NOVATEUR PUBLICATIONS
INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING RESEARCH AND TECHNOLOGY [IJIERT]
ISSN: 2394-3696
VOLUME 3, ISSUE 12, Dec.-2016

IMPLEMENTATION OF SPEECH RECOGNITION SYSTEM

MR. MAINDARGI L. C.
PG student, Department of E&TC, VVPIET, Solapur

PROF. MANTRI D.B.
Assistant Professor, Department of E&TC, VVPIET, Solapur

ABSTRACT
Speech recognition is an important and active analysis area of the recent years. This analysis aims to make a system for speech recognition with the help of dynamic time wrapping algorithm program, by examining the speech signal of the speaker with pre-stored speech signals within the stored database, and extracting by using Mel-frequency Cepstral coefficients which is the main features of the speaker speech signal and one of the most necessary factors in achieving high recognition accuracy. The process of extraction and matching is implemented after the Pre Process or filtering signal is performed. The non-parametric methodology for modeling the human perception system, MFCCs (Mel Frequency Cepstral Coefficients) are utilizing as extraction techniques. The non linear sequence alignment called DTW (Dynamic Time Warping) which is introduced by Sakoe Chiba has been used for features matching techniques. As it is acceptable that the speech signal tends to have various temporal rates, the alignment is essential to provide the enhanced performance. This analysis presents the MFCC viability for feature extraction and DTW for comparing the test patterns.

INTRODUCTION
The speech signal of a person is unique and never changing. The signal taken as an input can be stored in template format and the pre-stored templates can be compared with unknown or the signal which is given as input and exact match can be found out. The output which is matched signal can be used for turning on different mechanical machines, for activation of machines in electronics or any industry and for recognition of speech of unknown person for security purpose.
Speech recognition is the technique of automatically recognizing the spoken words of person based on information in speech signal. Each expressed word is generated by using the phonetic combination of a set of vowel, semivowel and consonant speech of sound units. Mel Frequency Cepstral Coefficients also known as MFCC is well known spectral based parameter used in recognition approach. MFCCs are coefficients that represent audio, supported perception of human auditory systems. The fundamental difference among the process of FFT/DCT and the MFCC is that in the MFCC, the frequency bands are to be found logarithmically (on the mel scale) that approximates the human auditory systems response closer than the linearly spaced frequency bands of FFT or DCT. Because of its advantage i.e. less complexness in execution of feature extraction algorithm, certain coefficients of MFCC related to the Mel scale frequencies of speech Cepstrum are extracted from spoken word samples stored in the database.

LITERATURE REVIEW
Speech Recognition analysis has been ongoing for more than 80 years. Over that period there have been at least 4 generations of approaches, and a 5th generation is being formulated based on current research themes. To cover the whole history of speech recognition is away from the scope of this analysis.
By 2001, speech recognition with the help of personal computer had accomplished 80% accurateness and after that no more progress was reported until 2010.
The development in Speech recognition technology began to edge into the front position with one most important event i.e. the introduction of the “Google Speech Search app for the iPhone”. During year 2010, Google added “personalized recognition” to Speech Search on Android phones, so that

the software could record users' speech searches and produces a more precise speech model. In addition with that the company also added Speech Search in its Chrome Browser during mid-2011. It represents its knowledge about the speaker to create a contextual reply and responds to speech input. Similar handing out methods with the use of mixture of HMMs and acoustic- phonetic approaches to identify and accurate linguistic irregularities are used to enhance recognition decision reliability and increase robustness to recognize of speech in noisy surroundings.

METHODOLOGY

A sensor or detector that makes acquisition of data and its consequent sampling: in the precise case the sensor or detector is a microphone (Mike), probably with a high SNR (Signal to Ratio) value. Since the signal is actually speech, the rate is typically set to 8Khz. A step of preprocessing that within the speech context is set up by the signal cleaning: simply de -noising algorithm program are often applied to recorded information after a procedure of normalization. With the intention of clean recorded speech signal from environmental additive noise, a spectral subtraction algorithm.

The extraction of the odd characteristics (feature extraction): in this stage evaluation of Mel frequency Cepstral coefficients are made using a Mel filter bank after transforming the frequency axis into logarithmic one. The generation of a particular sample for each speaker: in this process we have decided to use the GMM (Gaussian Mixture Models) where model parameters are estimated with the maximum similarity make with the use of Expectation and Maximization (EM) algorithm. In case, the user is registering (enrollment) for the first time to the system, this template will be added to the database, using some database programming techniques.

Otherwise, in case of test with users whoever already there in the database, a matching (comparison) decides which profile matches the generated template of the test speech. The matcher applies a correspondence test, obtaining by a ratio value that can be acknowledged if it is higher than a decision threshold. The typical ASR (Automatic Speech Recognition) system is shown in Figure 1. The technologies used for the development of the biometric system are the MFCC for the extraction of the characteristics and the GMM for the statistical analysis of the data obtained, for the templates generation and for the comparison.

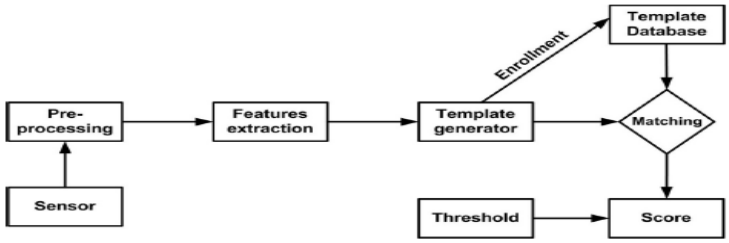

Fig.1- ASR System

A. RECOGNITION MODULE

Isolated word detection involves two digital signal Processes which are Feature Extraction and Feature Matching. Feature extraction involves calculation of MFCCs for each frame. MFCCs are the coefficients that put together to represent the temporary power spectrum of a sound, derived from a linear cosine transform of a log power spectrum on a nonlinear MEL scale of frequency. DTW method is used for feature matching.

B. FEATURE EXTRACTION: MFCC (MEL FREQUENCY CEPSTRAL COEFFICIENTS) :

MFCC is selected for the below given reasons

1. MFCC has most significant features, which are necessary among different kinds of applications in speech.

2. It provides high accuracy outcomes for clean speech.
3. MFCCs are often considered as the "standard" features not only in speaker but also in speech recognition.
4. In the MFCC frequency bands are positioned logarithmically that approximates the human auditory systems response closer than the linearly spaced frequency bands of FFT or DCT.

Cepstral analysis refers to the process of finding the Cepstrum of a speech sequence. Cepstrum, whose spelling is created by shuffling the characters of the word "spectrum", is a "time domain" representation of a signal. However, a special term of "frequency", which is also created by shuffling the spelling of the word "frequency", is often used instead of "time". It was explained in the year 1963 by Bogert et al. The inverse Fourier transform of the logarithm of the spectrum of a signal is called as Cepstrum.

$$x_c(n) = \mathrm{DFT}^{-1}\{\log|\mathrm{DFT}\{x(n)\}|\} \quad (1)$$

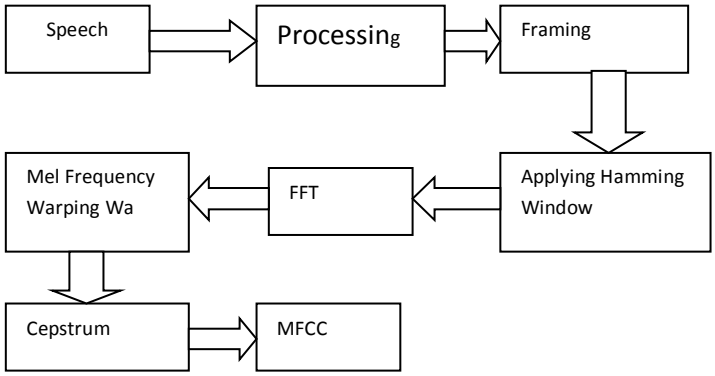

Fig.2 - MFCC flow diagram

The Cepstrum transforms the signal from the frequency domain into the frequency domain. When Cepstrum is applied to the voice, its strength is to be able to divide excitation and transfer function. In a signal y(n) based on the source-filter model, in this specific context, respectively the vocal cords and the vocal tract, Cepstrum allows separation in" y(n)=x(n)*h(n), where the source x(n) passes through a filter described by the impulse response h(n).

The spectrum of y(n) obtained by the Fourier transform is y(k)=X(k) H(k), where k index of discrete frequencies, i.e. the product of two spectra, correspondingly the source and the filter one. Separating these two spectra is complicated. On the contrary, it is possible to split the real envelope of the filter from the remaining spectrum by formulating all the phase at the beginning. The properties of the logarithm that can transform the product of the argument in sums of logarithms is the base of Cepstrum. Starting from the logarithm of the modulus of the spectrum:

$$\log|Y(k)| = \log(|X(k)\,H(k)|)$$
$$= \log(X(k)) + \log(H(k)) \quad (2)$$

Frame with Triangular MEL Filter banks shown in Fig. 3.

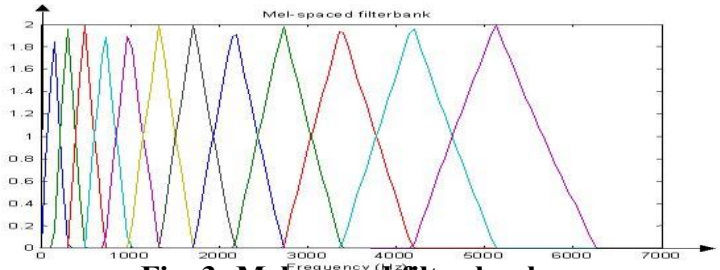

Fig. 3- Mel-spaced filter banks

It is feasible to split the fast oscillating component from the slow one, correspondingly by means of a high and low pass filter, obtaining:

$$c(n) = DFT^{-1}(\log|Y(k)|)$$
$$= DFT^{-1}(\log|X(k)|)+DFT^{-1}(\log|Y(k)|) \qquad (3)$$

That is the signal Cepstrum in the frequency domain. In the low frequencies are described the transfer function information, in the high frequencies there is data about excitation. Hence the original wave of percussion generated by the vocal cords and formed by the throat, nose and mouth can be evaluated as a sum of a source function (given by the excitation of the vocal cords) and a filter (throat, nose, mouth). The separation between high and low frequency, can be obtained by a high pass lifter (filter) for the fast oscillation and a low pass lifter for the slow one. Psychoacoustic research have shown that the mind perception of the frequency content of the sound follows a nearly logarithmic scale, the Mel scale, which is linear up to 1 kHz and logarithmic there after:

$$mel(f) = \begin{cases} f \text{ if } \leq 1 \text{ KHz} \\ 2595 \log\left[1 + \frac{f}{7000}\right] \text{ if } f > 1 \text{ KHz} \end{cases} \qquad (4)$$

Where Fmel = resulting frequency on the mel scale measured in mels and FHz = normal frequency measured in Hz.

The Mel scale is shown in Figure 4.1, where it is clear the compression of the Mel scale, as reported in y-axis, with respect to the Hertz scale, as reported in x-axis, for frequencies greater than 1 kHz. In this scale pitches are reviewed by listeners to be equal in distance from one another

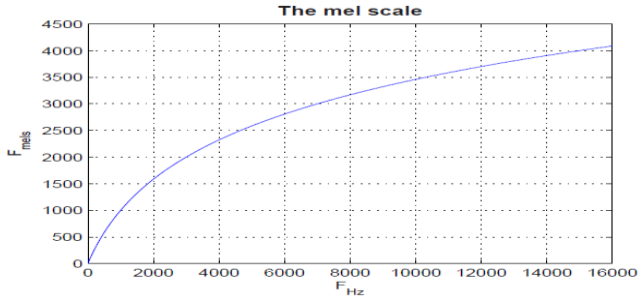

Fig 4 Frequency scale and Mel scale relationship

Mel-Cepstrum estimates the spectral envelope of the output of the filter bank. Let Yn represent the logarithm of the output energy from channel n, applying the discrete cosine transform (DCT) we obtain the Cepstral coefficients MFCC through the following equation":

$$c_k = \sum_{n=1}^{N} Y_n \cos\left[k\left(n - \frac{1}{2}\right)\frac{\pi}{N}\right] \forall k = 0, \dots, \dots K \qquad (5)$$

The spectral envelope is rebuilt with the
First Km coefficients, with Km < K

$$C(mel) = \sum_{k=1}^{K_m} c_k \cos\left(2\pi k \frac{mel}{B_m}\right) \qquad (6)$$

where Bm is the bandwidth analyzed in Mel domain and Km = 20 is a ordinary value assumed by Km. c_0 which is the mean value measured in dB of the energy of the filter bank channels, therefore it is in direct relation with the energy of the sound. It can be useful for the estimation of the energy.

Schematically, the coefficients are derived in the following way:

The spectrum of the original signal is computed with the Fourier transform.

The obtained spectrum is summarized in Mel making use of suitable overlapping windows.

For each obtained function the logarithmic calculations are made..

The discrete cosine transform is calculated (DCT): In order to give emphasis to the low frequencies, DCT is chosen.

The coefficients obtained at the end are the amplitudes of the resulting spectrum.

MEL FREQUENCY CEPSTRUM COEFFICIENTS PROCESS

The speech input is typically recorded at a sampling rate more than 10 KHz. This sampling frequency was mainly chosen for minimizing the effects of aliasing in the A to D conversion i.e. analog to digital conversion. The obtained sampled signals are able to capture all the frequencies up to 5 kHz, which covers the majority energy of sounds which are generated by humans. The main reason of the MFCC processor is to copy the performance of the human ears. Additionally, instead of speech waveforms themselves, MFFC's are shown to be less susceptible to specified variations. MFCC process is shown in following figure

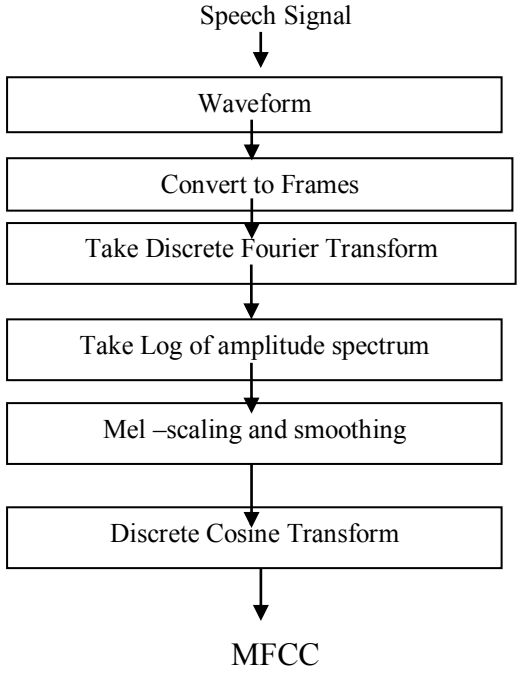

Fig 5- Process of MFCC

C. FEATURE MATCHING (DTW)

Dynamic time warping (DTW) which is used for time alignment was first introduced by Sakoe and Chiba in 1978 when it was used in conjunction with dynamic programming techniques for the recognition of isolated words. The core problem for distance measurement in speech recognition is the time alignment of different utterances. Even a small shift leads to inaccurate recognition. Dynamic Time Warping also known as DTW in short form, is efficient method to solve the time alignment problem. DTW algorithm desires at aligning two sequences of feature vectors by warping the time axis repeatedly till an optimal match between the two sequences is found. This algorithm functions a piece wise linear mapping of the time axis to align both the signals.

The timing differences between speech patterns can be removed, with the help of DTW algorithm, by straightening the time axis of one speech pattern until it maximally coincides with the other. All pattern vectors are warped against a reference pattern vector of the same category that has the same number of feature vectors, as there are frames in the input layer of the neural network. Once the relevant feature extraction has taken place, speech patterns can be processed as a sequence of feature vectors.

$x = x_1, x_2 \ldots\ldots\ldots\ldots x_K$ (7)

$y = y_1, y_2 \ldots\ldots\ldots\ldots y_M$ (8)

Two sequences are aligned on the sides of a grid, with one on the top and other on the left hand side. Both sequences starts on the bottom left of the grid. In every cell, a distance measurement is placed, comparing the respective elements of the two mentioned sequences. The distance between the two points is calculated with the help of Euclidean distance formula.

$$\text{Dist}(x, y) = |x - y|[(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots \ldots + (x_n - y_n)^2]^{1/2} \quad (9)$$

The most excellent match or alignment between these two mentioned sequences is the path through the grid, which reduces the total distance between them, which is called as Global distance. The overall distance i.e. Global distance can be calculated by finding and going directly through all the possible routes through the grid.

The global distance is the minimum distance which is calculated after addition of the distances (Euclidean distance) between the individual elements on the path divided by the sum of the weighting function. For any considerably long sequences the number of possible paths through the grid will be very large. Global distance measure is obtained using a recursive formula"

$$GD_{xy} = LD_{xy} + \min (GD_{x-1\,y-1}, GD_{x-1\,y,}\, GD_{x\,y-1}) \quad (10)$$

Here,
GD = Global Distance (overall distance)
LD = Local Distance (Euclidean distance)

The asymmetric dynamic time warping algorithm only provides compression of speech patterns. This stands for; a linear algorithm must be used with any speech patterns that need to be expanded. This is satisfactory since no feature vectors are being deleted when a linear algorithm is used in this form and there is no danger of losing important features of the speech.

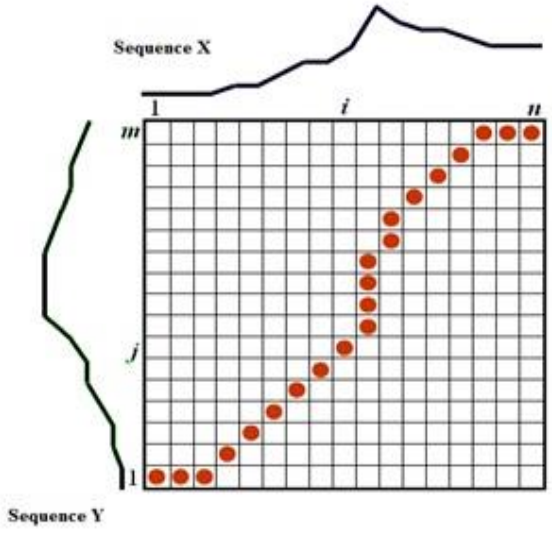

Fig. 6- Dynamic Time Wrapping of two speech samples

RESULTS

The database, which is used for experimentation, is in English Language. Recorded wave files (from speakers) are stored in the database. This database is noted as sentences. Each sentence was uttered by speaker and recorded with sampling frequency 16 kHz. Also database is collected from five different speakers as dataset. This dataset is used for speaker independent recognition recorded with same sampling frequency. Dataset recorded is used for testing purpose only.

1.SHRI GANESH

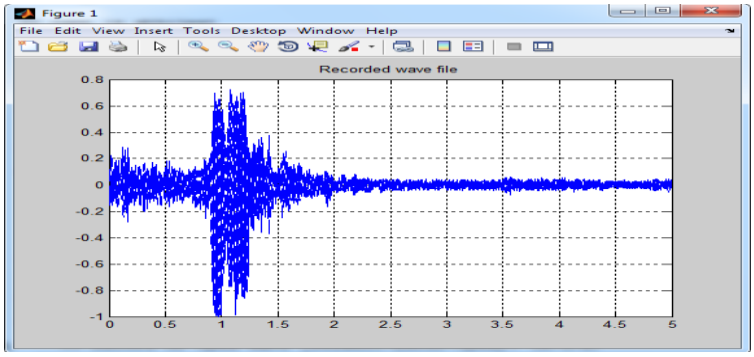

Fig 7- Recorded wave for sentence "shri ganesh"

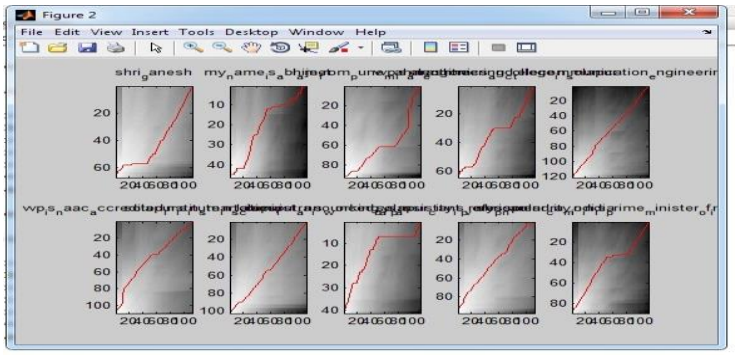

Fig. 8- DTW path taken for sentence "shri ganesh" with respect to other sentences

2. MY NAME IS ABHIJEET

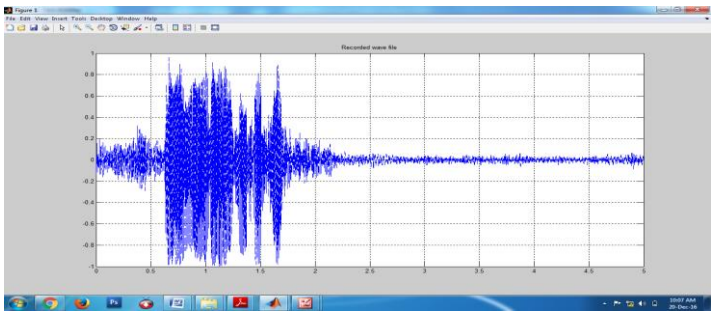

Fig 9- Recorded wave for sentence "my name is abhijeet"

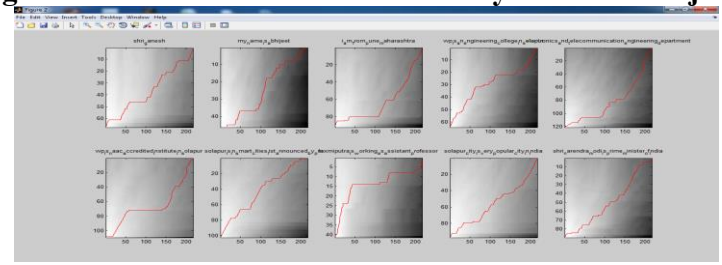

Fig. 10 - DTW path taken for sentence "my name is abhijeet" with respect to other sentences

CONCLUSION

This paper introduces a new method for speech recognition system by feature extraction method to improve the recognition accuracy and security. Many parameters can be used for speech recognition but MFCC appears to be the best method for the same. It is found to have a good recognition rate as well as to improve the accuracy of the system to implement it for number of applications. The work is planned to reveal the efficiency of MFCC for speech recognition purpose. Here we can plan for Gaussian filters instead of triangular to get the higher level of accuracy as all the elements of speech signals can be treated effectively without loss of information.

FUTURE SCOPE

The performance can further be improved by proper choice of mixing proportion of two streams in combined model. The combination of MFCC & IMFCC if used and implemented practically, the percentage accuracy can again be increased. Also if the system is designed with wireless microphone and wireless automatic system with high gain, an efficient system can be designed in future with fine accuracy.

REFERENCES

1) P.K. Sharma, B.R. Lakshmikantha and K.S. Sundar, "*Real Time Control of DC Motor Drive using Speech Recognition", Proceedings of the 2010 India International Conference on Power Electronics (IICPE), Jan. 28-30, 2011, New Delhi, India, pp. 1-5.*

2) MFCC retrieved on Jan 23rd, 2013 from, http:// practical cryptography. com/miscellaneous/ machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/Daryl Ning*, "Developing an Isolated Word Recognition System in MATLAB", article retrieved from http://www. math works.in/company/newsletters/articles/developing - an- isolated- word – recognition - system – in - matlab .html.*

3) B.P. Das, R. Parekh, "*Recognition of Isolated Words using Features based on LPC, MFCC, ZCR and STE, with Neural Network Classifiers", International Journal of Modern Engineering Research, Vol. 2, No. 3, June 2012, pp. 854-858.*

4) L. Rabiner, "*A tutorial on Hidden Marcov Model and selected applications in Speech Recognition", Proceedings of the IEEE, Vol.77, No.2, 1989, pp. 257-286.*

5) L. Muda, M.Begam and I. Elamvazuthi, "*Speech Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniq*ues", *Journal of Computing, Vol. 2, No. 3, March 2010, pp. 138-143.*

6) A. Bala, Abhijit kumar, Nidhika Birla, "*Speech Command Recognition System Based On MFCC And DTW", International Journal of Engineering Science and Technology, Vol. 2, No. 12, 2010, pp.7335-7342*

7) M.R. Hasan, M. Jamil, M.G. Rabbani and M.S. Rahman, "*Speaker Identification Using MEL Frequency Cepstral Coefficient"*, Proceedings of 3rd International conference on Electrical and Computer Engineering (ICECE), December,28-30, 2004, Dhaka, Bangladesh, pp. 565-568.

8) N.N.Lokhande , N.S.Nehe , P.S.Vikhe "*MFCC Based Robust Features for English Word Recognition" IEEE International Conference, 978-1-4673-2272-0/12/2012.*

9) Nguyen Viet Cuong, Vu Dinh, Lam Si Tung Ho, "*Mel frequency Cepstral Coefficients for Eye Movement Identification*", IEEE 2012.

10) Shivanker Dev Dhingra, Geeta Nijhawan , Poonam Pandit, "*ISOLATED SPEECH RECOGNITION USING MFCC AND DTW" , International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Aug 2013.*

11) "*Kernel-Based Feature Extraction with Speech Technology Application*" *Andras Kocsor and Laszlo Toth, Associate Member, IEEE Transactions On Signal Processing, Vol. 52, No. 8, August 2004.*

12) S. Chakroborty and G. Saha, "*Improved Text-Independent Speech recognition Using Fused MFCC and IMFCC feature Sets Based on Gaussian Filter*," International Journal of Signal Processing, Vol. 5, No. 1, 2009, pp. 11-19.

13) Alfredo Maesa, Fabio Garzia, Michele Scarpiniti, Roberto Cusani, "*Improved Text-Independent Speech recognition Using MFCC feature Sets & Gaussian Mixer Model," Journal of information security, 2012.*

14) Namrata Dave "*Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition", G H Patel College of Engineering, Gujarat Technology University, India, International Journal For Advance Research In Engineering And Technology, Volume 1, Issue VI, July 2013*