Enhanced MFCC Algorithm using Lookup Table and Kaiser Window

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering
in
Software Engineering

Submitted By
VIKRAM OJHA
(Roll No. 801231030)

Under the supervision of:
Mr. RAJ KUMAR TEKCHANDANI
Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004

June 2014

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*Enhanced MFCC Algorithm Using Lookup table and Kaiser Window*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Software Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Raj Kumar Tekchandani* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Vikram Ojha
ME (Software Engineering)
801231030

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Raj Kumar Tekchandani
Assistant Professor
Computer Science and Engineering Department

Countersigned by

(Dr. Deepak Garg)
Head
Computer Science and Engineering Department
Thapar University
Patiala

(Dr. S. K. Mohapatra)
Dean (Academic Affairs)
Thapar University
Patiala

ACKNOWLEDGEMENT

Here, I would like to thank everyone who helped with this project my guide *Mr. Raj Kumar Tekchandani* and *Dr. Ravinder Kumar* and the whole CSED Department of Thapar University.

Here, I will also like to thank those who had introduced this topic at my graduation *Dr. Sapan Naik*, Asst Professor at NIT-Surat *Dr. Himanshu Rana* HOD Computer Dept. at MGITER-Navsari and *Dr. Mukesh Patel* HOD Computer Department at SCET-Surat. These were the people who had brought my interest in domain of Artificial Intelligence and special thanks to *Dr. Mahesh Goyani* Asst Professor at L.D college of Engineering –Ahmadabad.

I would take pleasure to pay my heartiest gratitude to my father Mr. S. B. Ojha, my mother Mrs. Vinita Ojha.

Vikram Ojha

ABSTRACT

Automatic Speaker Recognition System has been quite fascinating for a man from quite long time. There are various feature extraction algorithm like Linear Predictive Coefficient (LPC), Mel- frequency Cepstrum Coefficient (MFCC). In this thesis, enhanced MFCC algorithm has been proposed which reduces the total time by almost 50 percent but accuracy decreases as compared to conventional algorithm from 94 percent to 94.93 percent. But this makes new algorithm to be implemented more effectively on hardware.

The proposed algorithm tries to explore in security where speaker recognition can be used for speaker identification. This algorithm uses lookup table and Kaiser Window instead of Hamming Window algorithm as in conventional MFCC algorithm which improves the accuracy of algorithm and lookup table which reduces the time complexity of algorithm also the formula for pre-emphasis has been modified which again reduces the copulation time for pre-emphasis of signal.

TABLE OF CONTENT

LIST OF FIGURES

List of Tables

Chapter 1
Introduction

From quite a long time, it has been fascinating for a man to make machines talk and listen. This interest stems from as far back as ancient Greek and Roman civilizations. Stanley Kubrick's famous movie *A Space Odyssey* and the robot R2D2 in George Lucas *Star Wars* series of movies shows the intense interest of human machine communication. The book by Dr. Bill Gates (founder of Microsoft), *"The Road Ahead*" calls Automatic Speaker Recognition (ASR) System as one of the most innovative research for future computer systems. However even with great effort and research in this field still is a far from reaching the great result in such field. The great motivation for researchers in this field has come from the development of *Siri* currently comes with all new versions of iPhone developed by two MIT graduates and lot of research is going in Google, one good example of it can be seen in Google search engine which has provided a feature called voice search which interprets the word or a complete sentence you speak and gives you the desired result but still 100% accuracy has not been achieved as this depends on many features like microphone used, background noise, accent of a person, mood of a person and many other features and due to these feature with interdisciplinary nature it is quite difficult to perform research in the area of speaker or speech recognition system and tendency of most researchers is to apply monolithic nature.

1.1 Speaker Recognition System

The ability of recognizing a person from his voice is known as speaker recognition. From technical point of view there are two types of ASR system direct voice input (DVI) and Large vocabulary continuous speech recognition (LVCSR) in both the systems the underlying technology is more or less the same with a little difference that DVI is used for command control system where as LVCSR systems are to fill forms or voice based document creation. DVI systems are typically confined to small vocabulary, typically thousands of words, and are usually assumed to respond immediately where as LVCSR systems contains hundreds of thousands of words and are supposed to understand the continuous speech. The general model for speaker or speech recognition system is shown in figure 1.1. This model as shown initially asks

user/speaker for his/her voice as input. From spoken signals, first words are recognized that are meaningful according to syntax, semantic and pragmatic of recognition task. The meaning of the recognized word is then sent to higher level processing block where it uses dynamic knowledge representation to modify the syntax, semantic and pragmatic in context in which has previously recognized the word [1].

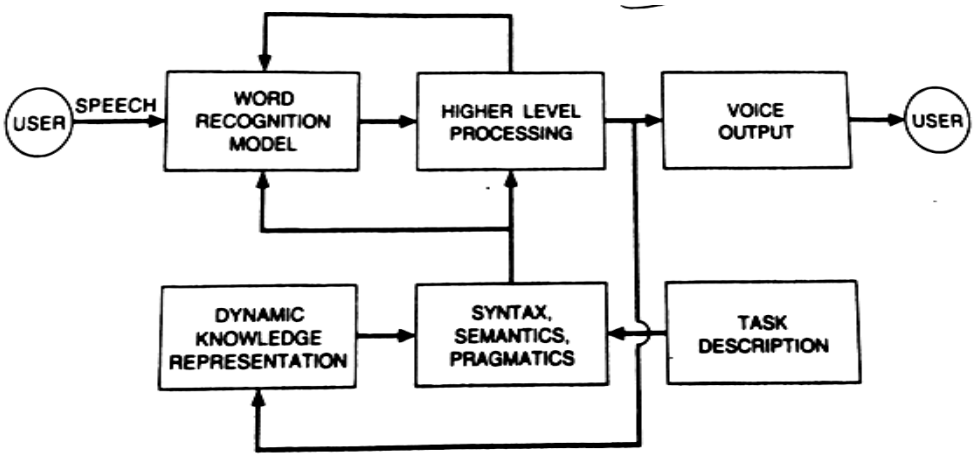

Fig 1.1: General Model for Speaker Recognition System [1]

In higher level processing phase, non-sequiturs are usually removed, like emotions, usually at risk of misunderstanding, but this reduces the error. The feedback from this higher-level-processing box reduces the complexity of recognition model by limiting the search for valid input sequence, speech, from the speaker/user. The recognition system responds to the user in the form of required output, it can be voice output or opening some application.

The Automatic Speaker Recognition System (ASR) has been the foremost interest of computer scientists since the advent of computer or even before the physical existence of computer. Initially it was started with Speech recognition System where Text to speech and Speech to Text was the main interest of these scientists. In earlier times devices were build which depicted the human vocal tract for speech production like musical instruments which were evident to these scientists [9]. However a device was needed to understand the word spoken by speaker as accurate as a human ears and give complex decision about gender, emotions and the most important is to recognize the word spoken by speaker. During the past twenty years there has been increasing

interest in identification of person through their speech or voice. But the speaker recognition systems have number of variation due to large number of application in upcoming areas. The application of greatest commercial interest seems to be confirming the identity of person carrying the business transaction, especially in cases where other means of identification are unavailable or are not appropriate. An example is business carried out on telephone like changing ATM pin card number through cell phone or landlines.

i. **Entry Control**

   A great improvement in entry control has been achieved in the past twenty through secret numbers and badges but they all are vulnerable because they all are based upon possession of artifact and thus vulnerable. Thus voice of a person can be used to detect his/her identity.

ii. **Business Transaction**

   It has now become more secure to carry transaction on telephone as the persons' identity can now be judged through their voice. Therefore more secure system is needed to prevent from any kind of frauds.

iii. **Securing Data**

   Today many systems are available in market with voice recognition software to prevent your data from intruders. As compared to traditional user id and password systems are not much secure as there are many techniques in window XP & others to hack password.

1.2 Motivation

Research in automatic speech recognition system has been going on from decades. So it is worthwhile to do research in this area. Many speaker recognition systems are in use today like in entry control, business transaction, data security. In today's speaker recognition systems Linear Predictive Coding (LPC) parameters are used to extract features with which neural network is trained. Then it is fed with a pattern where it extracts the LPC parameter of the given speech and compares it with the patterns in its database and computes the output. The early work on speaker recognition was totally limited to human listening. A part of this research gave rise to vocoders which are machine to synthesize speech. Although the synthesized speech is sufficiently intelligible, it is often deficient with respect to speaker

recognizibility [1]. This problem enhanced interest in search for those than the utterance itself. Although with varying degree of accuracy, machines have been made to recognize speakers, the fundamental question of "how human recognize different speakers?"[8] has remains unsolved and the above mentioned parameters are not fully understood. Today both text independent and text dependent speech recognition systems are present. Text dependent speaker recognition are those in which speaker has to utter the particular word for his/her identification. While in text independent speaker recognition system speaker can utter any word or sentence from which system identifies the identity of the person. Today's speaker recognition systems are more reliable than their predecessors as they are more adaptable to their environment, as many algorithms like end-point detection algorithms which detects the start and end of speaker voice. Also algorithms filter the input voice and extract only the required features from the speech of a person. It extracts only the voiced speech and leaves the unvoiced part or white noise present in speech. In short, today's speech recognition systems have become more intelligent and reliable. The only limitations in today's algorithms are that as the speaker or person grows up, the vocal tract of a person changes, thus there is a need to update the voice of speaker every after 6 months or a year. As mentioned earlier, the vocal tract of individual changes with time and system fails to recognize the speaker identity. One aspect of the use of computer for speaker recognition is very interesting; while in many areas of research machine could not duplicate the accuracy exhibited by human, in this particular area of speaker recognition, machine has surpassed human performance.

1.3 About the Thesis

In this thesis, the attempt was in-depth research about the Automatic Speaker Recognition Technique (ASR) which is currently used today in all real time application for security purposes and Entry control systems. Here mainly focus is on improving the time and cost complexity of computation of an Mel-frequency Cepstrum Coefficient (MFCC) algorithm. So the rationale was to improve Mel-frequency Cepstrum Coefficient (MFCC) to reduce the number of multiplication and replace these multiplications by addition which is less costly compared to multiplication and also use different window algorithm to improve the efficiency. This code for this is implemented using MATLAB.

1.4 Thesis Outline

The remaining chapters are summarized in following division.

Chapter 2: is literature Survey, this Chapter introduces different algorithms for extracting features from speech, brief history how it all started and till where it has reached.

Chapter 3: is problem definition. This chapter explains the problem statement.

Chapter 4: is a proposed solution, this chapter explains the proposed solution and proposed solution reduces time and cost of MFCC algorithm

Chapter 5: is an experimental result and discussion. This chapter explains the problem to the solution practically and analyses the result.

Chapter 6: is conclusion and future scope. This chapter explains conclusion of the complete report and describes future scope for the further experiments.

Chapter 2
Literature Survey

2.1 A Brief History of Speaker-Recognition System

Research in automatic speech recognition system has been going on from decades. So it is worthwhile to discuss some of the research highlights briefly. Speaker recognition technique has been relatively new subject extensive research has been carried out during the last two decades due to its promising usefulness specially in the fields of business and criminology. As a result a considerable amount of literature on this topic is available. As it has been pointed out earlier, the problem of speaker recognition has a number of variation factors (such as noisy or noise free case, test speaker's cooperation etc.) and the available literature can be divided into a corresponding number of categories.

The earliest attempt in Automatic recognition system was made in 1950s at Bell Laboratories, Davis, et al. [2] built a system for isolated digit recognition system for a single speaker [2]. This system was heavily depended on measuring spectral resonance during vowel region of each speech. In 1956 Olson and Belar in RCA laboratories tried to recognize 10 different syllables of single speaker this system also relied on spectral density of each vowel [3]. In 1959 Fry and Denes tried to build recognition which could recognize four vowels and nine consonants at University College of England [4]. Another major effort in this field during this period was made by Frogie and Frogie, constructed at MIT Lincoln Laboratories again a filter bank spectral information was used and time varying estimate of the vocal tract resonances was made to decide the vowel spoken.

In the 1960s, several fundamental ideas in speech recognition was published. This was the era in which several Japanese laboratories jumped into arena of speaker recognition system, described by Suzuki and Nakata of the Radio Research lab in Tokyo [8]] was a hardware vowel recognizer. In the same year three key research projects were initiated that had major impact on the research and development of speaker recognition system (SR). The first of these projects was efforts of Martin and his colleagues at RCA laboratories. A major achievement in Speaker Recognition

System in 1960s was by the pioneering research of Reddy which later spawned a long and highly successful speech recognition program at Carnegie Mellon University.

Another successful step was taken in 1970s by researchers of IBM. Finally at, AT&T Bell Labs, researchers began a series of experiments aimed at making a successful speaker independent system [6]. In 1980s the focus from isolated word recognition shifted to continuous word recognition. Speech research in 1980s shifted to new technology from template based approaches to statistical modeling methods – especially the HMM (Hidden Markov Model) approach [7][8]. Another new technology which was applied in 1980s was Neural Network to train Automatic Speaker Recognition System.

Finally in 1980s major focus was on continuous speech interpretation and large vocabulary speaker recognition system by Defense Advanced Research Projects Agency (DARPA) community, which sponsored a large research program. The DARPA program continued in 1990s with emphasis shifting from natural language front ends to the recognizer, and the task shifting to retrieval of air travel information.

2.2 The Process of Speech Production and Perception in Human Being

The figure 2.1 shows the process of speech production and speech perception in human being. The production process begins when someone (speaker) formulates a sentence in his mind that he wants to speak to the listener via speech. This is same as a text expressing the words in a machine. The next process in the communication is the conversion of speech which human wants to transmit into the language code. This roughly corresponds to the phoneme sequence representing the words of a message and prosody makers donating the sound pitch, loudness of a word. Once the language is chosen neuromuscular commands are given to the vocal cords to vibrate accordingly and to shape the vocal cords so that proper sound is created.

Once the speech signal is generated and propagated to the listener, the speech perception (or speech recognition) process begins. First the listener processes the acoustic signal along the basilar membrane which provides the analysis of incoming signal then the activity along the neuron is converted to the language code.

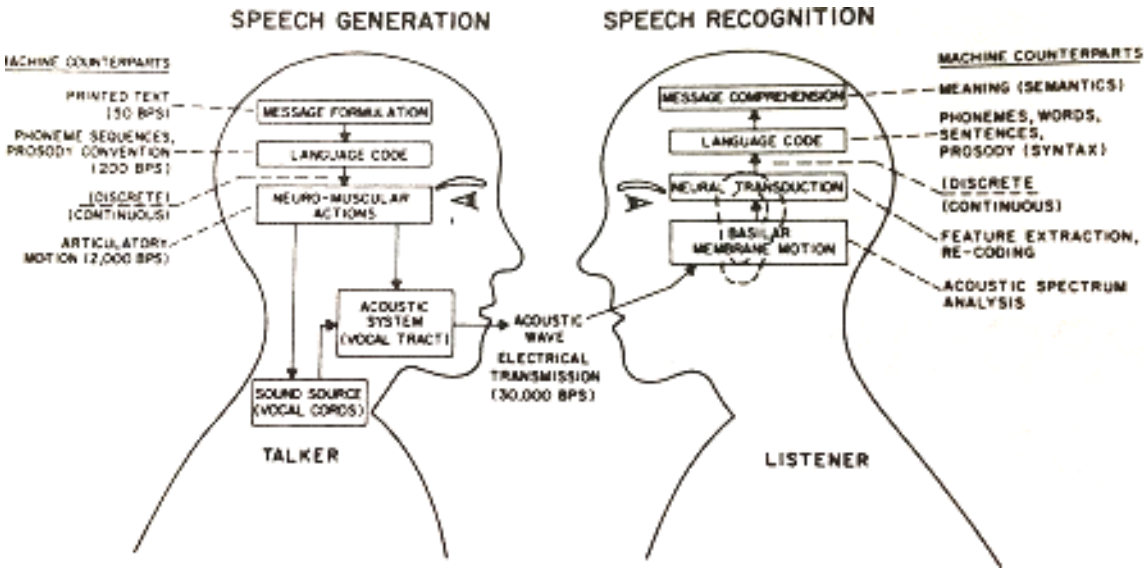

Fig 2.1: The Process of Speech Production and Speech Perception [3]

2.3 Parameters in Speaker Recognition

Efforts have been made to search for the parameters responsible for speaker recognition by human listeners [4], but the psycho-acoustic dimensions of clarity, roughness, magnitude, and animation offer very little hints of physical correlates [1]. Another major drawback in using voice prints is that an expert is needed for their use as there is the problem with finger prints. Ideally, an effective speaker recognition algorithm should measure some aspect of speech that reflects the unique properties of the speaker's vocal system and contain no information about the speech or word spoken itself. A group or an individual can be identified based upon their speech or voice because all of us have got different voice. This voice can be classified based on different parameters like:

i. **Pitch:** The sensation of frequencies is commonly referred to as the ***pitch*** of a sound. A high pitch sound corresponds to a high frequency sound wave and a low pitch sound corresponds to a low frequency sound wave. [6]
ii. **Tone:** Tone refers to the ascent of any individual. As seen people from all over the world have got different pronunciation for each and every word, like American or British accents. [6]
iii. **Rate:** Rate is defined as the speed with which the speaker pronounces the particular word [6].

Apart from these parameters in speech some other parameters are also present in speech like Linear predictive coefficient (LPC) parameter, Mel frequency cepstral coefficient (MFCC), Parcor coefficient, Log area coefficient which can be extracted directly from speech data or after converting speech signal to spectrogram. These parameters are grouped into two parts:

i. The parameters which are used for speaker recognition.
ii. The parameters which are used for speech processing.

This gives rise to one of the two most important steps towards successful speaker recognition, that is, the selection of those parameters that efficiently represent the speaker dependent information in speech. Procedures for the selection can be found in [1][5]. An ideal set of recognition parameters should exhibit at least the following characteristic [1][3].

i. High efficiency in representing speaker dependent information.
ii. Easy to measure.
iii. Stable over time.
iv. Frequent and natural occurrence in speech.
v. Insensitivity to speaking environment.
vi. Insusceptible to mimicry.

Recognition error rate is the key parameter in adequately describing or specifying the performance of system. Other factors responsible for the user acceptance are total recognition transaction time and overall system performance. Recognition errors may be categorized as Type I and Type II. A type I error refers to the situation where a claimed identity is rejected when in fact the claim was true. A Type II error is the acceptance of the claimed identity when in fact the identity claim was false. Recognition errors may occur due to the following three reasons:

i. Variation in speech by the same speaker.
ii. Similarities in speech between speakers.
iii. External influences like problem in acoustic measurement comparison and recording environment.

Speaker recognition system determines the identity of a speaker on the basis of speech

signals or voice signals. Speaker recognition systems are usually divided into two parts: speaker identification and speaker verification [5]. In Speaker identification, it matches the closest identity of the speaker based on the pool of speakers data saved after extracting the voice features. And background noise or DC noise is removed from the speaker and padded with zero to maintain the length [7]. Speaker verification aims to identify whether the claimed speaker is genuine or not. Like other pattern recognitions problems, speaker verification system has both testing and training phase.

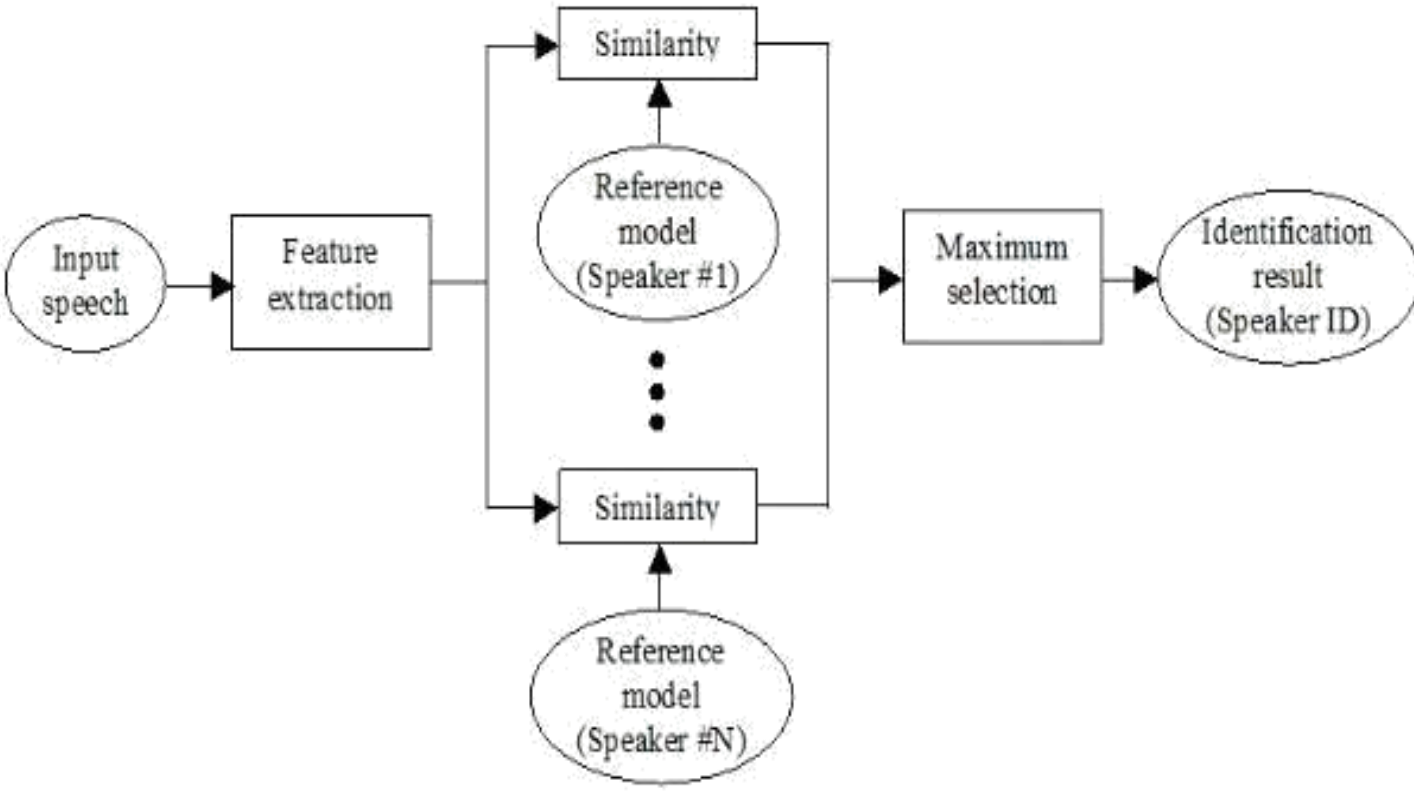

Fig 2.2: Speaker Identification [3]

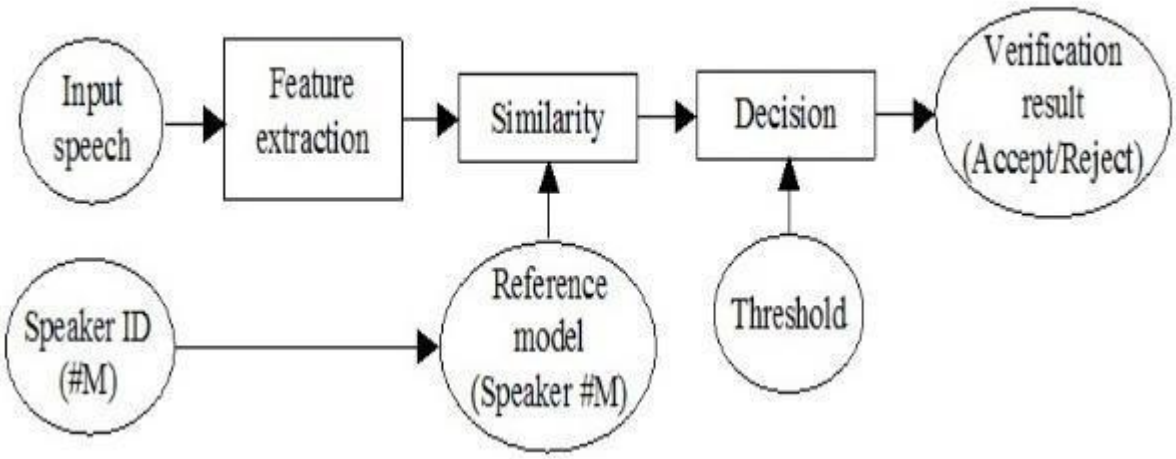

Fig 2.3: Speaker Verification [3]

In testing phase system is tested with some given set of data bits different from training phase. In training phase Speaker Recognition System (SRS) are trained with some set of speaker's voices after filtering noise from them like humming sound and DC offset, which normally occurs due to the conversion from analog to digital conversion of speech when saved in system's hard disk, and finding zero crossing to detect the end point of speech. This is done through End-Point detection Algorithm which is now replaced by Dynamic Time Wrapping Algorithm (DTW), End point detection fails in some cases like for example if users' password is 'Project' and user say it as 'Prrrooject' here first three letters of the password are said in slower version and last three with the same pace here. Speaker Recognition System are divide into two parts [9].

i. Text-Independent Speaker Verification System (TI-SV).
ii. Text-Dependent Speaker Verification System (TD-SV).

In TI-SV as the name suggest speaker is not restricted to speak a given word speaker. Here speaker has freedom of speech whereas in case of TD-SV speaker has to speak a a particular phrase or from set of phrases with which system was trained. Gaussian Mixture Model (GMM) is used in case of TI-SV whereas in case of TD-SV Hidden Markov Model (HMM) is used.

2.4 Algorithms Used in Speaker Recognition System

Speaker Recognition System consists of the following five steps:

i. Input Speech.
ii. Normalized Capture Speech.

iii. Feature Extraction.
iv. Similarity Matching.
v. Decision/ Threshold.

In the first step of input speech, there are various factors which affect the input speech or voice of a speaker. There are some hardware factors like microphone, further noisy environment or background cannot be away from background noise in real time applications. Also when audio signal is converted to digital signals some noise get added is called DC content, which is removed with the help of removed by removing or setting all the amplitudes in speech below 0.05Hz. There are other factors which affect input voice of human like accent, each person's voice is unique and voices can differ significantly from dialect to dialect, other factors are the speed or rate with which people speak having different pronunciation at different times like people during stress may speak the same word or sentence with different rate then when they are normal condition or in relaxed mood.

In the second step of Speaker Recognition is normalizing the speech DC content of the voice is removed which arises due to conversion of analog signal into digital signal and then end point detection algorithm is applied to find the end and start of the speech this removes the noise from the speech. This step is also called pre-emphasis of speech. The end point detection algorithm is as follows

i. The algorithm first removes the DC content then finds the zero crossing rate to identify where the speech exits and puts zero to the non-voice area of the signal.
ii. Compute the average magnitude and zero –crossing rate of the signal and background noise this enable us to remove noise from the signal.
iii. Then search for the signal magnitude which exceeds the previously threshold value which is marked as beginning of the speech
iv. From this point search backwards, search backwards until amplitude with less than threshold is not found.
v. From here, search the original twenty- five frames of the signal to locate the beginning of the speech
vi. The above process will be repeated for the end of speech signal to locate the end of signal.

The third step of speaker recognition system process is feature extraction where different algorithm like Linear Predictive Cepstral Coefficients (LPC) algorithm, Mel-frequency Cepstrum Coefficient (MFCC) are used or applied to extract features from speech. Approach applied in this thesis, tries to improve MFCC algorithm by changing the window algorithm and dynamic approach to reduce the time complexities and cost of computation for each frame. Initially Linear Predictive Cepstral Coefficient was used for feature extraction then came MFCC which mimics human hearing behavior. At this stage speech features are extracted some procedures require filter bank at this stage to cover the useful frequencies and discard the rest. The output of each filter is processed and processed at suitable frequency rate to obtain the energy from each filter. The energy formula is given by, where k represents the number of samples in a single frame and *Sn* represents amplitude of each frame [9].

$$\mathrm{E} = log \sum_{\mathrm{k}=0}^{160} {S_n}^2$$

Although filter banks provide highly effective and reliable features but there are some disadvantages to this, because once the speech is passed through the filter bank all the input depends upon combining different filter outputs. This makes procedure like pitch and format analysis difficult to measure. Hence in most of the methods that uses pitch other features like pitch, frequency and overall energy estimation is also included.

In case of using dynamic features for speaker recognition, the time function are brought, the time functions are brought into time registration with reference functions. The results of certain experiments show's that there is a slight difference in recognition accuracies compared to statistical features and dynamic features. Since the amount of calculation necessary for recognition using statistical feature is only one-tenth of the calculation needed by using dynamic features.

A large class of procedures requires segmentation of the input speech. The class contains both text dependent and text independent procedures. In case of text independent speaker recognition procedures, segmentation of speech is precisely aligning the occurrence of similar text events in reference and text utterances. This

process makes it possible to compare equivalent events and also compensate for the variation of recognition in different speaker The technique of dynamic time wrapping has found a wide spread use in speaker recognition technique the matching process in which the unknown input pattern coefficient is compared with stored coefficient reference pattern. The purpose of the time wrapping algorithm is to match the speaker even when the time registration of the same word is different at two different times. Speaker recognition system determines the speaker identity on the basis of his/her speech signals or voice signals. Speaker recognition systems are usually divided into two parts: speaker identification and speaker verification.

Chapter 3
Problem Statement

As discussed in literature survey, the interest of human to interact with machine is decade long. It started in 1950s but still even after the intense research on this topic there is a scope of improvement and research in the field of Speaker Recognition System. These gaps need to be filled to improve the accuracy and make these systems reliable to real world. Here in this thesis the algorithm which is used to extract features from voice is tried to improve. Following are some of the gaps which were found in conventional MFCC algorithm and are tried to improve in proposed algorithm.

i. The formula for pre-emphasis has been modified which reduces the number of multiplications keeping the recognition accuracy to be same
ii. The number of computation of sine and cosine which are used in windowing and for calculating Fast Fourier transformation (FFT) has been reduced by half compared to conventional algorithm
iii. Kaiser Window gives more accuracy instead of hamming window which gives more precision compared to hamming window

Chapter 4
Design and Implementation

As discussed in literature survey, Speaker Recognition is the process of automatically recognizing the speaker by extracting various features from the speaker voice like pitch, frequency rate and then with the help of pattern matching those features with the voices already stored in the databases or the voices with which system is already trained with. The speaker recognition system has two phases this has already been discussed in literature survey and introduction called enrollment session or training phase and operation phase or testing phase [13][14]. So in this chapter deals with proposed work to enhance the mel-frequency cepstrum Coefficient (MFCC) algorithm to reduce the time complexity and computation cost by using dynamic approach, i.e. creating lookup table for values which are already computed in previous loops for values of sine, cosine and log. In turns reduces the time complexity of conventional MFCC algorithm. Also different window algorithm Kaiser Window algorithm instead of Hamming Window algorithm which gives more precision as compared to hamming window algorithm.

In Automatic Speaker Recognition the mel-frequency cepstrum coefficient (MFCC) is widely used algorithm as compared to other feature extracting algorithms like LPC and others [14]. Due to increase in computation of speaker recognition system , the capacity of memory has been restricted in this field. Now to improve the most time-consuming Fast Fourier Transformation (FFT) establish look up table to improve the efficiency of algorithm [21]. The speech extraction feature module is to convert speech to some type of parametric representation which are then used as input to neural network. The block diagram of conventional MFCC algorithm is given below:

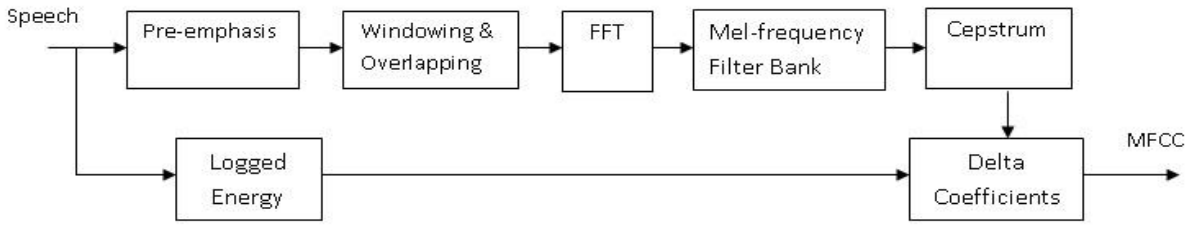

Fig 4.1: Block Diagram of MFCC [4]

The speech input is typically recorded at a sampling rate above 10000Hz, which is taken it as 16000Hz in code. The sampling frequency was chosen to minimize the effect of aliasing, i.e. overlapping of signal when converting from digital from analog

to digital. This sampling frequency can capture all the frequencies up to 5000Hz which cover most of the sound that are generated by humans. The main aim of MFCC is to mimic the behavior of human hearing, like the way humans perceive. The MFCC as shown above is figure 4.1 divided into 5 blocks frame blocking, Windowing and overlapping, FFT and Cepstrum coefficient. The process of computation of MFCC features for a given speech signal consists of following steps:

i. Speech signals are firstly divided into overlapping frames.
ii. Hamming window is applied to prevent discontinuities in each frame.
iii. Energy of each frame is calculated using energy formula.
iv. The power spectrum samples are computed for each samples
v. The logarithm coefficients are obtained for each sample in a frame.

4.1 The Optimization of MFCC

4.1.1 Establishing Sine Table

After analyzing the process of MFCC algorithm, it was known that computation of sine values were only used to weight cepstrum and FFT.

i. Weighted Cepstrum

The formula for weighted cepstrum is given below

$$Cn = \sum_{k=0}^{n-1} \log(s_k) \cos\left[\frac{n\left(k-\frac{1}{2}\right)\pi}{K}\right], \qquad n = 1,2,\ldots,K$$

Where S_k is the output power spectrum of a filters. The value of K is chosen to be 12, which is considered to be standard [15].

As can be seen from the figure 4.2 value of cos ranges from 0-1 and repeats after interval of 10, 33 filter banks have been used, so there is no need to compute values 33 times which reduces computation time.

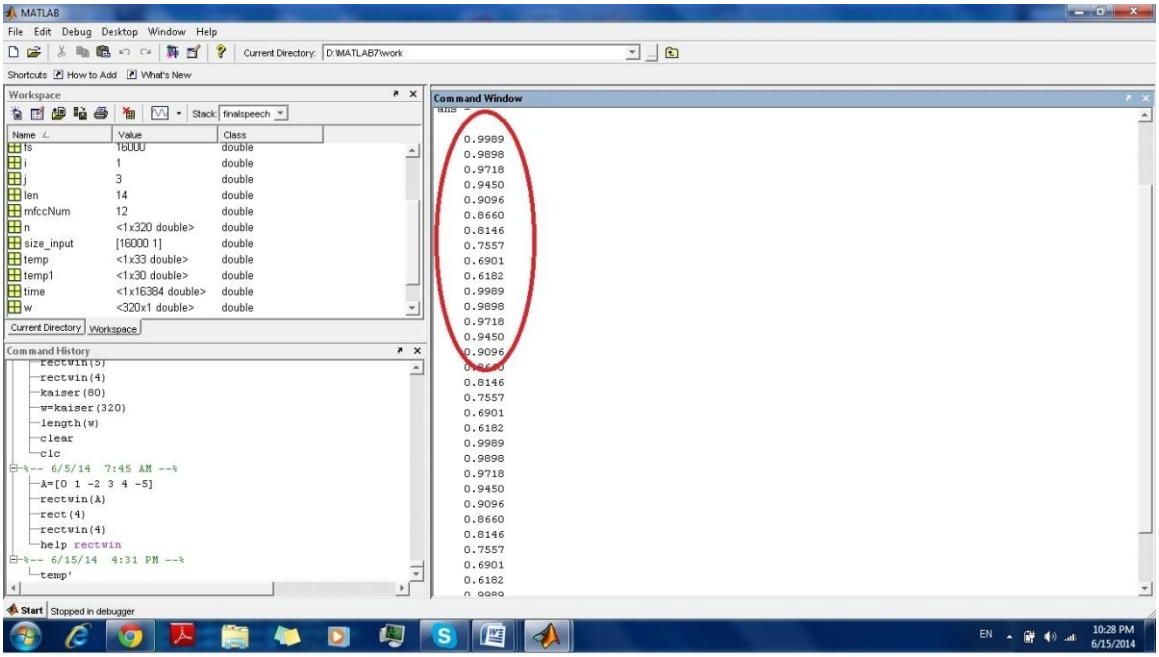

Fig 4.2: Cos values

As seen from the figure 4.2 the value of cos is getting repeated after 10 values so it decreases the number of computations by using look up table. Algorithm for which is given it can be implemented in MATLAB, C, C++ or java.

Algorithm :Lookup table for Cosine

Ouput: Lookup table for Cosine

1. Set i <- 1, mfccNum <-12
2. for i<-1 to mfccNum
3. set tempmat <- [];
4. Compute tempmat -> cos((pi/filterNum)*i*((1:10)-0.5));
5. tempcolmat=[];
6. for j->1 to 3
7. tempcolmat -> [tempcolmat tempmat];
8. end for
9. temp <-[tempcolmat tempcolmat(1) tempcolmat(2) tempcolmat(3)];
10. coef <-[coef;temp];
11. end for

The mfccNum in this code is taken as 12 and filterNum as 33.

ii. Fast Fourier Transform

The formula for FFT is given by:

$$\mathrm{X}(k) = \sum_{n=0}^{N-1} \mathrm{x(n)} {W_N}^{kn}, 0 < k < N-1$$

Here ${W_N}^{kn} = e^{-j\,2\pi kn\,/N}$, n is number of speech samples.

This formula can be written in terms of cos and sin by Euler's formula as

$$\mathrm{X}(k) = \sum_{n=0}^{N-1} \mathrm{x(n)} \left[\sin\left(-\frac{2\pi kn}{N}\right) + j\cos\left(-\frac{2\pi kn}{N}\right)\right], 0 \le \mathrm{k} \le \mathrm{N}-1$$

As seen from the above formula like sin and cosine look up table will decrease the time complexity of a program.

4.1.2 Window Algorithm

The next step in the processing of MFCC algorithm is windowing, like window each individual frame to minimize the signal discontinuities at the beginning and end of each frame. The objective here is to minimize the spectral distortion by using the window. For this hamming window algorithm is used, the formula for it is given by

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi \mathrm{n}}{\mathrm{N}-1}\right), \qquad 0 \le n \le N-1$$

where N is the number of samples in each frame, the result of windowing can be calculate with the help of the following equation.

$$\mathrm{y}_l(n) = x_l(n)w(n), \qquad 0 \le n \le N-1$$

The result of applying window algorithm to the signals can be seen in figure 4.3, the original signal had distortion which has been removed during after applying hamming window algorithm.

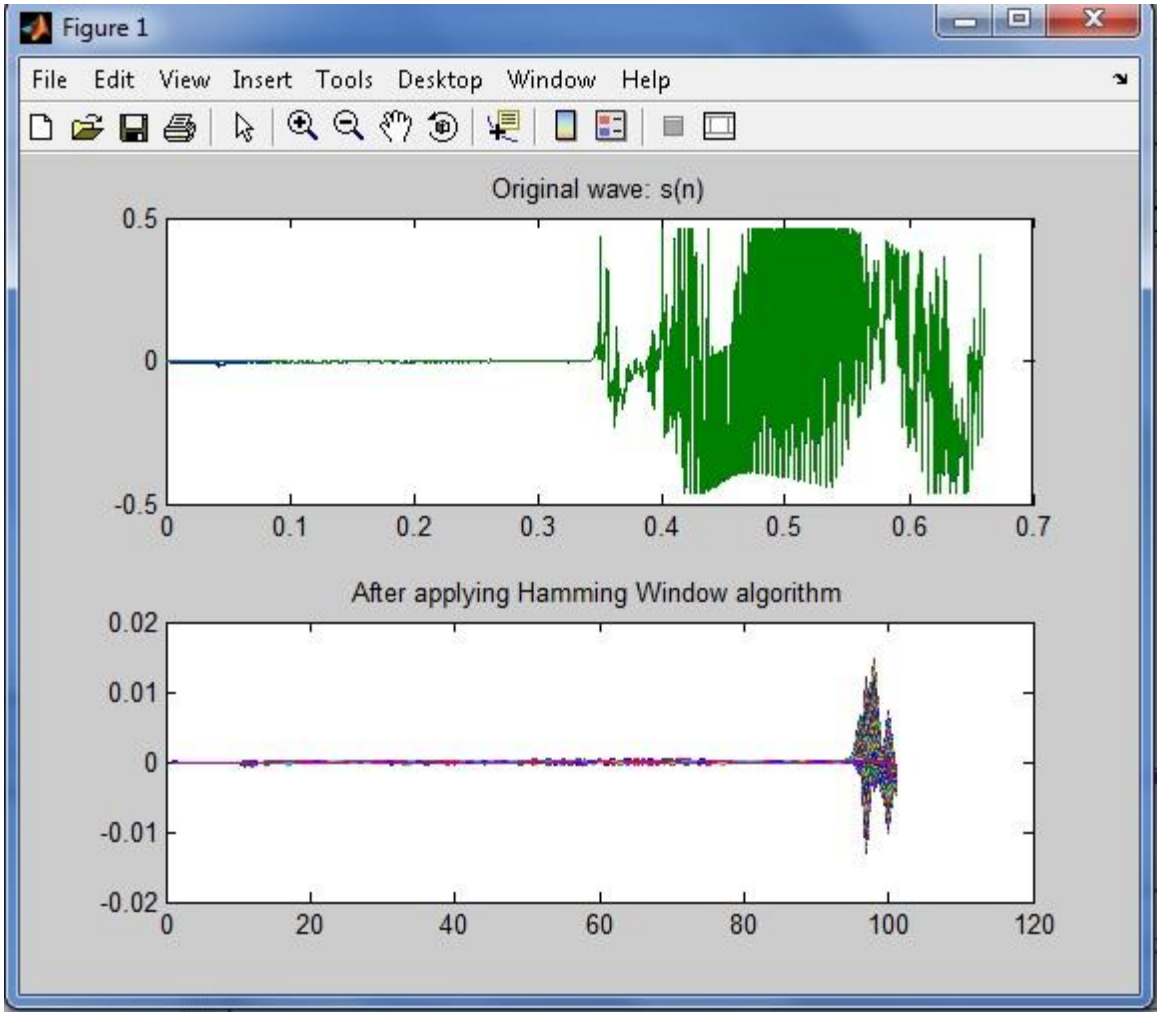

Fig 4.3: Signal after Applying Hamming Window Algorithm

The MATLAB has in built function for finding hamming window, hamming().

The Matlab source code for hamming window algorithm on signal is shown below. As seen from the code below the size of each frame is taken to be 160, this will be explained in later why it has been taken 160 as frame size instead and no overlapping is done. In new approach Kaiser Window algorithm is applied instead of hamming window algorithm the formula for which is given by:

$$w_k = \begin{cases} \dfrac{I_0\left(\beta\sqrt{1-\left(\frac{2k}{N}-1\right)^2}\right)}{I_0(\beta)} & 0 \le k \le N \\ 0 & \text{otherwise} \end{cases}$$

MATLAB Source Code for Hamming Window

Algorithm : Hamming Window Algorithm

Input: Wave file

Ouput: Breaking wave file into Windows of length 160

```
1.  Set  y <- 0, fs <-16000Hz
2.  [y fs]=wavread('C:\Users\Shri Ram\Desktop\viki.wav');  % Reading Wave file
3.  Set size_input <-size(y);
4.  subplot(2,1,1);
5.  plot(time, y);
6.  title('Original wave: s(n)');
7.  %Hamming window calculation
8.  for n<-0 to319
9.      w(n+1)=0.54-0.46*cos(2*pi*n/319);  %Formula for Hamming Window
10. end for
11. % Deviding into frames
12. L<-100;
13. Set temp=[],frame=[];
14. for i=0 to L
15.     j <-160*i+1:320+160*i;
16.    n<-0:319;
17.    temp(n+1)<-y2(j).*w(n+1);
18.    frame<-[frame;temp];
19. end for
20. subplot(2,1,2);
21. plot(frame);
22. title('After applying Hamming Window algorithm');
```

The figure obtained from above code is shown above. In case of Kaiser algorithm will remain same only the kaiser window equation window will be used instead of hamming window.

4.1.3 Pre-emphasis of Signal

The speech is first pre-emphasized to flatten the speech signal, the formula to pre-emphasize the speech signal is given by following equation:

$$S_n{}' = S_n - aS_{n-1}$$

The value of '*a'* varies from 0.9-1, which is chosen as 0.98, now in this if chosen '*a'* is chosen as 31/32, which is approximately 0.97, then replaces multiplication by simple addition or shift which is less costly compared to multiplication, so the equation becomes:

$$S_n{}' = S_n - \left(S_{n-1} - \frac{S_{n-1}}{32}\right)$$

This replaces the complex multiplication with simple shift operation without affecting the recognition accuracy. The signal after applying pre-emphasis can be seen figure 4.4.

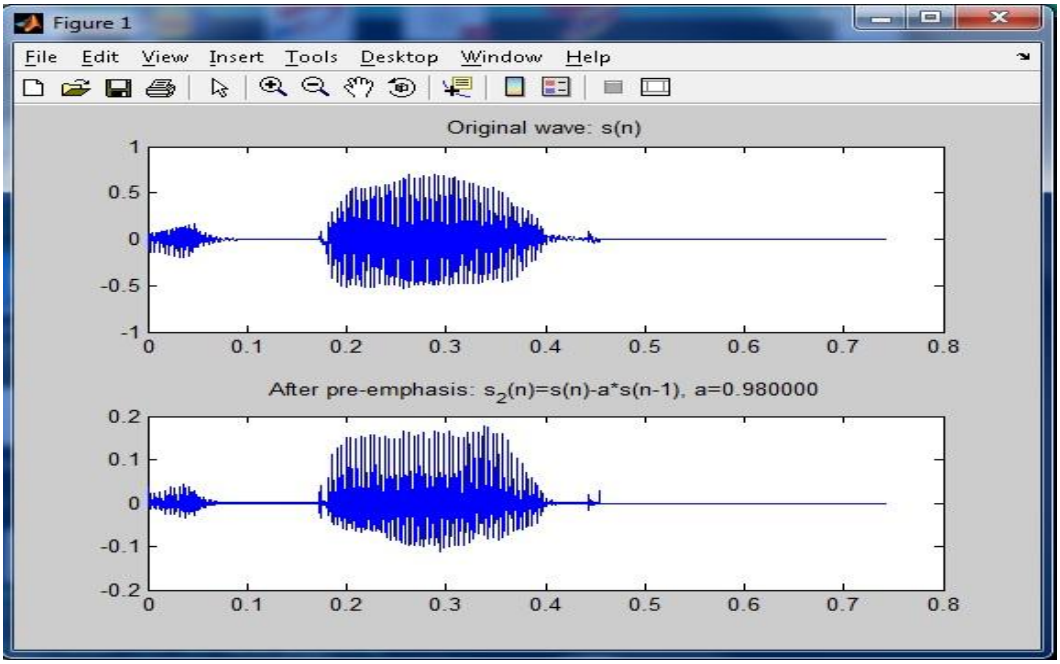

Fig 4.4: Pre-emphasis of a signal

4.1.4 Frame Blocking

In conventional MFCC algorithm two frames are overlapped as shown in general block diagram of MFCC algorithm, now in case of improvised algorithm the frame of length 160 is taken instead of 320. And overlapping is moved after mel-frequency cepstrum. The diagram of new model of MFCC is shown below in figure 4.5

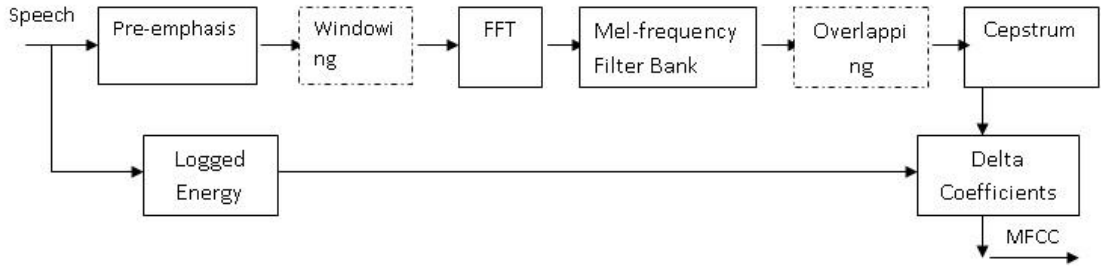

Fig 4.5: New Model for MFCC Algorithm

Initially, there was frame overlapping as can be seen from following figure 4.6 which has been removed in proposed algorithm which in turns reduces the number of multiplications,

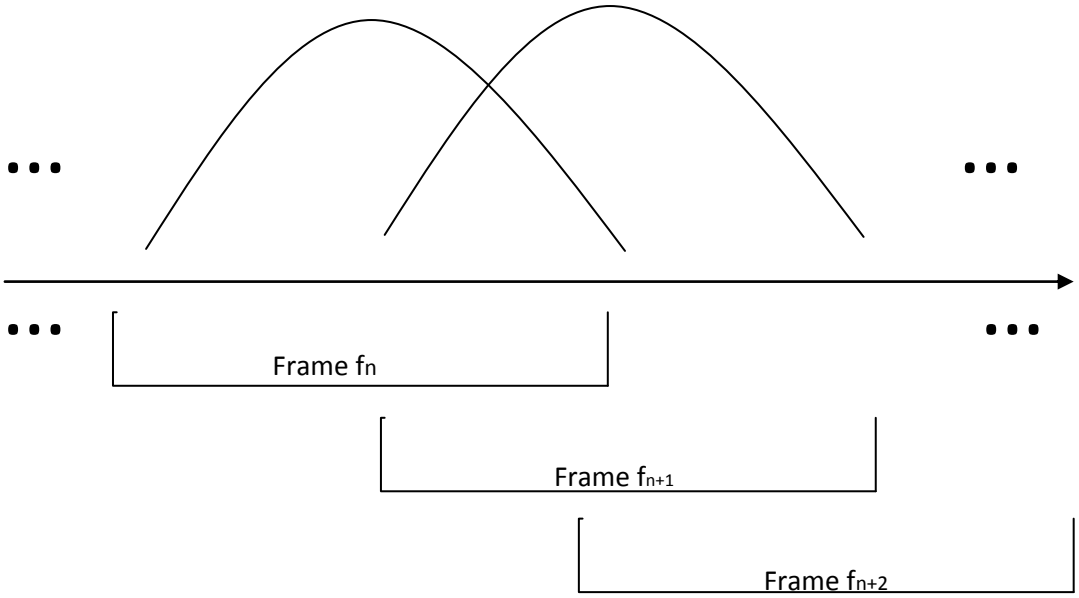

Fig 4.6: Frame Overlapping in Conventional MFCC algorithm [15]

Now as seen from figure 4.6 speech signals is first divided into frames, in new approach overlapping has been after finding MFCC cepstrum, so in new approach the frame which were initially 320 samples of each are now divided into 160 samples of sub frames, this is shown in figure 4.7

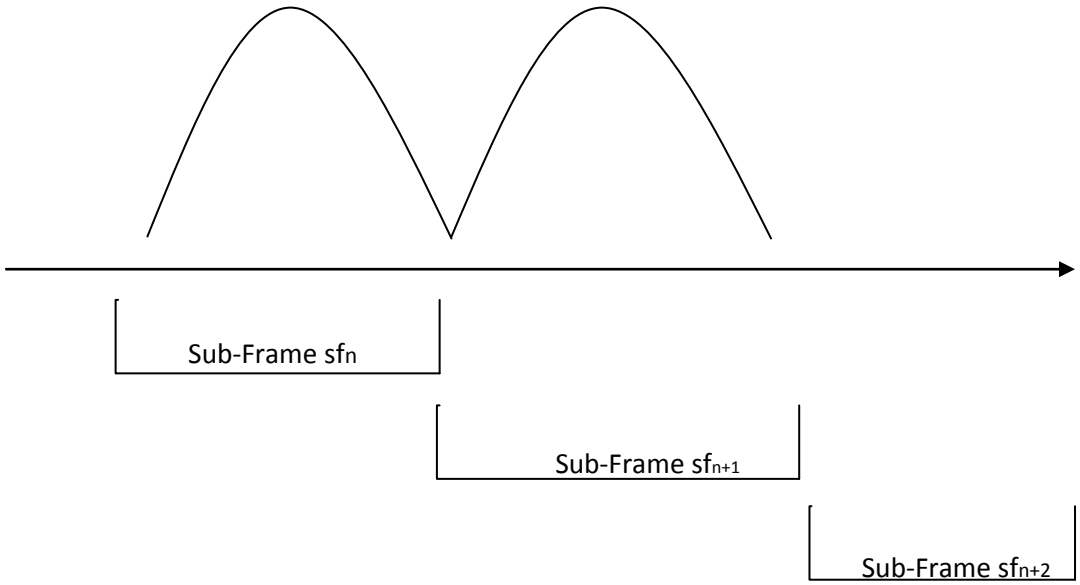

Fig 4.7: Frame Blocking of Proposed MFCC Algorithm [15]

With this the calculation is reduced by 128 points because of the new window size, so only the first 64 coefficients need to be calculated because of the FFT symmetry. In case of proposed algorithm instead of taking triangular filter banks and then multiplying them with the output of each frame after applying hamming window algorithm, rectangular window has been used whose outputs are normally in terms of 0 or 1, thus reducing the time complexity by replacing multiplications by addition of frame. This reduces the CPU time for multiplication and in turn reduces time complexity. The rectangular window is shown in figure 4.8. The MATLAB code for triangular filter bank is freely available on internet with getfilterbank.m file. The function in MATLAB triang() is also available to find triangular parameters.

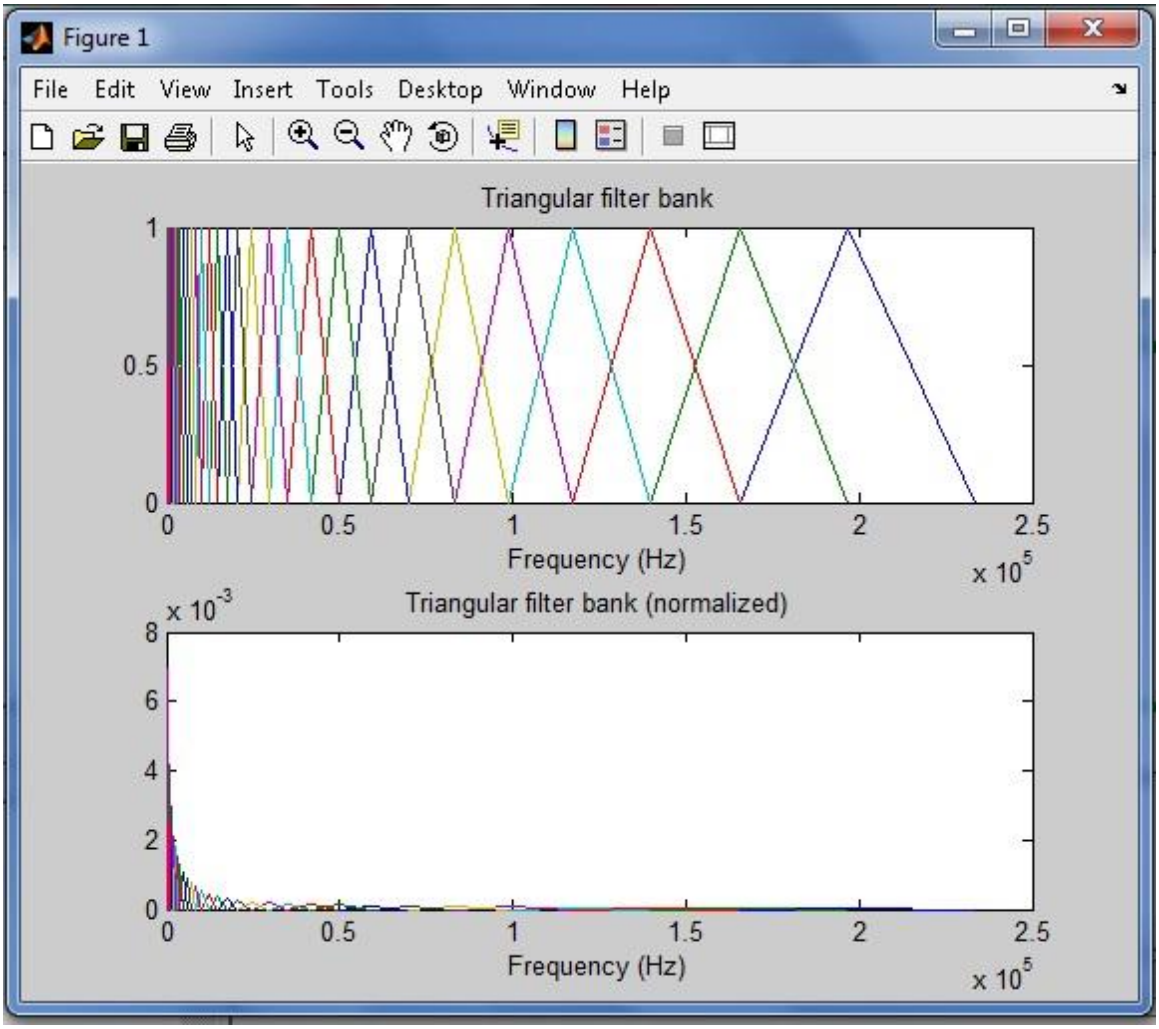

Fig 4.8: Triangular Filter

Before applying filter banks mel-scale need to be applied to the speech signals, mel scale is normally used as it mimic the human hearing system. The formula for mel-scale is given by:

$$mel(f) = 2595 * log\left(1 + \frac{f}{700}\right),$$

This 2595, 16000/(1+log16000/700) , where16000 is sampling frequency.

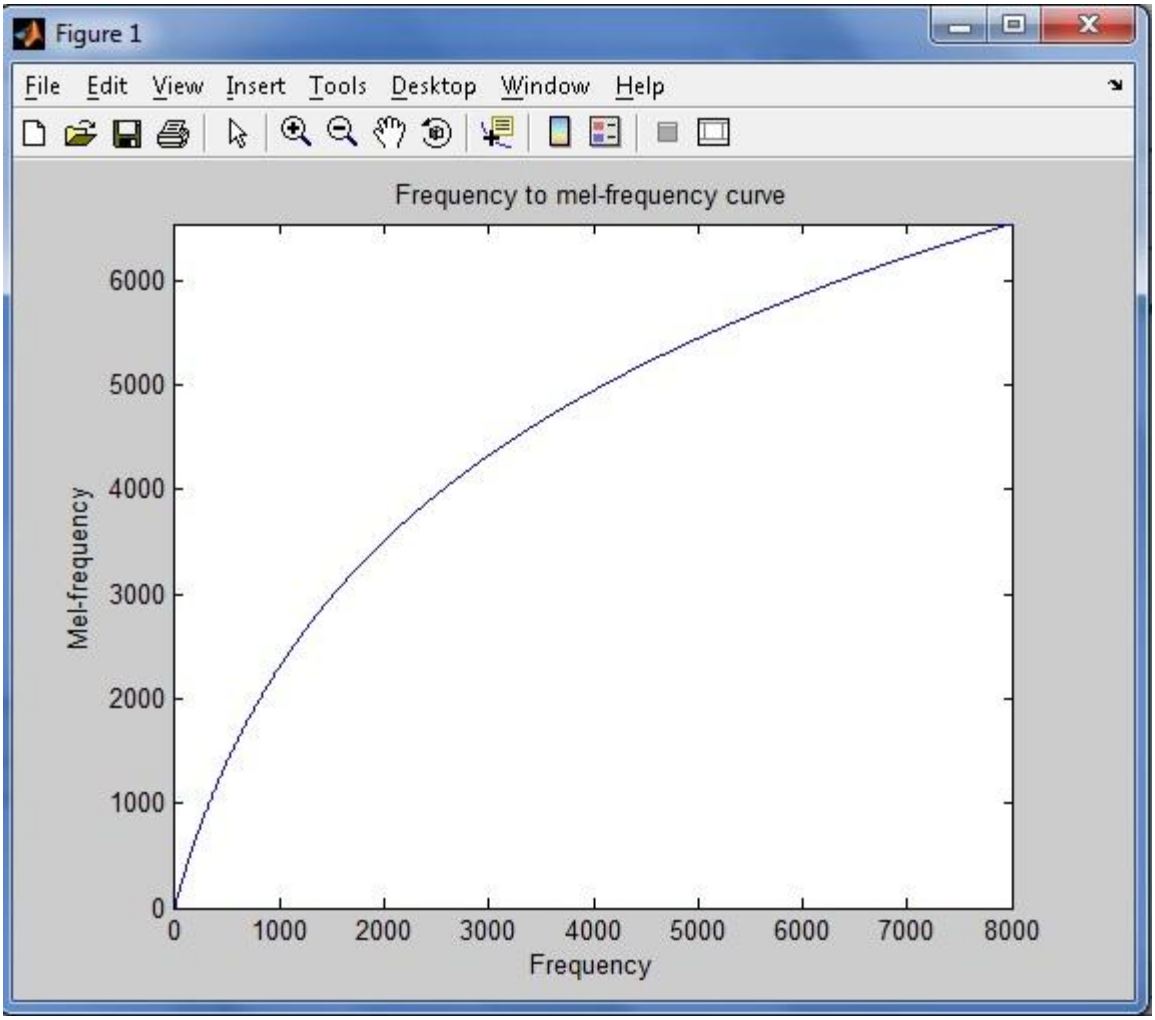

Fig 4.9: Diagram for mel-scale

Chapter 5
Experimental Results and Discussion

The project has been implemented on MATLAB, which is a high performance language for technical and advanced computing. MATLAB has neural network toolbox which consists of simple elements operating in parallel, called neurons. This neural network is inspired by the biological nervous systems. Neural network can be trained to perform various functions by adjusting the values of weight between different elements. Commonly neural networks are adjusted or trained so that specific input leads to the particular target vector. The following pseudo code illustrates the simple example of neural network; here network has been trained to find the square of

Algorithm: Example of Neural Network

Input: Matrix containing values from 1 to 10

Output: Sqaure of value, with which it is simulated

```
1. Set P<-[1 2 3 4 5 6 7 8 9 10];   % Input Matrix
2. Set T<-[1 4 9 16 25 36 49 64 81 100];  %Output Matrix
3. net=newff(minmax(P),[5 1],{'tansig' 'purelin'});  % train network
4. net.trainParam.goal=0.01;
5. net.trainParam.epoch=100;
6. net=train(net,P,T);
7. Y=sim(net,11)
```

a number after training. Initially network is trained with the square of first 10 natural numbers. Here P is a training set, and T, is target vector. Newff() is a built in function in MATLAB to train neural network. It gives more accurate result if the network is trained with more input vectors. Figure 5.1 show the training of neural network.

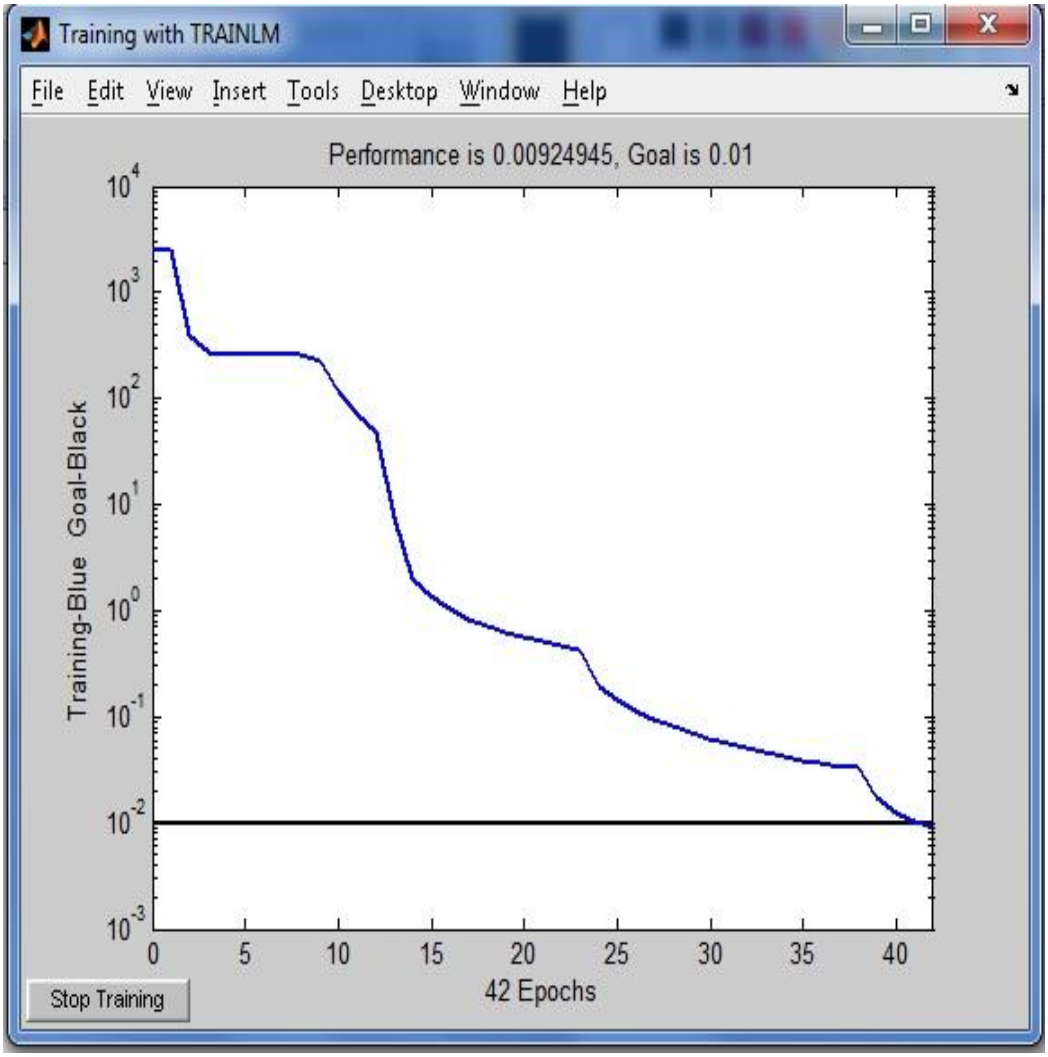

Fig 5.1: Training Network

```
Command Window
TRAINLM, Epoch 0/100, MSE 2592.78/0.01, Gradient 1358.87/1e-010
TRAINLM, Epoch 25/100, MSE 0.141694/0.01, Gradient 318.279/1e-010
TRAINLM, Epoch 42/100, MSE 0.00924945/0.01, Gradient 55.4343/1e-010
TRAINLM, Performance goal met.


Y =

  119.4807

>>
```

Fig 5.2: Output of Neural Network

As it can be seen from the figure 5.2 and MATLAB code even though system is not trained with output vector of 11 as can be seen from the from figure 5.2 the output of

Y=sim(11,net), comes out to be 119.48, which is near to 121. To find accurate result network need to be trained with more input vectors.

5.1 Test Results after Establishing Sin, Cos, and Log Table

The experiments were performed on MATLAB, intel core2Duo processor T600 2.10GHz, with 4GB RAM and on Windows 7 as operating System. The following table 5-1 illustrates the comparison of conventional MFCC algorithm or Roger Jang's method and enhanced MFCC algorithm:

Table 5.1: Frame Execution Time of MFCC

Parameters	Roger Jang's method	Proposed Method
Compute Energy	3139ns	3159ns
Pre-emphasis	1019ns	1006ns
Hamming	1737ns	1546ns
FFT	52834ns	53935ns
LOG	3619ns	2618ns
DCT	12275ns	4515ns
Delta	706ns	706ns

The table 5.1 shows that the algorithm increases the execution time, although the execution time for MFCC has increased. Now checking the accuracy after using Kaiser Window and replacing α in pre-emphasis state by 31/32 which is nothing but 0.97, but this comes with advantage that it reduces the number of multiplication thus reducing CPU cycle and improving time complexity of algorithm. The following table 5.2 shows the comparison.

Table 5.2: Recognition Accuracy with different Feature Set

Feature Set	α	Window Length	FFT point	Filter Shape	No. of Filter	Recognition Accuracy
F1	0.97	320	256	Triangle	33	94.43%
F2	31/32	320	256	Triangle	33	94.43%
F3	31/32	160	128	Triangle	33	92.29%
F4	31/32	160	128	Rectangle	33	92.08%
F5	31/32	160	128	Rectangle	33	92.93%

Chapter 6
Conclusion and Future Scope

6.1 Conclusion

In this thesis more precise method has been applied like look-up table of sine, cosine, and log which are used for calculating FFT and windowing, and log look-up table is used in calculating cepstrum. This look-up table reduces the time complexity or execution time of an algorithm. Also the formula for pre-emphasis has been changed. Previously in conventional MFCC algorithm α was taken as 0.98 which has been replaced by 31/32 which is approximately equal to 0.97 (decimal value of 31/32). This also improves the time complexity of algorithm without affecting the recognition accuracy. Here a more efficient Kaiser Window algorithm is used as compared to hamming window algorithm which is more cost effective but are more precise.

6.2 Future Scope

Even though Speaker Recognition is a decade old and researchers have shown lot of interest in human-machine communication but there is still a scope of improvement and research in this field. There is lot of scope in this field to work on Emotion Neutralization system, in which if a person speaks in sad voice or angrily the system may fail to detect the person similarly in case of accent neutralization [27][28]. People from different region have different accent, even in India people from different state may speak same word differently, so there's a lot of scope is this field of speaker recognition system [3].

References

[1] L. Rabinier and B. H. Juang, "Fundamental of Speech Recognition," *10th Edition Prentice Hall International*, pp 3.

[2] K. H. Davis, "Automatic recognition of Spoken Digits," *Journal of Acoust. Society America*, vol. 24(6), pp. 637-642, 1952.

[3] L . Rabinier and B. H. Juang, " Fundamental of speech recognition," *10th Edition Prentice Hall International*, pp 17-20.

[4] H. F. Olson and H. Belar, "Phonetic Typewriter," *Journal of Acoustic Society America*. vol. 28(6): pp 1072-1081, 1956.

[5] D. O'Shaughnessy, "Invited paper: Automatic speech recognition: History, methods and challenges," *Pattern Recognition*, vol 41(10), pp. 2965-2979, 2008.

[6] N. S. Dey, R. Mohanty, K. L. Chugh, "Speech emotion recognition using Hidden Markov Model," *International Conference on Communication System and Network Technologies*, *IEEE Computer Society*, 2012.

[7] I. R. Murray and J. L. Arnott, "Towards the simulation of Synthetic Speech: A review of the literature on Human vocal emotion," *J. Acoustic Society of America*, vol 93(2), pp. 1907-1108, 1993.

[8] A. Hagen and A. Morris, "Recent advances in multi-stream HMM/ANN hybrid approach to noise robust ASR," *Computer Speech and Language*, vol 19(1), pp 3-30, 2005.

[9] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol 16(3), pp. 261-291, 1995.

[10] D. B. Fry, "Theoretical aspects of Mechanical Speech Recognition," *J. British Inst. Radio Engineering*, vol.19(4), pp. 211-299, 1959.

[11] J. Suzuki and K. Nakata "Recognition of Japanese Vowels – Preliminary to the Recognition of Speech," *J. Radio Research Lab*, vol.37 (8), pp. 193-212, 1961.

[12] L. Rabinier and B. H. Juang, "Speaker Independent Recognition of isolated Words using Clustering Techniques," *IEEE Trans. Acoustics, Speech Signal Proceedings* ASSP-27, pp.336-349, 1979.

[13] J. Ferguson, and L. R. Rabinier, "Hidden Markov Model fo Speech," *IDA Princeton*, NJ 1980.

[14] L. R. Rabinier, "A tutorial on Hidden Markov Model and selected Application in Speech Recognition," Proc IEEE, vol.77 (2), pp. 257-286, 1989.

[15] A. E. Rosenburg, "Automatic Speaker Verification: A Review," Proc. IEEE vol. 64(4), 1976.

[16] C. G. K. Leon, "Robust computer voice Recognition Using Improved MFCC Algorithm", *International Conference on Information and Service Science*, 2009.

[17] A. K. Abbasi, "A paper on Speaker Recognition Using Orthogonal LPC Parameter in Noisy Environment," *MS Thesis, King Fahd University of Petroleum & Minerals*, 2005.

[18] W. Han, "An efficient MFCC Extraction Method in Speech Recognition," *ISCAS'*06.

[19] F. Zheng. "Comaprison of Different Implementation of MFCC," *Journal of Computer Science and Technology*, vol. 16(6), pp. 582-589, 2001.

[20] G. Zhang and F. Zheng, "The Fixed-Point Optimization of Mel Frequency Cepstrum Coefficients for Speech Recognition," *6th International Forum on Strategic Technology*, 2011.

[21] S. B. Davis and P. Mermelestien, "Comparision of Parametric Representations for Monosyllabic Word Recognition in Continiously Spoken Senteneces," *IEEE Transaction on Acoustic, Speech, and Signal Processing*, vol. 28(24), 1980.

[22] X. Zhang, "A Speech Endpoint Detection Algorithm Based on Entropy and RBF Neural Network," *International Conference on Granular Computing, IEEE Computer Society*, 2011.

[23] Q. Lank and X. Zhang "A High Performance Auditory Feature for Robot Speech Recognition," in *6th International Conference on Spoken Language Processing*, Beijing, 2000.

[24] J. J. Wolf, "Efficient Acoustic for Speaker Recognition," *Journal of Acoustic Society of America*, vol. 51, pp 2044-2055, 1972.

[25] Z. Zhang, "Speaker Verification: Text-Independent vs. Text-Dependent,"

Available at: research.microsoft.com/en-us/um/people/zhang/Speaker%20Verification/default.htm, 2006.

[26] S. W. Smith, "The Scientist and Engineer's Guide to Technical Publishing," *California Technical Publishing*, pp 169-174, 1997.

[27] I. R Murray and J. L. Arnott, "Towards the simulation of Synthetic Speech: A review of the literature on Human vocal emotion," *Journal of Acoustic Society of America*, vol 93(2), pp 1907-1108, 1993.

[28] A. Hagen and A Morris, "Recent advances in multi-stream HMM/ANN hybrid approach to noise robust ASR," *Computer Speech and Language*, vol 19(1), pp 3-30, 2005.

[29] J. J. Wolf, "Efficient Acoustic for Speaker Recognition," *Journal of Acoustic society of America*, vol. 51, pp 2044-2055, 1972.

[30] G. Saha, S. Chakroborty, S. Senapati," A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications", Available at: citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.138.623&rep=rep1&type=pd

LIST OF PUBLICATIONS

Communicated:

[1] V. Ojha and R. K. Tekchandani "Enhanced MFCC Algorithm using Lookup Table and Kaiser Window" in Third International Conference on Advances in Computing, Communications and Informatics (ICACCI-2014)

[2] V. Ojha and R. K. Tekchandani "Literature Review on Speaker Recognition System" in Third International Conference on Advances in Computing, Communications and Informatics (ICACCI-2014).