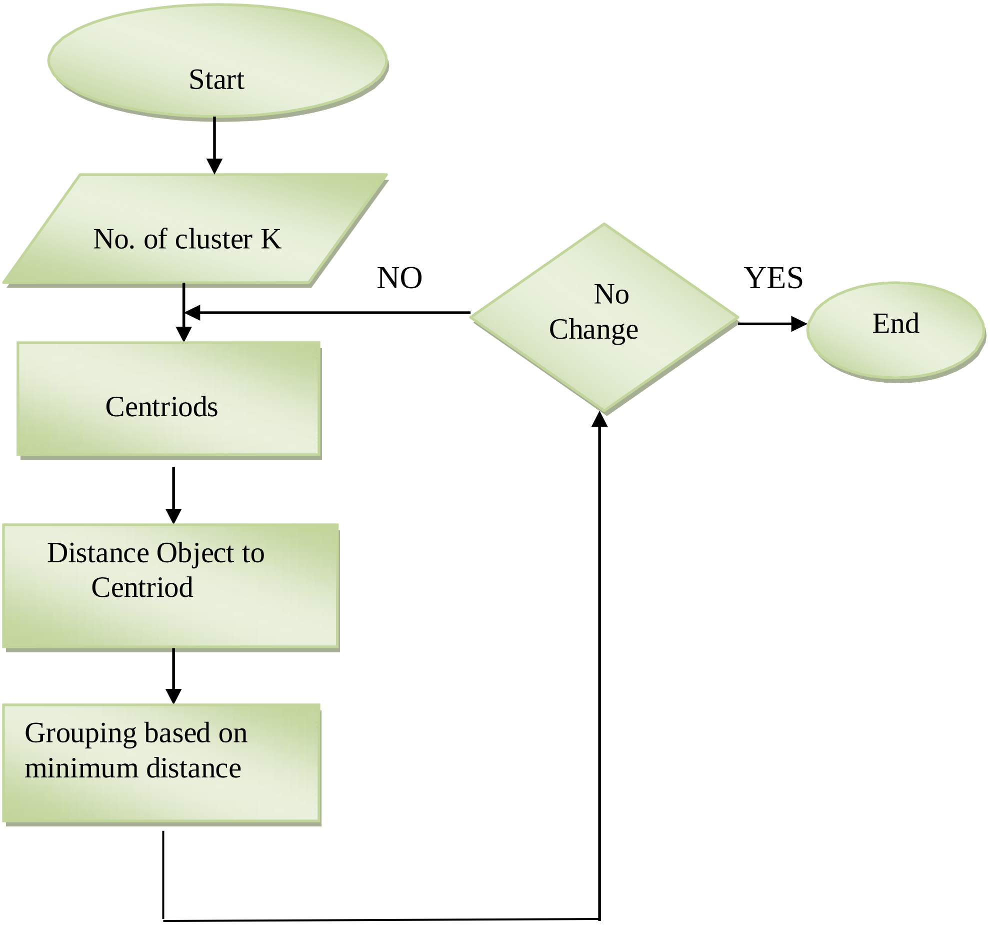

Start
No. of cluster K
NO
No Change
YES
End
Centriods
Distance Object to Centriod
Grouping based on minimum distance

Figure 4.3: K-means clustering

Algorithm step-

Step 0: Define the number of clusters K

Step 1: randomly chosen the cluster centroids ($\mu_1 . . . \mu_k$)

Step 2: Classify the data samples according to the nearest centroid (μ_k)

Step 3: Grouping the data sample based on minimum distance.

Step 4: To find the minimum distance we use Euclidian distance formula

Step 5: recomputed the centroid until ($\mu_1 . . . \mu_k$) until it is convergence.

K-means algorithm is an unsupervised learning algorithm. In this we first considered the speaker feature vectors as set of data points which is divided in to k number of cluster. Then centroids are randomly allocated to the k number of cluster. After defining the k number of centroid we find nearest data points to the centroid by the formula of minimum Euclidean distance and grouping these data points and calculate the mean of the data points.

These means defined as a new centroid. This process is repeated until the means are converge. Here converge satisfied the result if the new value of mean and old value of mean is same.

4.4 Overview

In this chapter, we have defined the proposed work in which we define each stage of the algorithm. In section 4.1, we describe the stochastic model and algorithm based on this model. In section 4.2, we discussed about the EM algorithm in which we describe what is the parameter for is calculated for method of Feature matching and how we are going to deal them. Now in next chapter we are going to discuss about the implementation step of our proposed work.

Chapter 5
Implementation

In this chapter we are going to discuss about the implementation of the algorithms that we had discussed in the previous chapters. In the previous chapters we discussed about the various stages of the algorithms in which we discussed each and every stage of the algorithms. Now, in this chapter we are going to discuss about the implementation of those stages and how we are going to use those stages in our algorithm.

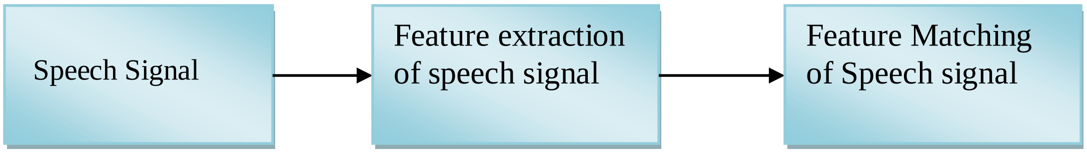
5.1 Algorithm

In this section, we will discuss about the algorithm which will be used to implement the speaker identification both in natural and whisper speech using Feature extraction method of speech signal is MFCC and EFCC and Feature Classification method of speech signal are GMM or K-means. We have implemented these algorithms in matlab. In matlab, we have implemented this algorithm on 10 sec of speech data which are phonetically balanced. In verilog we have also introduced speech identification scheme and how to identify the speaker using 32-bit floating point representation.

5.2 Speech Identification in Matlab

5.2.1 Training Algorithm in Matlab

These algorithms implemented in matlab on 10 sec of speech data. Steps of algorithms and block diagram of implementation as shown below:-


Speech Signal
Feature extraction of speech signal
Feature Matching of Speech signal

Step 1:- First we have recorded speech of 10 sec in '.wav' file of each speaker using matlab. Speech of speaker was digitizing with sampling frequency of 16 KHz and 16 bits per sample.

Step 2:- In the next step we created a database using 'db.mat'file in matlab which is contain of cell of matrix to store all the feature values of speech data.

Step 3:- After creating a database each cell of matrix is denoted the number of speaker which is store in the database. Each cell contain the MFCC feature vector, length of the signal ,mean, variance, weighted average of data, name of the speaker and type of their speech.

Step 4:- In the next step we will remove the silence part of the speech and send the speech signal in to MFFC algorithm for training phase.

- First we have done pre-emphasis of the signal and then divided the signal in to number of frame. One frame time is 32ms and one frame shift is 16ms and for calculating the sample duration in each frame multiply the frame duration with sampling frequency of 16 KHz and shift in frame sample multiply the sampling frequency with frame shift duration. If speech signal is not divided in to even number of frame then we padding it with zero to make even number of frame.
- After framing the signal we applied hamming window function to make it continuous. Here we given a window function which is a toolbox of window and it's default window is hamming window.
- Then in next step we convert time domain signal in to frequency domain using FFT and then calculated the magnitude spectrum of FFT signal.
- Magnitude spectrum of FFT signal is applied the 26 triangular bandpass filter which is equally spaced on Mel scale and calculate the filter bank energy of the unique part of the magnitude spectrum.
- Taken the log this filter bank energy and multiply with DCT to get cepstral co-efficient. And taken the cepstral co-efficient in to CC variable.
- Then these CC variables multiply with cepstral lifter which gives the lifter cepstral co-efficient.

- After performing all the operation we will get feature vector of speaker's speech and we will send this feature vector to the Testing phase.

5.2.2 Testing Algorithm in Matlab

In testing algorithm we are going to perform the speaker identification process on the extracted feature vector of MFCC algorithm by GMM and K- means algorithm.

Testing with GMM algorithm -

Step1:- First we calculate the MFCC feature vector of the testing speech and give these vectors in to GMM.

Step2:-. For testing of GMM we have already calculated the mean, variance and weighted Sum of all the speaker and store in database.

Step-3:- Now we are calculated the all parameter of given speech for testing phase of GMM Like- mean, variance and Weighted sum.

Step4:- These parameters of testing speech are given in to Expectation Maximization Algorithm to calculate the Log-likelihood and probability to find the true speaker.

Step5:- If log-likelihood new and old log-likelihood is less than 10^-5 then the algorithm is converge and give the feature of speech as a probability of the speaker.

Step6:- Then we calculate the mean and absolute value of these features and find the minimum difference with mean and absolute value of all the store speaker’s feature.

Testing with K-means classifier

Step1:- In first step we define number of cluster. Here we are using two clusters of the feature vectors of the speech.

Step2:- Next step we find the initial centroid of the clusters which is randomly initialize by the K-means function present in matlab .

Step3:- In third step we find the minimum distance of vector to given clusters centroid which is nearest to the centroid assign the vector to that centroid.

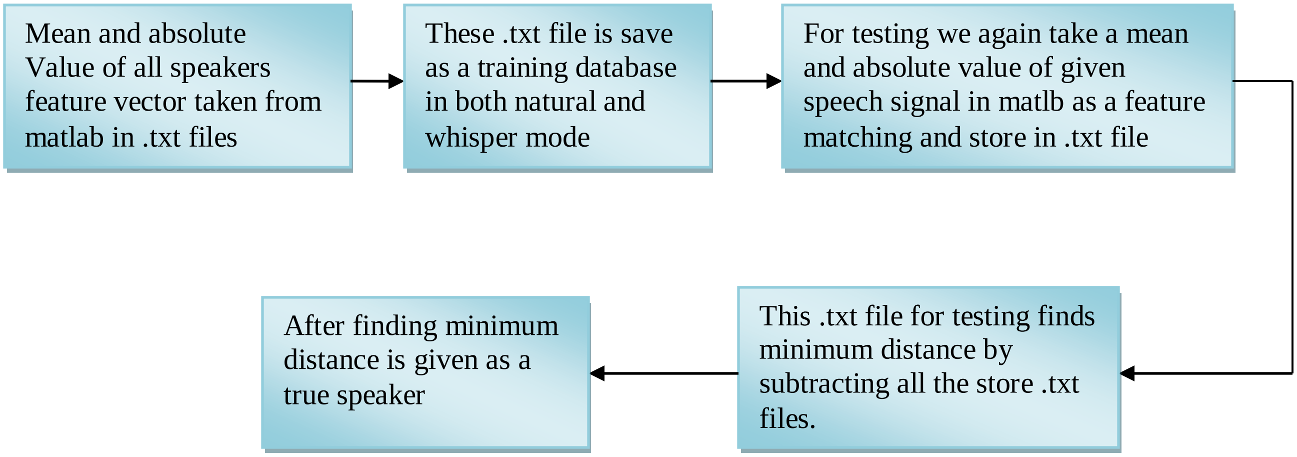
Step4:- Then we calculate the mean and covariance of the clusters around the centroid and define that mean and covariance as new centroid.

Step5:- This process continue until the value is not converge means the mean of the centroid is approached the same value again and again.

Step6:- In the last step find the minimum difference of the store mean value of all the speaker and calculated mean with k-means which mean value of the speaker is minimum that is the true speaker.

5.3 VLSI implementation of Speech Identification

Testing and training steps and block diagram in verilog is shown bleow-


Mean and absolute Value of all speakers feature vector taken from matlab in .txt files
These .txt file is save as a training database in both natural and whisper mode
For testing we again take a mean and absolute value of given speech signal in matlb as a feature matching and store in .txt file
This .txt file for testing finds minimum distance by subtracting all the store .txt files.
After finding minimum distance is given as a true speaker

Step1:- In this step we first find the mean and absolute value of the extracted MFCC feature vector of each speaker and store in one variable.

Step2: This mean value is a floating point number so we first convert these floating point number in to hexadecimal then hexadecimal to decimal and then decimal to binary because in matlab there is no matlab function which convert floating point number directly to binary.

Step3:- In step third the binary value of all the speaker is given in to text files to verilog. But these binary values is only 31 bits so we add extra sign bit to all the text file to make it 32-bit representation of floating point number because in verilog we do not deal the floating point number directly.

Step4:- To access these value in verilog we first developed the IEEE-754 standard 32-bits floating point addition and subtraction module.

Step5:- In the next step we find the difference between the store text files of each speaker as training and testing text file of the speech. If the difference of probability is equal which is found by matlab of any speaker text files then speaker is identify.

5.4 Overview

In this chapter 5, we discussed about the implementation of the speech identification algorithm. In section 5.2.1 and 5.2.2, we discussed about the training and testing algorithm and its implementation in Matlab. In section 5.3 we discussed about the implementation of speech identification in Verilog. Now in next chapter 6, we are going to discuss about the results obtained from those algorithm using verilog and matlab.

Chapter 6

Simulation Result

In the previous chapter, we have discussed the implementation of the algorithms of speech identification. These algorithm are implemented in matlab platform. We also have done speech identification in verilog using the result of matlab. Now in this chapter we are going to discuss the simulation results which were obtained from the training and testing algorithm using matlab and verilog.

6.1 Simulation Results of Matlab

6.1.1 Training Results

In this we will discuss about the results of the MFCC algorithms of the extracted feature of the speech in which we are proceeds the result step by step.

Step1:-The original speech of the speaker is containing sample of 10 sec.

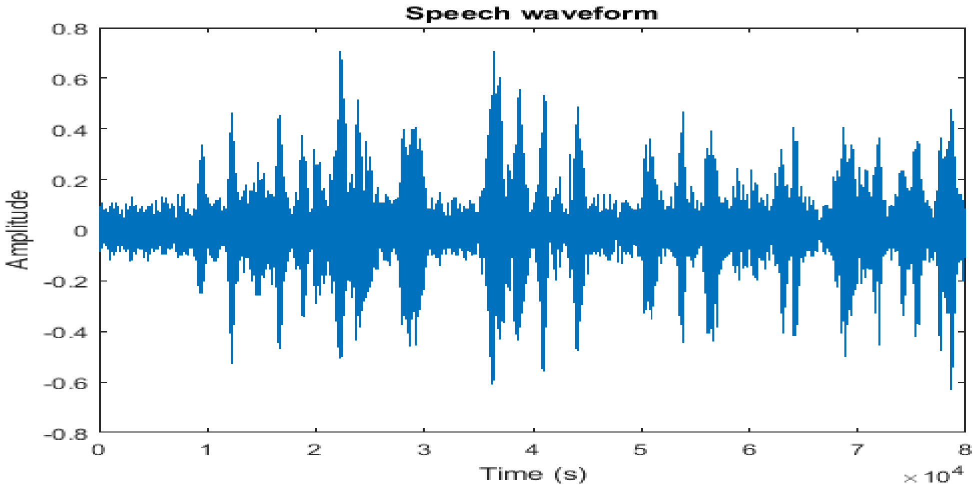


Figure 6.1 Original speech of speaker

Step2:- The Pre-emphasis result of the original speech to improve the high frequency of the speech signal.

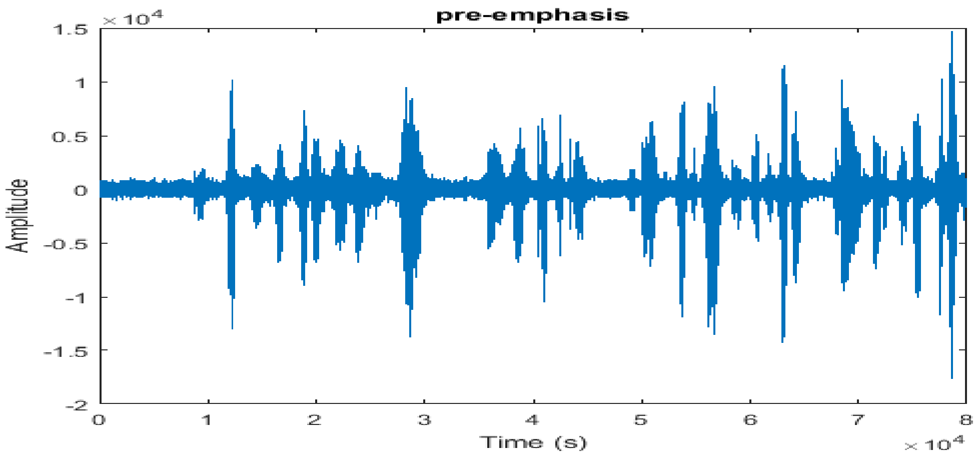


Figure 6.2 Pre-Emphasis speech of speaker

Step3:- Framing the speech signal after the pre-emphasis to make a signal stationary in short time analysis because speech signals behavior is non-stationary. Here we using the one frame time is 32ms and overlapping the frame is 16ms.

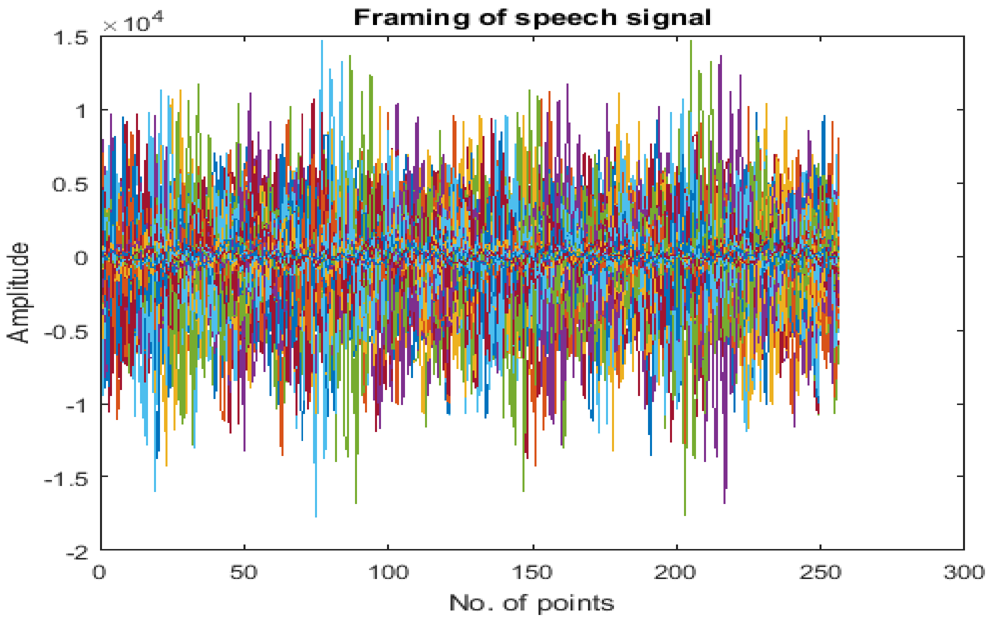
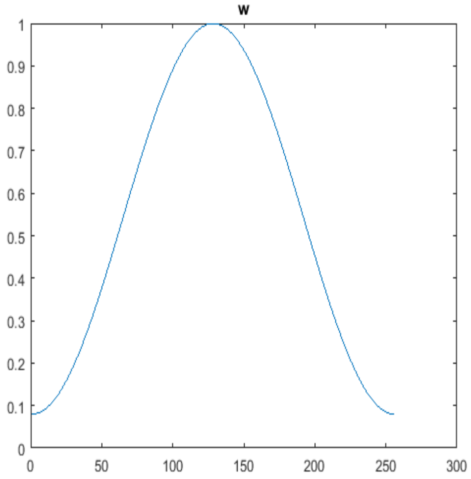


Figure 6.3 Frame the speech signal

Step4:- After Frame the signal we apply hamming window function as shown below and get the resultant framing and windowing function.



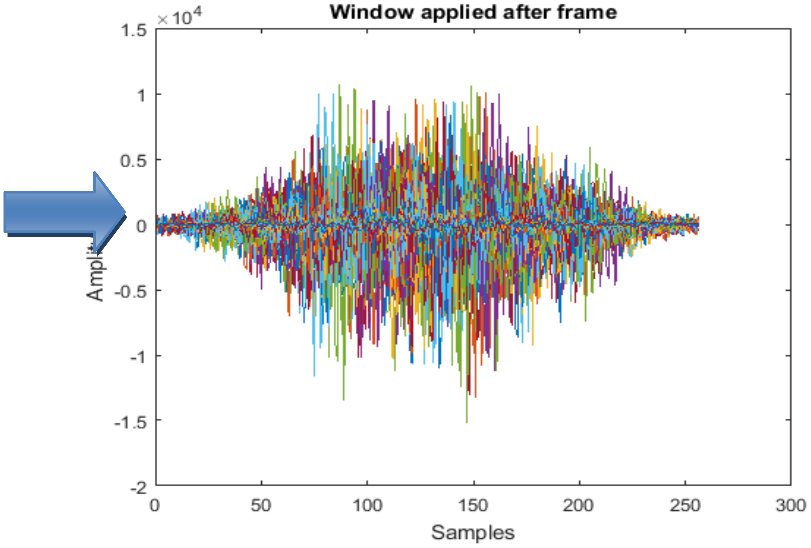


Figure 6.4 Window function

Figure 6.5 Windowing after framing

Step5:-After the FFT of signal we calculate the magnitude of the spectrum of the signal like that-

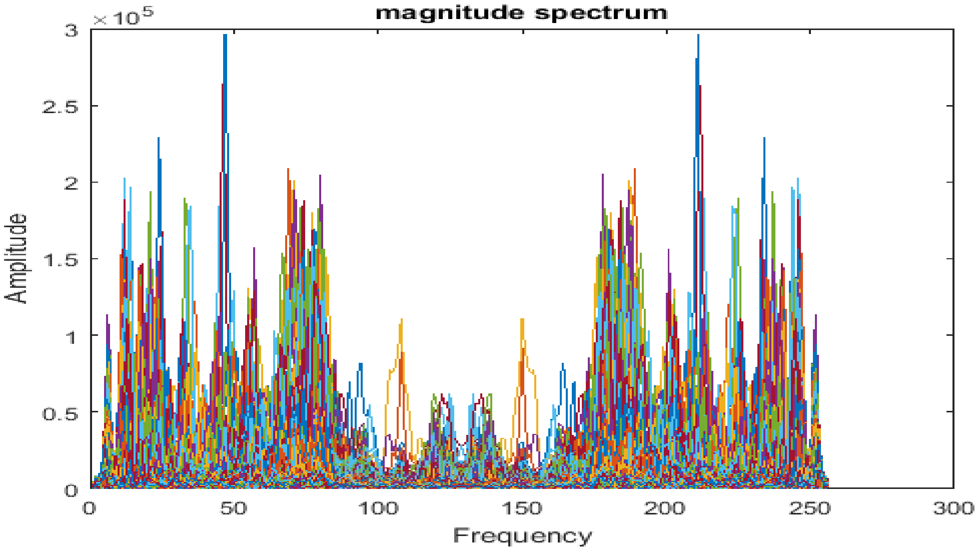


Figure 6.6 Magnitude Spectrum

Step:-6 Magnitude spectrum of the speech signal is applied in to 26 triangular filter bank of the band pass filter and the result we got as shown below-

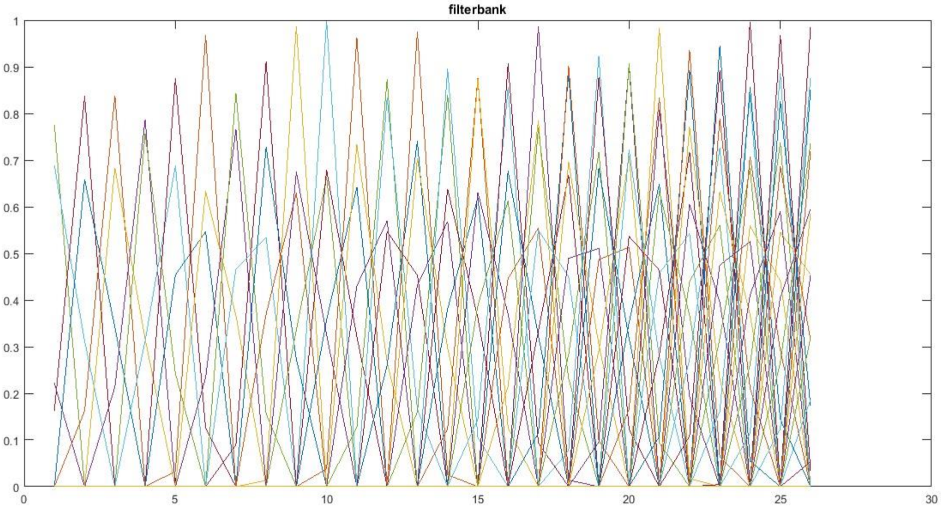


Figure 6.7 Triangular Filter bank

Step7:- After the process of triangular filter bank we multiply with DCT to convert frequency domain to time domain. Then the resultant extracted feature vector of speech is shown below-

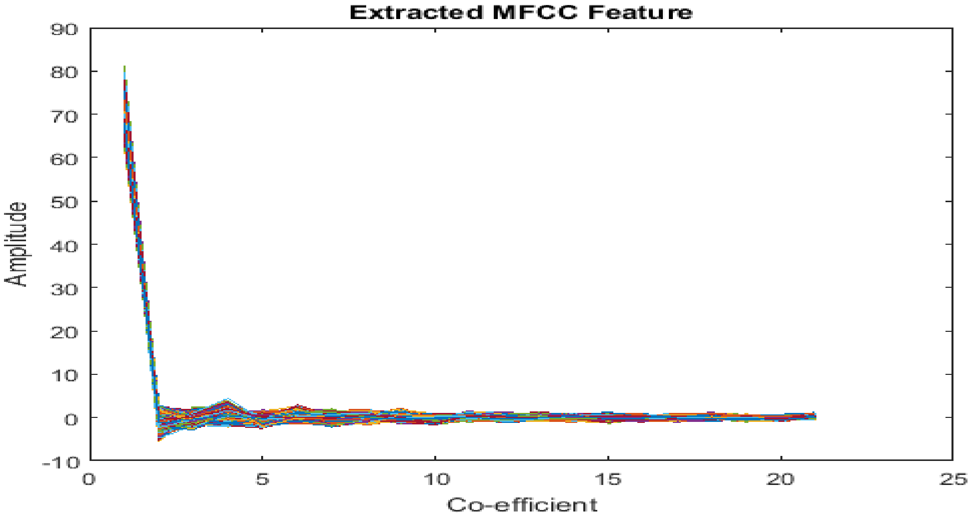


Figure 6.8:- MFCC feature vector

6.1.2 Testing Results

The identity of speaker is given by the maximum log likelihood result of the GMM. The figure show the matched speaker maximum log likelihood function-

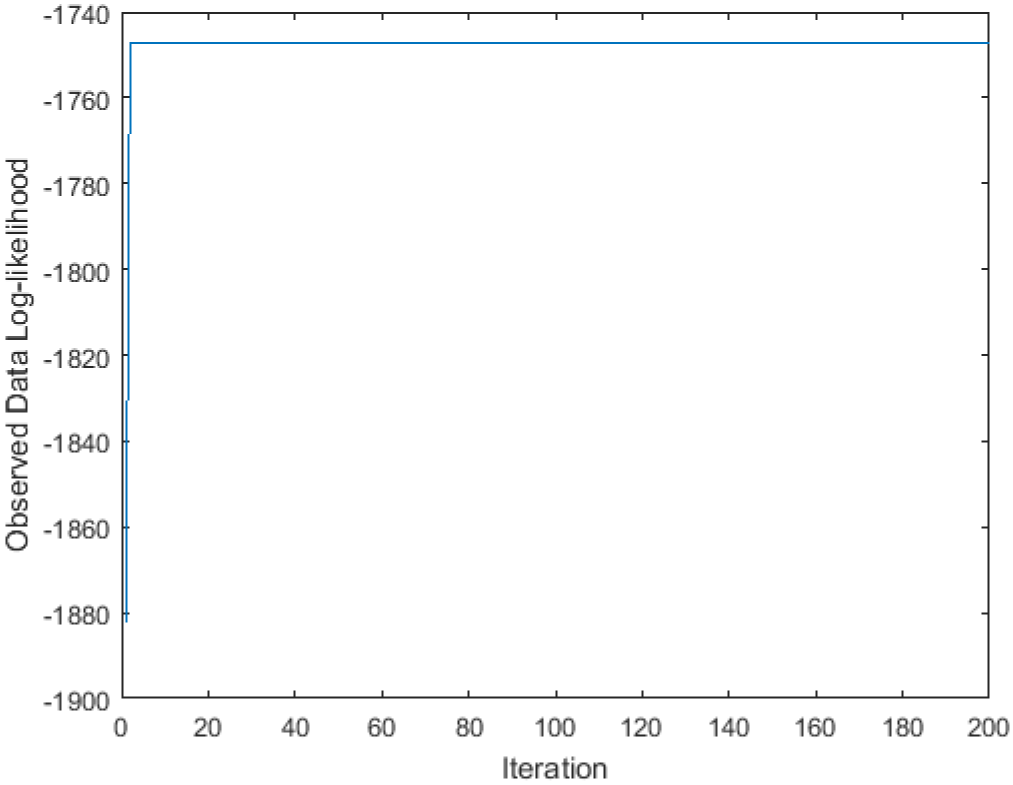
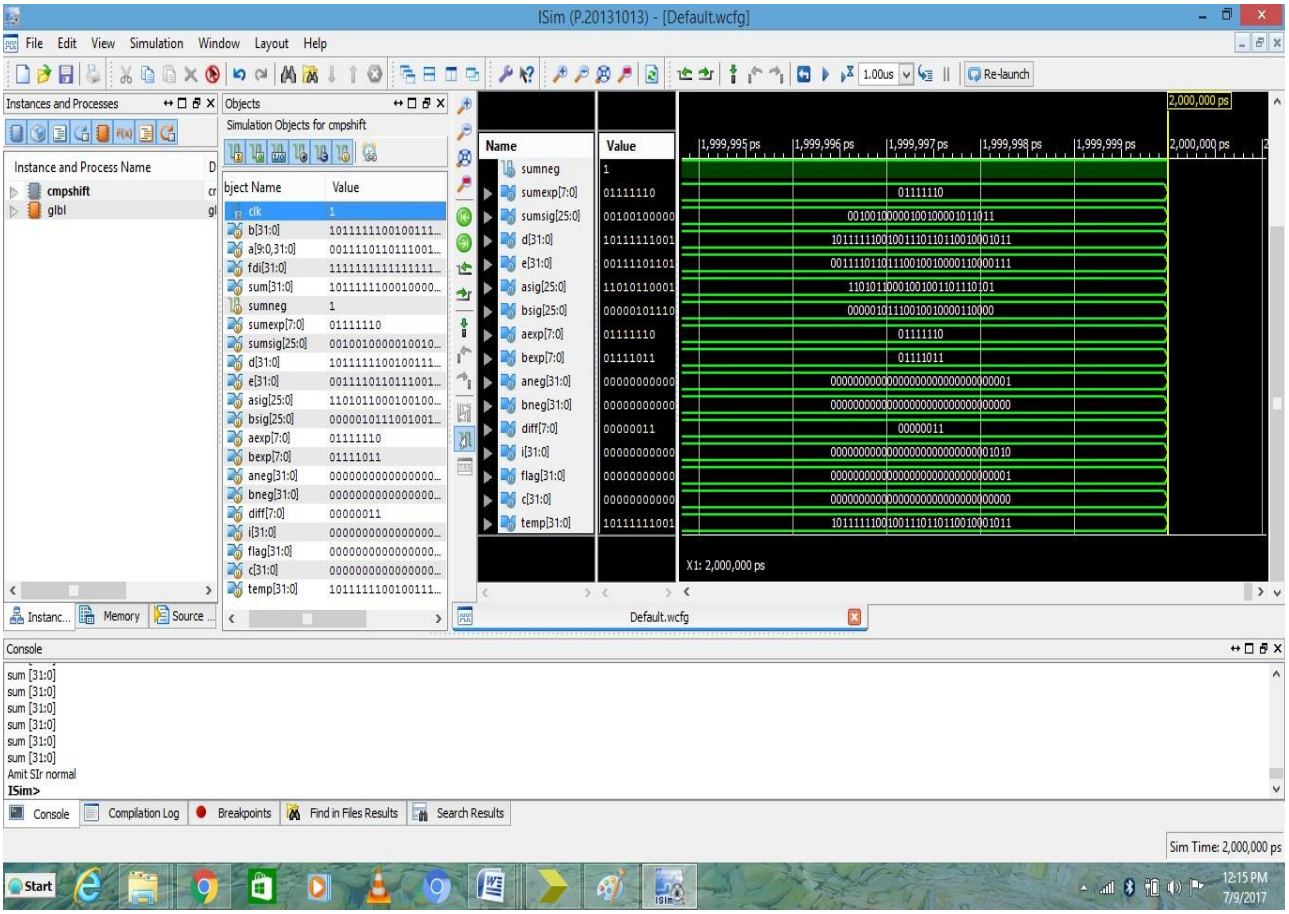


Figure 6.9:- Maximum likelihood function

6.2 Simulation Results of Verilog

In verilog we create five speaker databases in normal and whispering speech. Both speech is represented the 32-bit floating point number. Then calculated the minimum distance of each speaker to recognize the true speaker.



bject Name	Value
clk	1
b[31:0]	1011111100100111...
a[9:0,31:0]	0011110110111001...
fdi[31:0]	1111111111111111...
sum[31:0]	1011111100010000...
sumneg	1
sumexp[7:0]	01111110
sumsig[25:0]	0010010000010010...
d[31:0]	1011111100100111...
e[31:0]	0011110110111001...
asig[25:0]	1101011000100100...
bsig[25:0]	0000010111001001...
aexp[7:0]	01111110
bexp[7:0]	01111011
aneg[31:0]	0000000000000000...
bneg[31:0]	0000000000000000...
diff[7:0]	00000011
i[31:0]	0000000000000000...
flag[31:0]	0000000000000000...
c[31:0]	0000000000000000...
temp[31:0]	1011111100100111...

Name	Value	
sumneg	1	
sumexp[7:0]	01111110	01111110
sumsig[25:0]	00100100000	00100100000100100001011011
d[31:0]	10111111001	10111111001001110110110010001011
e[31:0]	00111101101	00111101101110010010000110000111
asig[25:0]	11010110001	11010110001001001101110101
bsig[25:0]	00000101110	00000101110010010000110000
aexp[7:0]	01111110	01111110
bexp[7:0]	01111011	01111011
aneg[31:0]	00000000000	00000000000000000000000000000001
bneg[31:0]	00000000000	00000000000000000000000000000000
diff[7:0]	00000011	00000011
i[31:0]	00000000000	00000000000000000000000000001010
flag[31:0]	00000000000	00000000000000000000000000000001
c[31:0]	00000000000	00000000000000000000000000000000
temp[31:0]	10111111001	10111111001001110110110010001011

```
sum [31:0]
sum [31:0]
sum [31:0]
sum [31:0]
sum [31:0]
sum [31:0]
Amit SIr normal
ISim>
```

Figure 6.10 Simulation result in Verilog

Natural and whisper speech identification of a person using K-means and GMM algorithm classifiers.

Speech samples of 10 sec are taken in room at noisy condition of both natural and whisper voice with normal microphone of Hp laptop. Speech of 5 persons with 10 different tones including natural and whisper samples are collected. Speeches are preprocessed by sinusoidal lifter filtering to remove noise. Speech database of persons with different tones and mode of all persons are separately created.

Stage1: feature extraction is done by MFCC and features (vectors or matrix) of each person are stored respectively.

Stage2: Distance between extracted train speech sample and extracted test speech sample is calculated. Classification is done based on minimum distance by mean and absolute value of probability of speaker using GMM classifier and K-means classifier.

RESULTS

Training database with Whisper speech and testing with natural and whisper speech using MFCC algorithm

Table:-1

Speaker1	**K-means**		**GMM**	
Samples	Natural	Whisper	Natural	Whisper
Sample1	Y	N	Y	Y
Sample2	Y	N	Y	Y
Sample3	Y	Y	Y	Y
Sample4	Y	Y	Y	Y
Sample5	Y	Y	Y	Y
Sample6	Y	N	Y	Y
Sample7	Y	N	Y	Y
Sample8	Y	N	Y	Y
Sample9	Y	Y	Y	Y
Sample10	Y	Y	Y	Y

Table:-2

Speaker2	**K-means**		**GMM**	
Samples	Natural	Whisper	Natural	Whisper
Sample1	N	N	N	Y
Sample2	N	N	N	Y
Sample3	N	N	N	Y
Sample4	N	Y	N	Y
Sample5	N	N	N	Y
Sample6	Y	Y	N	Y
Sample7	Y	N	N	Y
Sample8	N	Y	N	Y
Sample9	Y	N	N	Y
Sample10	Y	N	N	Y

Table:-3

Speaker3	K-means		GMM	
Samples	Natural	Whisper	Natural	Whisper
Sample1	Y	Y	N	Y
Sample2	Y	Y	N	Y
Sample3	Y	Y	N	Y
Sample4	Y	Y	N	Y
Sample5	Y	Y	N	Y
Sample6	Y	Y	N	Y
Sample7	Y	Y	N	Y
Sample8	Y	Y	N	Y
Sample9	Y	Y	N	Y
Sample10	Y	Y	N	Y

Table:-4

Speaker4	K-means		GMM	
Samples	Natural	Whisper	Natural	Whisper
Sample1	Y	Y	N	N
Sample2	Y	Y	Y	N
Sample3	Y	Y	N	Y
Sample4	Y	Y	Y	N
Sample5	Y	Y	N	N
Sample6	Y	Y	Y	N
Sample7	Y	Y	N	N
Sample8	Y	Y	Y	Y
Sample9	Y	Y	N	Y
Sample10	Y	Y	Y	N

Table:-5

Speaker5	K-means		GMM	
Samples	Natural	Whisper	Natural	Whisper
Sample1	Y	Y	Y	Y
Sample2	Y	Y	N	Y
Sample3	Y	Y	N	Y
Sample4	Y	Y	Y	Y
Sample5	Y	Y	N	Y
Sample6	Y	Y	Y	Y
Sample7	Y	Y	N	Y
Sample8	Y	Y	N	Y
Sample9	Y	Y	N	Y
Sample10	Y	Y	N	Y

Now we are taking the natural speech of all speakers and testing with natural and whisper mode of speakers using MFCC –

Table:-1

Speaker1	**K-means**		**GMM**	
Samples	Natural	Whisper	Natural	Whisper
Sample1	Y	N	Y	Y
Sample2	Y	N	Y	Y
Sample3	Y	Y	Y	Y
Sample4	Y	Y	Y	Y
Sample5	Y	Y	Y	Y
Sample6	Y	N	Y	Y
Sample7	Y	N	Y	Y
Sample8	Y	N	Y	Y
Sample9	Y	Y	Y	Y
Sample10	Y	Y	Y	Y

Table:-2

Speaker2	**K-means**		**GMM**	
Samples	Natural	Whisper	Natural	Whisper
Sample1	N	N	Y	Y
Sample2	N	N	Y	Y
Sample3	N	N	Y	Y
Sample4	N	N	Y	Y
Sample5	N	Y	Y	Y
Sample6	Y	N	Y	Y
Sample7	Y	Y	Y	Y
Sample8	N	N	Y	Y
Sample9	Y	N	Y	Y
Sample10	Y	Y	Y	Y

Table:-3

Speaker3	**K-means**		**GMM**	
Samples	Natural	Whisper	Natural	Whisper
Sample1	Y	N	Y	N
Sample2	Y	N	Y	Y
Sample3	N	N	Y	Y
Sample4	Y	N	Y	Y
Sample5	Y	N	Y	Y

Sample6	Y	N	Y	Y
Sample7	N	N	Y	Y
Sample8	N	N	Y	Y
Sample9	Y	N	Y	Y
Sample10	Y	N	Y	Y

Table:-4

Speaker4	**K-means**		**GMM**	
Samples	Natural	Whisper	Natural	Whisper
Sample1	Y	N	Y	N
Sample2	Y	N	Y	N
Sample3	Y	N	Y	N
Sample4	Y	N	Y	N
Sample5	Y	N	Y	N
Sample6	Y	N	Y	N
Sample7	Y	N	Y	N
Sample8	Y	N	Y	N
Sample9	Y	N	Y	N
Sample10	Y	N	Y	N

Table:-5

Speaker5	**K-means**		**GMM**	
Samples	Natural	Whisper	Natural	Whisper
Sample1	Y	N	Y	Y
Sample2	Y	N	Y	Y
Sample3	Y	Y	Y	Y
Sample4	Y	N	Y	Y
Sample5	N	Y	Y	Y
Sample6	N	Y	Y	Y
Sample7	Y	N	Y	Y
Sample8	Y	Y	Y	Y
Sample9	Y	N	Y	N
Sample10	Y	Y	Y	N

Now we take training as a normal speech and testing with normal and whisper speech using EFCC algorithm for feature extraction.

Table:-1

Speaker1	**K-means**		**GMM**	
Samples	Natural	Whisper	Natural	Whisper
Sample1	Y	N	Y	Y
Sample2	Y	N	Y	Y
Sample3	Y	N	Y	Y
Sample4	Y	N	Y	N
Sample5	N	Y	Y	Y
Sample6	N	Y	Y	Y
Sample7	N	N	Y	Y
Sample8	N	N	Y	N
Sample9	Y	N	Y	Y
Sample10	Y	N	Y	N

Table:-2

Speaker2	**K-means**		**GMM**	
Samples	Natural	Whisper	Natural	Whisper
Sample1	Y	N	Y	Y
Sample2	Y	N	Y	Y
Sample3	Y	N	Y	Y
Sample4	N	N	Y	Y
Sample5	N	N	Y	Y
Sample6	Y	N	Y	Y
Sample7	N	N	Y	Y
Sample8	N	N	Y	Y
Sample9	Y	N	Y	Y
Sample10	Y	N	Y	Y

Table:-3

Speaker3	**K-means**		**GMM**	
Samples	Natural	Whisper	Natural	Whisper
Sample1	Y	Y	N	N
Sample2	Y	Y	N	N
Sample3	Y	N	N	N
Sample4	N	N	N	N
Sample5	N	Y	Y	Y
Sample6	N	Y	Y	N
Sample7	N	N	N	Y
Sample8	Y	N	Y	Y
Sample9	Y	N	Y	N

Sample10	Y	Y	Y	N

Table:-4

Speaker4	**K-means**		**GMM**	
Samples	Natural	Whisper	Natural	Whisper
Sample1	Y	N	Y	N
Sample2	Y	N	Y	N
Sample3	Y	Y	Y	N
Sample4	Y	Y	Y	N
Sample5	N	Y	Y	N
Sample6	N	Y	Y	Y
Sample7	Y	N	Y	Y
Sample8	Y	Y	Y	Y
Sample9	Y	N	Y	N
Sample10	Y	Y	N	N

Table:-5

Speaker5	**K-means**		**GMM**	
Samples	Natural	Whisper	Natural	Whisper
Sample1	N	N	Y	N
Sample2	N	N	Y	N
Sample3	Y	N	Y	N
Sample4	Y	N	Y	N
Sample5	Y	N	Y	N
Sample6	Y	N	Y	N
Sample7	Y	N	Y	N
Sample8	Y	N	Y	N
Sample9	Y	N	Y	N
Sample10	Y	N	Y	N

Now we take training as a whisper speech and testing with normal and whisper speech using EFCC algorithm for feature extraction.

Table:-1

Speaker1	**K-means**		**GMM**	
Samples	Natural	Whisper	Natural	Whisper
Sample1	Y	N	Y	Y
Sample2	Y	N	Y	Y
Sample3	Y	Y	Y	Y
Sample4	Y	N	Y	Y

Sample5	N	Y	Y	Y
Sample6	N	Y	Y	Y
Sample7	Y	N	Y	Y
Sample8	Y	Y	Y	Y
Sample9	Y	N	N	Y
Sample10	Y	Y	N	Y

Table:-2

Speaker2	**K-means**		**GMM**	
Samples	Natural	Whisper	Natural	Whisper
Sample1	Y	N	N	N
Sample2	Y	N	N	Y
Sample3	Y	Y	N	Y
Sample4	Y	N	N	Y
Sample5	N	Y	N	Y
Sample6	N	Y	N	Y
Sample7	Y	N	N	Y
Sample8	Y	Y	N	Y
Sample9	Y	N	N	N
Sample10	Y	Y	N	N

Table:-3

Speaker3	**K-means**		**GMM**	
Samples	Natural	Whisper	Natural	Whisper
Sample1	Y	N	N	Y
Sample2	Y	N	N	Y
Sample3	Y	Y	N	Y
Sample4	Y	N	N	Y
Sample5	N	Y	N	Y
Sample6	N	Y	N	Y
Sample7	Y	N	N	Y
Sample8	Y	Y	N	Y
Sample9	Y	N	N	N
Sample10	Y	Y	N	N

Table:-4

Speaker4	**K-means**		**GMM**	
Samples	Natural	Whisper	Natural	Whisper
Sample1	Y	N	N	N
Sample2	Y	N	N	N

Sample3	Y	Y	N	Y
Sample4	Y	N	Y	Y
Sample5	N	Y	N	N
Sample6	N	Y	N	N
Sample7	Y	N	N	N
Sample8	Y	Y	N	Y
Sample9	Y	N	N	N
Sample10	Y	Y	N	N

Table:-5

Speaker5	K-means		GMM	
Samples	Natural	Whisper	Natural	Whisper
Sample1	Y	N	N	Y
Sample2	Y	N	N	Y
Sample3	Y	Y	N	Y
Sample4	Y	N	N	Y
Sample5	N	Y	Y	Y
Sample6	N	Y	N	Y
Sample7	Y	N	Y	Y
Sample8	Y	Y	N	Y
Sample9	Y	N	N	Y
Sample10	Y	Y	N	N

From above all the table analysis of all speaker in different classifier and both training as a natural and whisper and testing with all natural to natural, whisper to whisper, natural to whisper and whisper to natural gives the overall accuracy as shown below.

This Table shows the overall accuracy of all speakers in all speech modes using MFCC algorithm

Speech Mode		**Accuracy (%)**	
Training	**Testing**	**GMM**	**K-means**
Natural	Natural	98%	74%
Natural	Whisper	74%	22%
Whisper	Whisper	84%	60%
Whisper	Natural	36%	10%

This Table shows the overall accuracy of all speakers in all speech modes using EFCC algorithm

Speech Mode		**Accuracy (%)**	
Training	**Testing**	**GMM**	**K-means**
Natural	Natural	80%	62%
Natural	Whisper	70%	20%
Whisper	Whisper	76%	16%
Whisper	Natural	24%	8%

Chapter 7
Conclusion and Future Work

From the above recognition accuracy result it is concluded that GMM algorithm is best for speech identification which provides accuracy in speech identification. Highest recognition accuracy of 98**%** for speech recognition in neutral to neutral and **74%** for whisper to whisper recognition is obtained with GMM classifier algorithm and MFCC algorithm. People normally lower their voice to overhear in public places so they normally speak in whisper voice. We proposed a method to improve the identity of person in whispering mode. From the above analysis we saw that identify a person in whispering mode is very difficult task. The spectral frequency for both neutral and whisper mode is scale on exponential filter bank. By looking the spectrum of the whisper and neutral voice we saw that both speech modes have a similarity in high frequency range. So we perform the exponential scale on below 1000Hz frequency range.

In future, we will work on improving the accuracy of the whisper speech which is very important fact nowadays scenario and execution time of the speech identification process. In the proposed algorithm problem is that the accuracy normal to whisper and whisper to normal is very less and if we increase the number of person it takes more execution time for the feature extraction time on training phase which depends on the time of the speech samples means as the time of the speech samples and number of person is increases then execution time will also increases. So in future, we will work on the execution time of the feature extraction of speech and most important is whisper analysis of speech so that it will take less time to execute the algorithm.

References

[1] Bageshree V. Sathe-Pathak, Ashish R. Panat, ―Extraction of Pitch and Formants and its Analysis to identify 3 different emotional states of a person‖, International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012

[2] Premakanthan P. and Mikhad W. B., ―Speaker Verification/Recognition and the Important of Selective Feature Extraction: Review‖, MWSCAS. Vol. 1, 57-61, 2001

[3] B. S. Atal, ―Automatic Recognition of Speakers from their Voices‖, Proceedings of the IEEE, vol.64,1976, pp 460 – 475.

[4] D.A. Reynolds, ―Experimental evaluation of features for robust speaker identification‖, IEEE Trans. Speech Audio Process. , vol.2(4), pp. 639-43, Oct. 1994.

[5] JOURNAL OF COMPUTING, VOLUME 2, ISSUE 3, MARCH 2010, ISSN 2151-9617 HTTPS://SITES.GOOGLE.COM/SITE/JOURNALOFCOMPUTING/

[6] F.Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin Chagnolleau, S. Meignier, T.Merlin J. Ortega-Garcia, D. Petrovska Delacreetaz, and D.A. Reynolds, "*A tutorial on text-independent speaker verification*," EURASIP Journal on Signal Processing, Hindawi Publishing Cor-poration, vol. 4, pp. 430–451, 2004

[7] http://mirlab.org/jang/books/audiosignalprocessing/speechFeatureMfcc.asp?title=122%

20MFCC

[8] Md. Rashidul Hasan, Mustafa Jamil, "Speaker Identification Using Mel Frequency Cepstral Coefficients", 3rd International Conference onElectrical & Computer Engineering, ICECE 2004, 28-30 December2004.

[9] J. R. Deller, Jr., J. G. Proakis, J. H. Hansen. (1993). Discrete-Time Processing of Speech Signals. New York: Macmillan.

[10] D., Jurafsky and J.H., Martin, "*Speech and Language Processing an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*".

[11] S. E. Bou-Ghazale and J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Trans Speech Audio Process.*, vol. 8, no. 4, pp. 429–442, Jul. 2000. Prentice Hall, Upper Saddle River, NJ, USA, 2000.

[12] G. Zhou J. H. L. Hansen J. F. Kaiser "Nonlinear feature based classification of speech under stress" <em>IEEE Transaction on Speech and Audio Processing</em> vol. 9 no. 3 pp. 201-216 Mar 2001.

[13] X. Fan and J. H. L. Hansen, "Speaker identification for whispered speech based on frequency warping and score competition," in *Proc. ISCA Interspeech*, 2008, pp. 1313–1316.

[14] https://en.wikipedia.org/wiki/Speech

[15] http://mirlab.org/jang/books/audiosignalprocessing/example/output/preEmphasis01.png

[16] http://mirlab.org/jang/books/audiosignalprocessing/example/output/hammingWindow01.p

[17] X. Li, "Reconstruction of Speech from Chinese Whispers, [PhD], Nanjing University, China, 2004.

[18] T. Ito, K. Takeda and F. Itakura, "Analysis and Recognition of Whispered Speech," Speech Communication 45, pp.139-152, 2005

[19] Jin, S. S. Jou and T. Schultz, "Whispering Speaker Identification, "IEEE International Conference on Multimedia and Expo, 2007.

[20] M. Matsuda and H. Kasuya, "Acoustic Nature of the Whisper,"EUROSPEECH,

pp.137-140, 199

[21] C. Zhang and J. H. L. Hansen, “Analysis and Classification of Speech Mode: Whisper through Shouted,” INTERSPEECH2007

[22] X. Fan and J. H. L. Hansen, “Speaker identification for whispered speech based on frequency warping and score competition,” INTERSPEECH 2008