IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC) BASED SPEAKER IDENTIFICATION IN NOISY ENVIRONMENT USING LBG VECTOR QUANTIZATION

Arun Kumar Choudhary, Jitendra Kumar Mishra
PG Student, Dept. of ECE, Patel College of science and technology, Bhopal, India
Assistant professor, Dept. of ECE, Patel College of science and technology, Bhopal, India

DOI: 10.5281/zenodo.46499

ABSTRACT

Recognizing A speaker can simplify task of translating speech in systems that have been trained on specific person's voices or it can be used to the authenticate or verify the identity of a speaker as part of a security process. This work discusses Implementation of an Enhanced Speaker Recognition system using MFCC and the LBG Algorithm. The MFCC has been used the extensively for purposes of Speaker Recognition. This work has augmented the existing work by using Vector Quantization and Classification using a Linde Buzo Gray Algorithm. A complete test system has been developed using MATLAB which it can be used for real time testing as it can take the inputs directly from the Microphone. Therefore, the design can be translated into the Hardware having the necessary real time processing Prerequisites. The system has been tested using VID TIMIT Database and using the Performance metrics of False Acceptance Rate(FAR), True Acceptance Rate(TAR) and False Rejection Rate(FRR). A system has been found to perform better than the existing systems under moderately noisy conditions.

Keywords : *VQ, LBG , MFCC, Mel Frequency Wrapping , Voice recognization.*

INTRODUCTION

In this age of modern Electronic gadgets, it iswell accepted fact that most people use high end electronic devices that use natural language, whether its English or otherwise. Whether it is Apple's Siri (speech recognition software for iPhone or Microsoft's Kinect (the gaming device for Xbox360® and windows-based platforms) it seems machines can’t do without the understanding human language. However, realize that mechanism, it is a essential to improve a accuracy of the speech directed applications even in the most ordinary tasks, such as the deciding if a person is even speaking at a particular instant of time or not. Processing of human speech therefore its holds utmost importance in modern world today and finds application in various fields of Robotics, Biometrics etc[1].

Speaker Recognition: Speaker Recognition is now possible using the range of different approaches each with costs and benefits. The SR is a very important activity research today encompasses the range of a difference approaches and for this reason there has been a classification of a approaches into classes. The SR approach classes are the[2]:

1. Conventional.
   a. Speaker identification
   b. Speaker verification
2. Text Conversion.
   a. Text independent recognition
   b. Text dependent recognition

Speaker Identification

A Speaker identification is defined as a process of determining which speaker provides a given utterance. speaker is registered into a database of speakers and utterances are added to the database that may be used at later time during the speaker identification process. The speaker identification process is shown in Figure 1. A steps as shown in Figure1 include the feature extraction from the input speech, a measure of similarity from the available speaker utterances and the decision step that identifies the speaker identification based upon the closest match algorithm used in previous step[3].

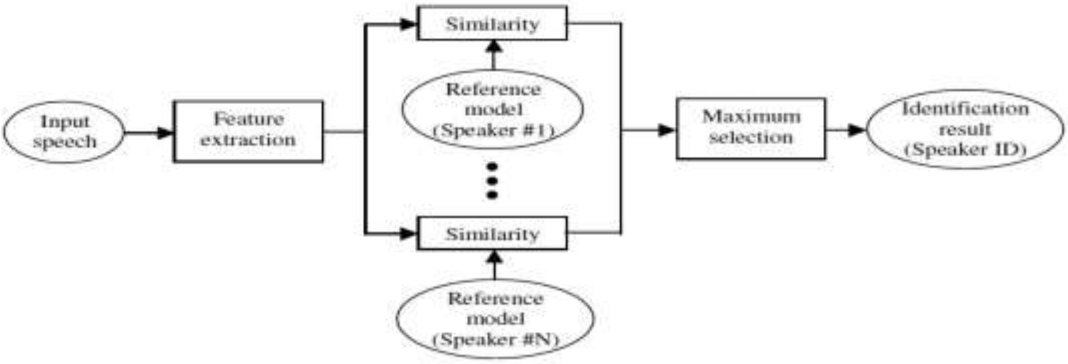
Input speech
Feature extraction
Similarity
Reference model (Speaker #1)
Similarity
Reference model (Speaker #N)
Maximum selection
Identification result (Speaker ID)

Fig 1 Speaker Identification

Speaker Verification

The acceptance or rejection of an identity claimed by the speaker is known as the Speaker Verification. The speaker verification process is shown in Figure 2 and includes the feature extraction from the source speech, comparison with speech utterances stored in a database from the speaker whose identity is now being claimed and a decision step that provides a positive or the negative outcome.

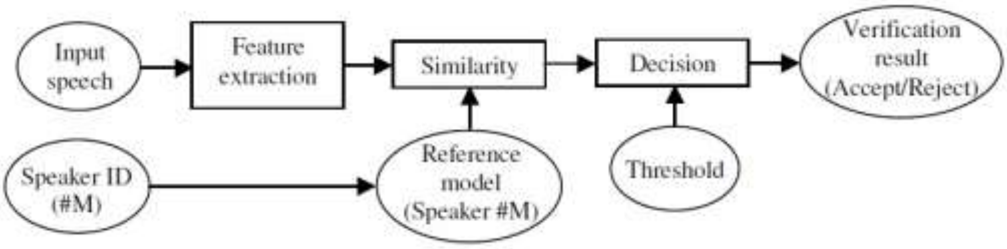

Fig 2 Speaker verification

Text-independent Recognition

In Figure 3, text-independent SR system as shown where the key feature of the system is speaker identification utilizing the random utterance input speech.

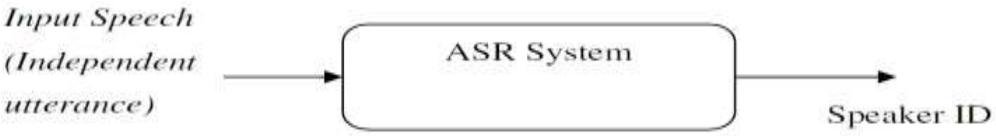

Fig 3 Text-independent Speaker Recognition

Text-dependent Recognition

In Figure 4 , a text-dependent SR system is shown where recognition of the speaker's identity is based on a match with the utterances made by the speaker previously and stored for later comparison. Phrases like a passwords, card numbers, PIN codes, etc. made be used.

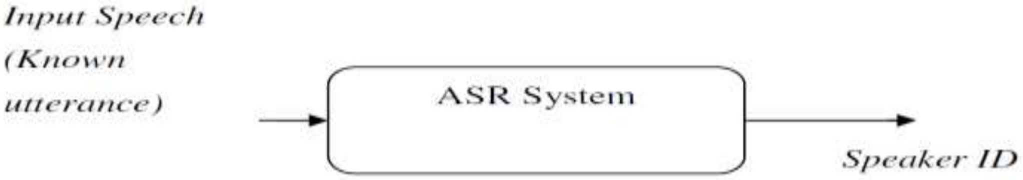

Fig 4 Text dependent Speaker Recognition

PATTERN RECOGNITION BASED VECTOR QUANTIZATION ALGORITHM

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

Therefore, The LBG algorithm proposed by the Linde, Buzo, and Gray is chosen. After taking the enormous number of feature vectors and approximating them with.

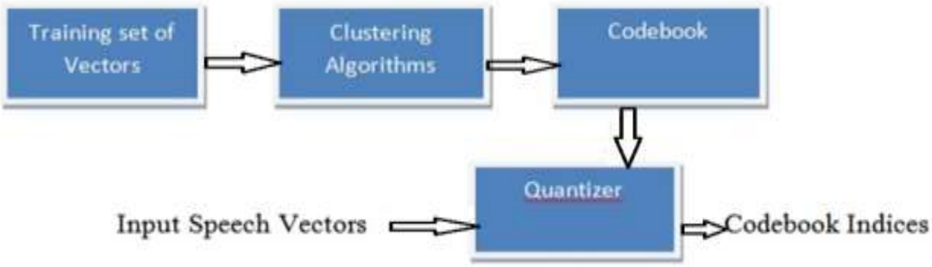

Fig 5 Block Diagram of the basic VQ Training and classification structure.

The smaller number of vectors, all of these vectors are filed away into a codebook, which is referred to as codeword's. The result of a feature extraction is a series of vector characteristics of a time varying spectral properties of the speech signal. These vectors are 24 dimensional and are a continuous. These can be mapped to the discrete vectors by the quantizing. However, as vectors are quantized, this is termed as Vector Quantization. VQ is potentially an extremely efficient representation of spectral information in the speech signal[4].

ASSIGN METHODOLOGY

The block diagram showing the computation of MFCC is shown in a Fig6.

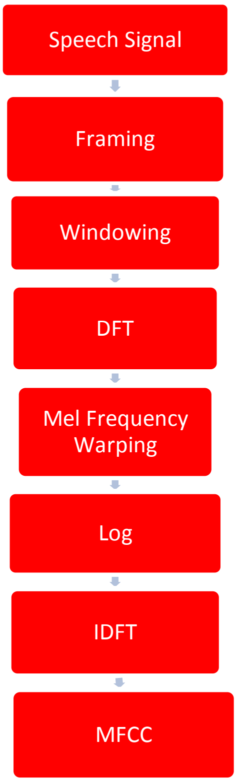

Fig 6 computation of MFCC is shown

MFCCs are obtained as follows [5].
1. Take a Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of a spectrum obtained above onto the Mel scale, using triangular overlapping windows.
3. Take logs of the powers at each of the Mel frequencies.
4. Take a discrete cosine transform of the list of Mel log powers, as if it were a signal.
5. The MFCCs are amplitudes of the resulting spectrum.
Speech signals are the normally pre-processed before features are extracted to enhance a accuracy and efficiency of extraction processes. A Speech signal pre-processing covers the digital filtering and speech signal detection. Filtering includes the pre-emphasis filter and filtering out any surrounding noise.

MEL-FREQUENCY CEPSTRUM COEFFICIENTS PROCESSOR

A block diagram of the structure of an MFCC processor is given in the Figure7. A speech input is typically recorded at a sampling rate above 12500 Hz. This sampling frequency was chosen to a minimize the effects of aliasing in analog to- digital conversion.

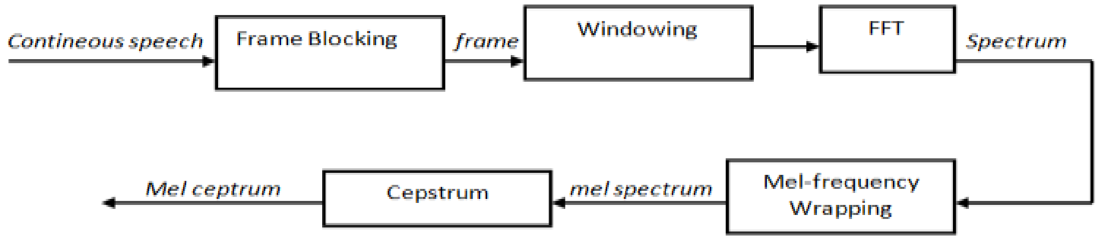

Fig 7 Block diagram of the MFCC processor

A- Frame Blocking:-, A continuous speech signal is blocked into the frames of N samples, with adjacent frames being separated by the M (M < N). The first frame consists of a first N samples. The second frame begins M samples after the first frame, and the overlaps it by N – M samples. Similarly, third frame begins 2M samples after the first frame (or M samples after the second frame) and the overlaps it by N - 2M samples. This process continues until all the speech is accounted for within one or more the frames

Frame blocking of the speech signal is done because when examined over a sufficiently short period of the time (between 5 and 100 msec), its characteristics are fairly stationary. However, over the long periods of time (on the order of 1/5 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken. A Overlapping frames are taken not to have much information loss and to maintain correlation between the adjacent frames

B.-Windowing

A next step in the processing is to window each individual frame so as to minimize signal discontinuities at the beginning and end of each frame. The concept here is to minimize spectral distortion by using the window to taper a signal to zero at the beginning and end of each frame. If we define the window asw(n), 0 ≤ n ≤ N −1, where N is number of samples in each frame, then the result of windowing is the signal

$$y_l(n) = x_l(n)w(n), \quad 0 \le n \le N-1 \qquad (1)$$

Typically the Hamming window is used, which has the form and plot is given in

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

C- Fast Fourier Transform:-

The third step is to the apply discrete Fourier transform on each frame. A fastest way to calculate a DFT is to use FFT which is an algorithm that can speed up DFT calculations by a hundred-folds.The resulting spectrum is then converted into mel scale[6].

D - Mel-frequency Wrapping:- psychophysical studies have shown that human perception of a frequency contents of sounds for the speech signals does not follow a linear scale. Thus for each tone with actual frequency, f, measured in Hz, the subjective pitch is measured on a scale called the ‘Mel’ scale. A Mel-frequency scale is linear frequency spacing below the 1000 Hz and a logarithmic spacing above 1000 Hz. As a

reference point, the pitch of a 1 kHz tone, 40 dB above a perceptual hearing threshold, is defined as 1000 mels [7]. Therefore we can use the following approximate formula to compute a mels for a given frequency f in Hz:
*mel(f) = 2595*log10(1+ f / 700)* (2)
One approach to simulating the subjective spectrum is to use a filter bank, spaced uniformly on the Mel scale . That filter bank has triangular band pass frequency response, and spacing as well as the bandwidth is determined by a constant Mel frequency interval. Modified spectrum of S(ω) thus consists of the output power of these filters when S(ω) is the input. The number of the Mel spectrum coefficients, K, is typically chosen as 20. This filter bank is applied in frequency domain, therefore its simply amounts to taking those triangle-shape windows in the spectrum. A useful way of thinking about the Mel-wrapping filter bank is to view each filter as an histogram bin (where bins have overlap) in the frequency domain.

E- Cepstrum:-

IN a final step, A log Mel spectrum has to be converted back to time. The result is called a Mel frequency cepstrum coefficients(MFCCs). Cepstral representation of the speech spectrum provides a good representation of a local spectral properties of a signal for the given frame analysis. Because the Mel spectrum coefficients are a real numbers (and so are their logarithms), they may be converted in time domain using a Discrete Cosine Transform (DCT). MFCC may be calculated using a equation-

$$C_s(n,m) = \sum_{i=1}^{M} \log Y(i) Cos\, [\frac{i2\pi}{N}\, n] \qquad (3)$$

where N' is a the number of points used to compute standard DFT[8].

RESULTS ANALYSIS AND DISCUSSION

Testing Database (VidTIMIT Database)

Vidtimit database is comprised of the video and corresponding audio recordings of the 43 volunteers (19 female and 24 male), reciting the short sentences. it was recorded in 3 sessions, with the mean delay of 7 days between session 1 and 2, and the 6 days between session 2 and 3. A delay between sessions allows for the changes in the voice, hair style, make-up, clothing and mood.

Speaker recognition system performance is measured using the various metrics such as recognition or acceptance rate and the rejection rate. Recognition rate deals with a number of genuine speakers correctly identified, whereas rejection rate corresponds to the number of the imposters (people falsifying other's identity) being rejected.

, 1 True Acceptance Rate (TAR) - The rate at which a legitimate speaker is accepted
2 False Rejection Rate (FRR) - The rate at which a legitimate speaker is rejected (FRR=1-TAR)
3 False Acceptance Rate (FAR) - The rate at which an imposter is accepted as a legitimate speaker

RESULTS ANALYSIS:-

True Acceptance Rate (TAR):- In below Fig(8) shows the True acceptance rate which has been tested on a similar database and has been found to have an accuracy of roughly 96%. However the acceptance rate shows a decline if the no of user inputs are from microphone owing to noisy environments

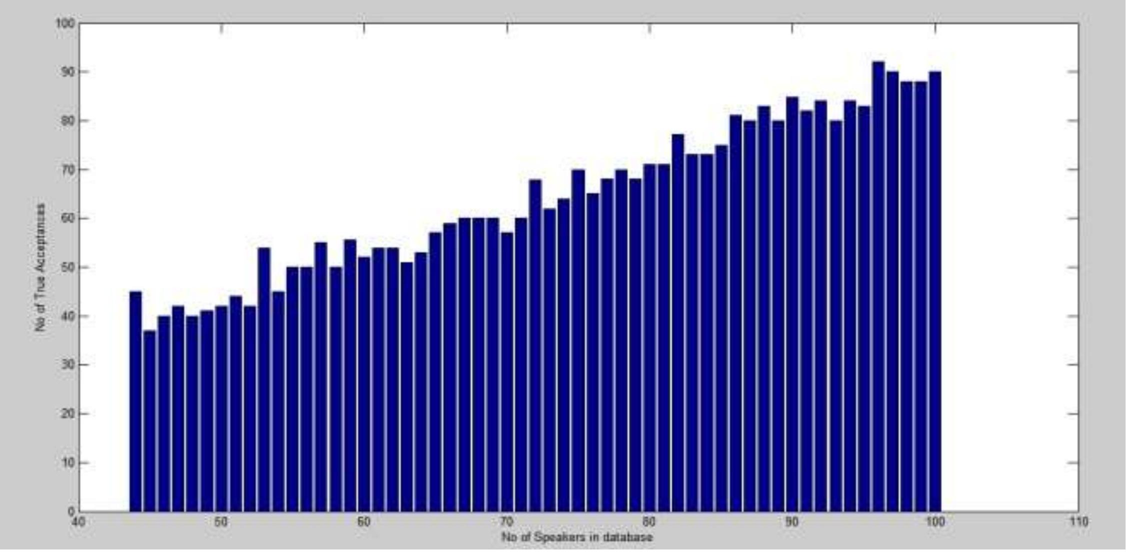

Fig8 True acceptance rate results

False Rejection Rate (FRR):-In below Fig(9) shows false rejection rate which is roughly 15% where legitimate speakers have been rejected as Imposters. The dataset has again been varied from 44 to 100 users.

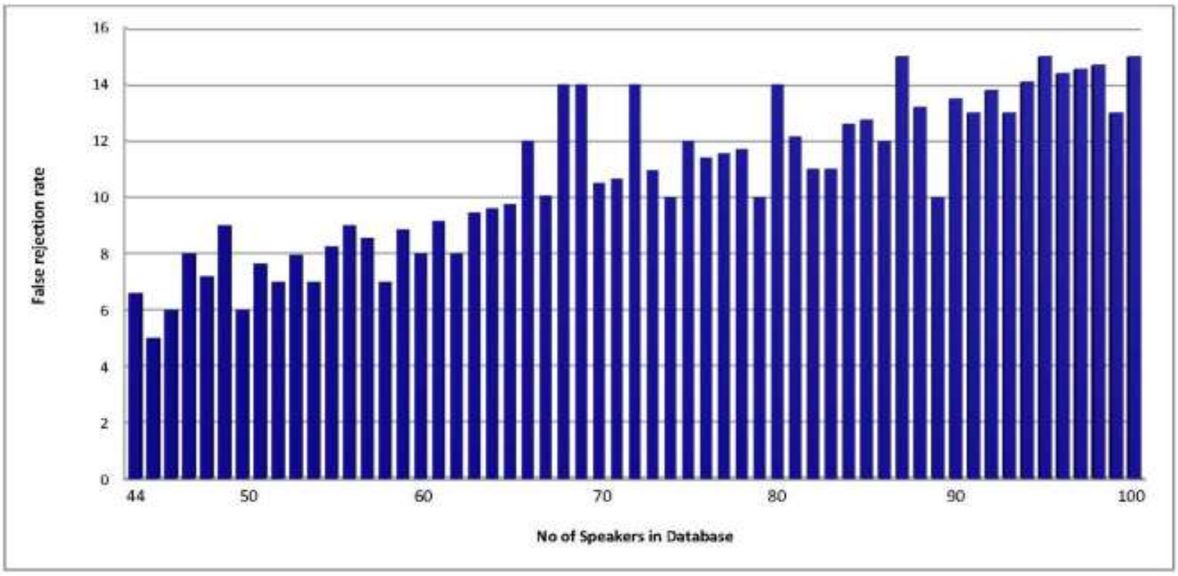

Fig 9False rejection rate results

False Acceptance Rate (FAR):- FAR Results have been shown in below fig(10) . As we can see that the False acceptance rate over a test set ranging from 44 to 100 users is roughly 9%.

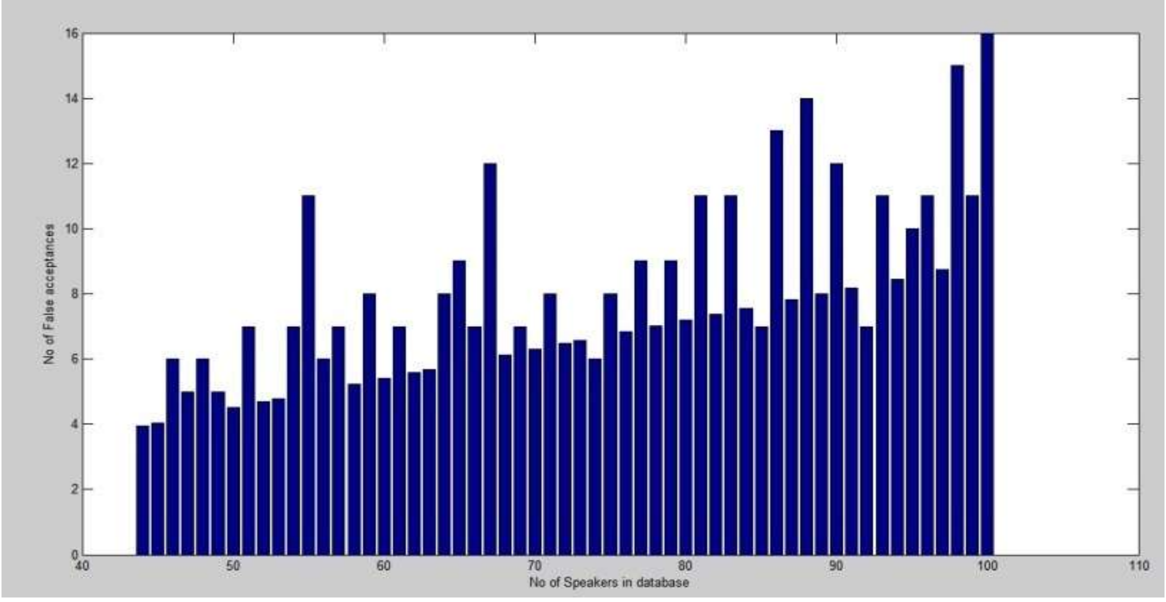

Fig 10 FAR for 44 to 100 user database

MEL FREQUENCY COEFFICIENTS ANALYSIS:- The Mel frequency wrapping and the calculation of MEL Frequency coefficients. This is done by using filter bank, spaced uniformly on the Mel scale. That filter bank has a triangular band pass frequency response, and the spacing as well as bandwidth is determined by a constant Mel frequency interval. The modified spectrum of the S(ω) thus consists of the output power of these filters when S(ω) is the input. A number of Mel spectrum coefficients, K, is typically chosen as the 20.

The filter bank is applied in the frequency domain, therefore its simply amounts to taking those triangle-shape windows in the Figure(11) on the spectrum. The useful way of thinking about this Mel-wrapping filter bank is to the view each filter as an histogram bin (where bins have overlap) in the frequency domain. Fig(12)shows the unmodified original spectrum and the spectrum which has been modified as a result of cepstrum analysis

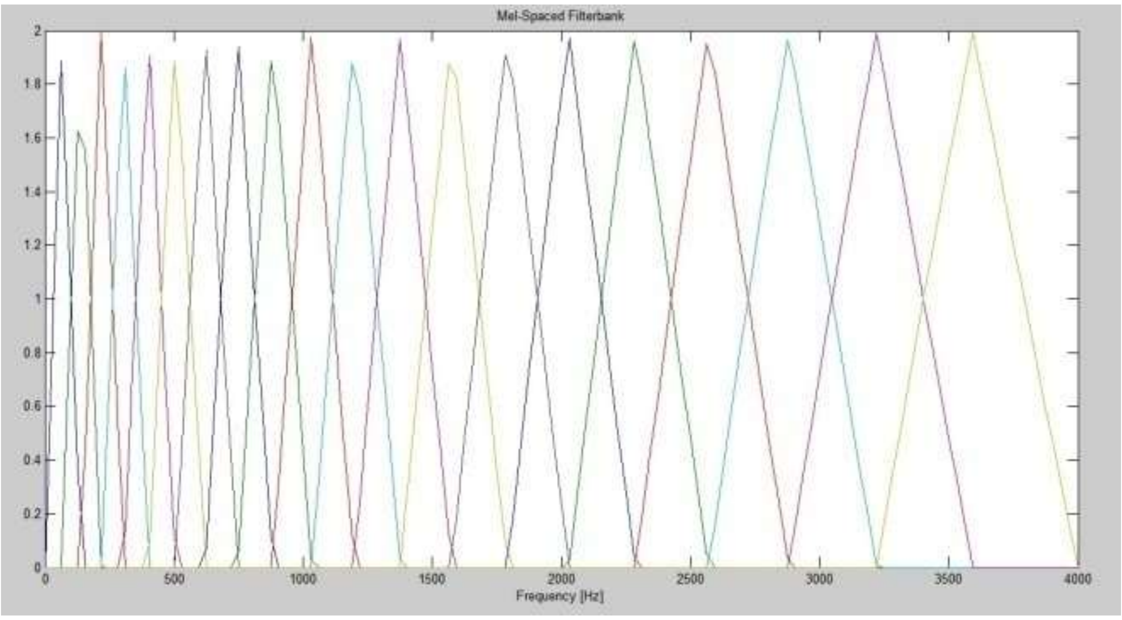

Fig. 11 Output from Mel filter bank

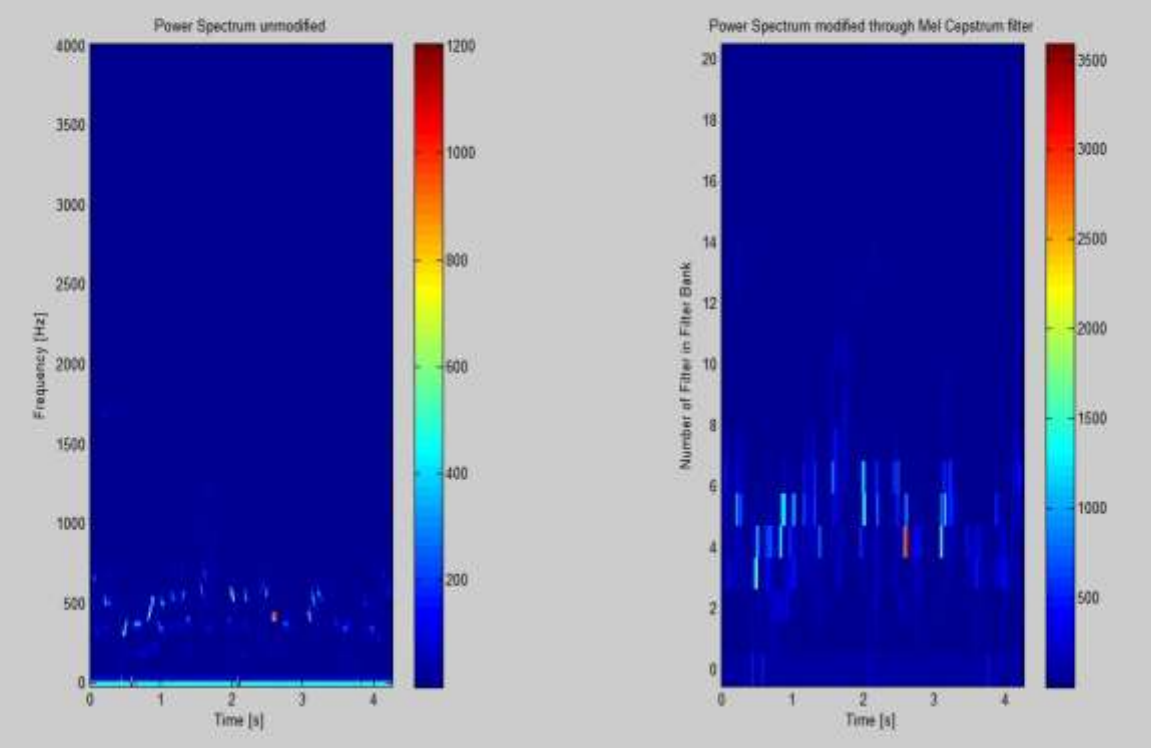

Fig .12.(a) Unmodified Spectrum(b) Power Spectrum after Mel Cepstrum filter

Windowing and Fast Fourier Transform :-

After windowing, A next processing step is the Fast Fourier Transform, which converts each of the frame of N samples from the time domain into the frequency domain. A result after this step is often referred to as spectrum or period gram. It has been shown in figs 13 (a) The figures are listed for M=100 and N=256. The logarithmic power spectrum has also been shown in fig 13(b)

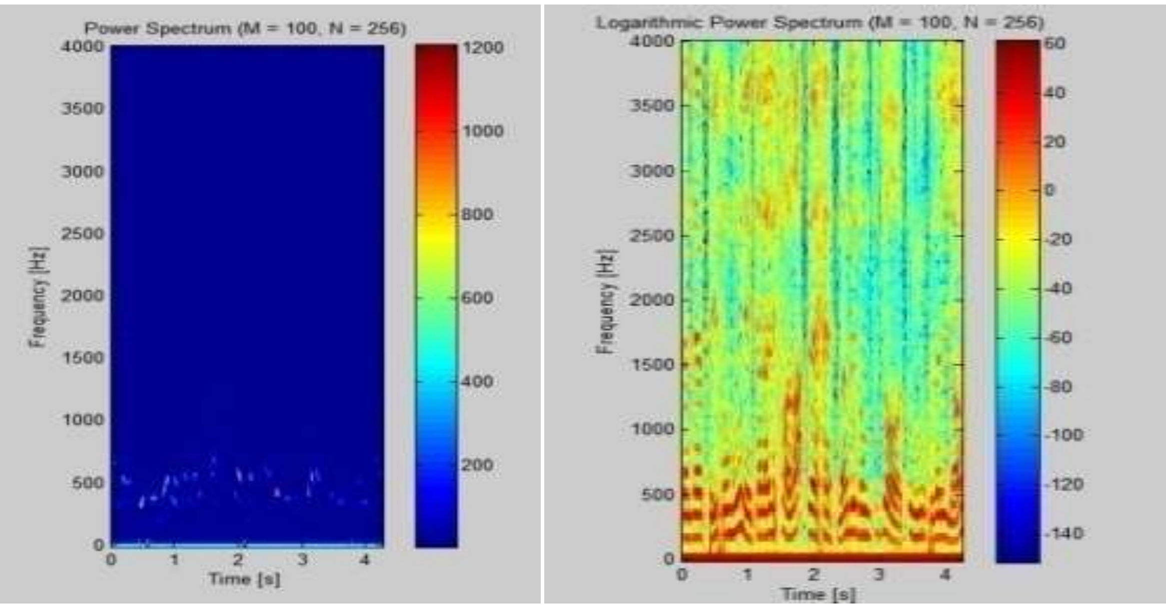

Fig.13(a) Power Spectrum (b) Logarithmic Power Spectrum

Speaker Amplitude Plots
Amplitude Plots

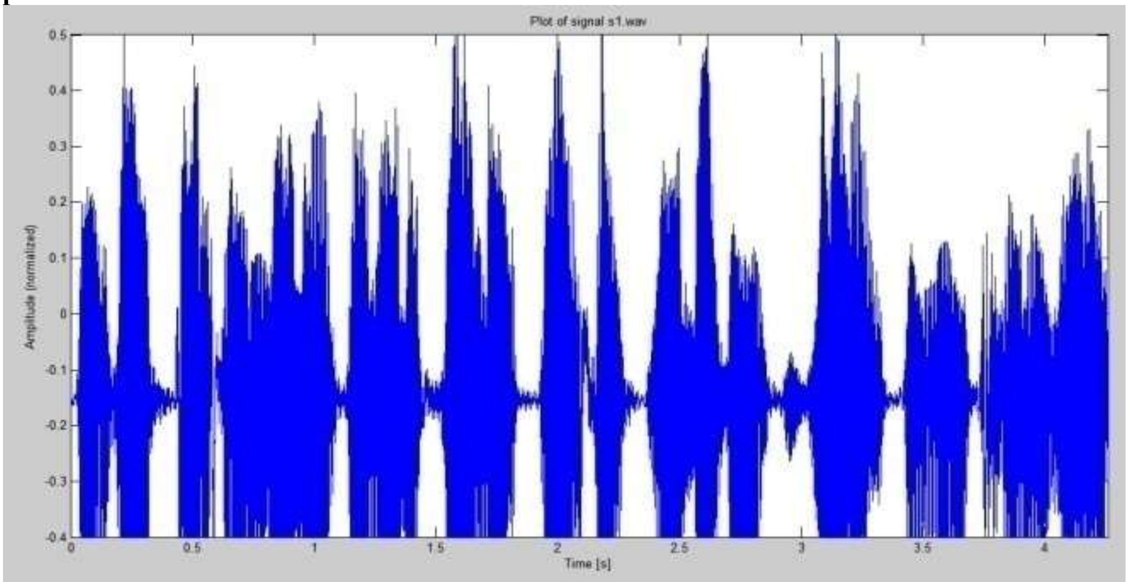

Fig.14 Speech Signal 1 from testing database VID TIMIT

The figures show the amplitude plots of the some of the chosen legitimate speakers and Imposters. The speech signal is a slowly time varying signal. When examined over a sufficiently short period of the time (between 5 and 100 msec), its characteristics are fairly stationary. However, over a longer periods of time (on the order of 1/5 seconds or more) a signal characteristics change to reflect a different speech sounds being spoken. Therefore, short-time spectral analysis is a most common way to characterize the speech signal

.

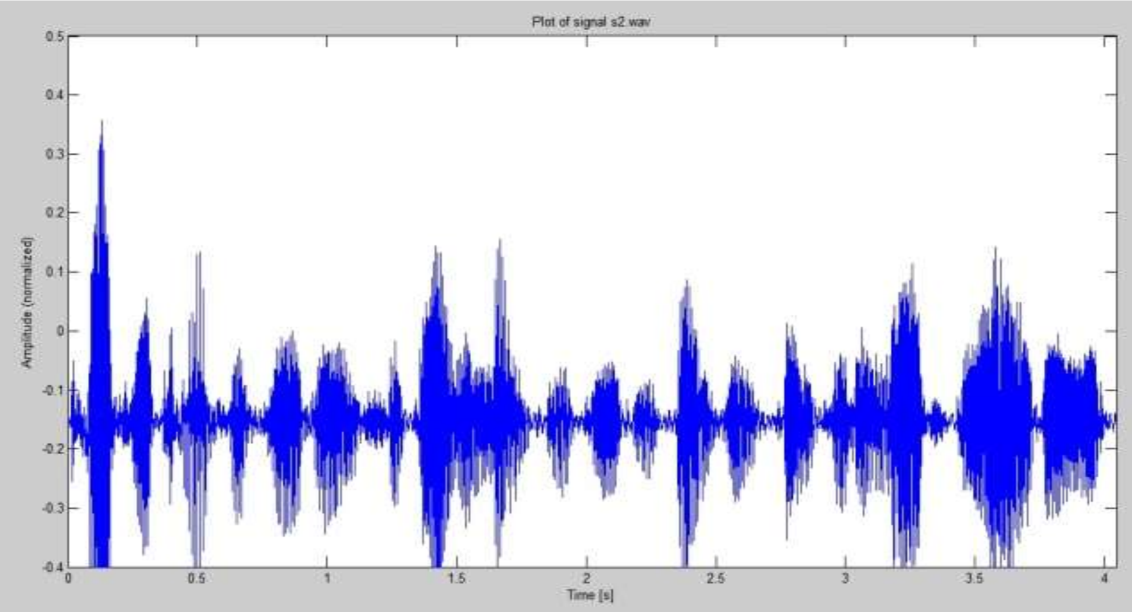

Fig.15 Speech Signal 2 from Testing Database VID TIMIT

2D Acoustic Vectors:-

The resulted acoustic vectors i.e. the Mel Frequency Cepstral Coefficients corresponding to fifth and sixth filters were plotted in the following figure:-

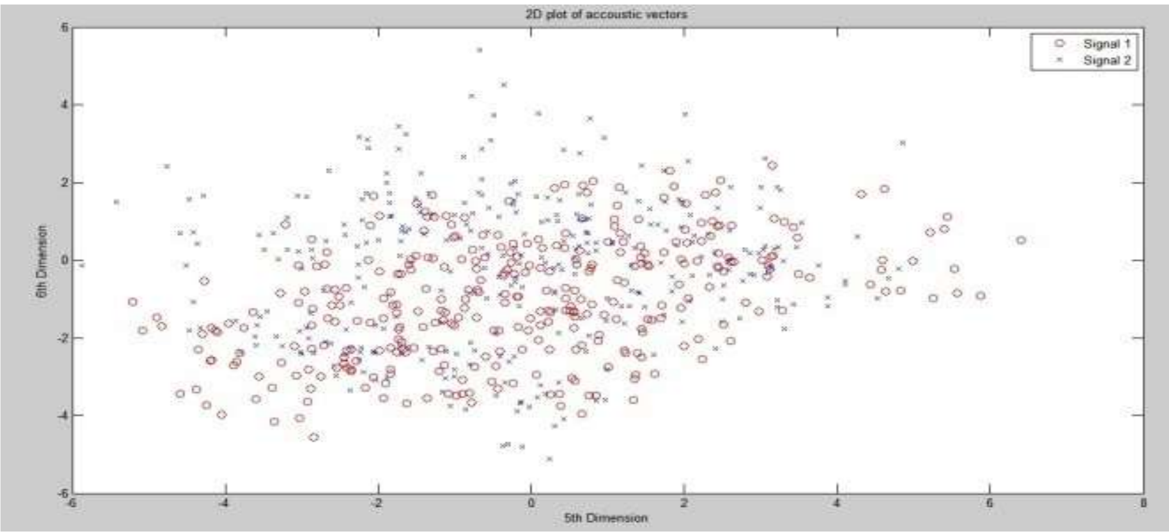

Fig.16 Plot of 2D Acoustic Vectors

Vector Quantization Based on Linde Buzo Gray Algorithm

Finally on the application of VQ, we get a set of the 2D trained VQ code words. The figure(17) shows a 2D Plot of the 2D trained VQ codeword corresponding to the fifth and sixth dimension.

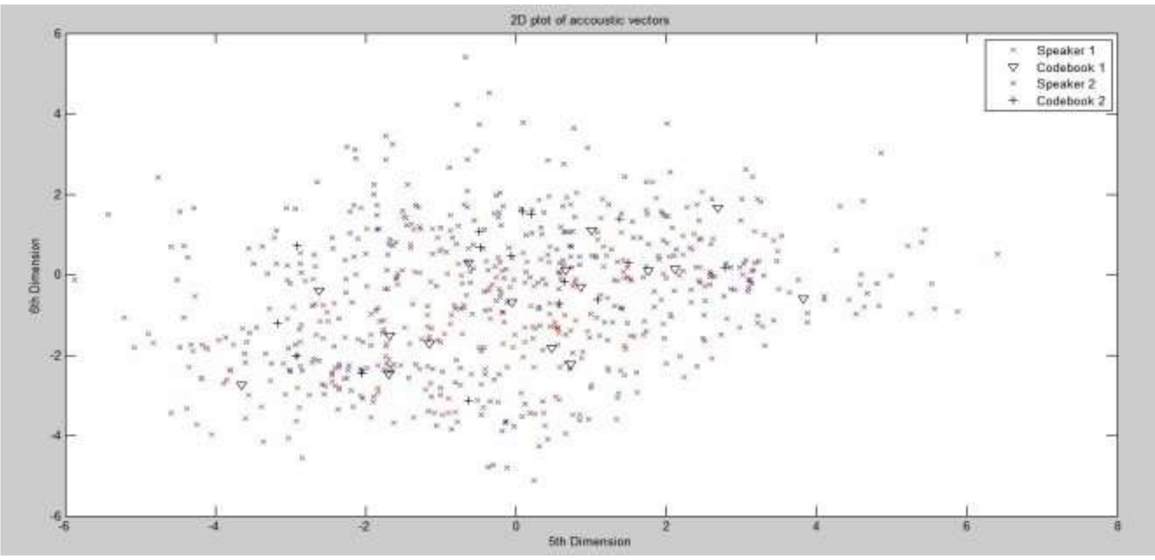

Fig.17 2D plot of Trained VQ Code words

COMPARISON ANALYSIS OF PRPOSED WORK:

Table

PARAMETERS	PAST WORK	PRESENT WORK
Algorithms Used	MFCC	MFCC with LBG(Vector Quantization)
Performance	Poor at noisy condition	Good at noisy condition
False rejection rate	More than 10%	15%
True acceptance rate	85%	96%
False acceptance rate	More than 15 %	Less than 9%
Compatibility	Poor	Good
No. of User	44 to 85 user testified	44 to 100 used testified
Self generated inputs	15 inputs	20 input
Drawback	Speaker and microphone both has been done in acoustic silent environment.	Testing has been done using standard microphone in acoustic silent environment.
Data base	Zdelcoul database	VID TIMIT Database

CONCLUSION

This work has presented an enhanced mechanism of Speaker Recognition using a combination of a well known MFCC algorithm as well as the LBG algorithm for generating the vector code words. In which training and testing was done on the VID TIMIT database and the system was found to perform efficiently as is visible from the False Acceptance Rate (FAR) True Acceptance Rate (TAR) , False Rejection Rate (FRR) results. The testing has been done by using standard Microphones in acoustically silent environments and then additional vq has been added for noise simulations. The GUI developed for the purpose has capabilities of real time speaker recognition, making it a significant contribution to the work.

The work has been simulated and tested using MATLAB R 2012. Although the GUI that has been developed takes inputs in real time, however the performance of the system needs to be tested on a Hardware platform F2812 Floating point Processor for its actual real time performance to be verified

REFERENCES

1. Douglas A Reynolds "An Overview of Speaker Recognition Technology", MIT Lincoln Laboratory, MA 2002
2. Campbell Jr., J.P., 1997. Speaker recognition: A tutorial. Proceedings of the IEEE, 85(9), pp.1437–1462
3. R. Rabiner,B.-H. Juang, C.-H. Lee ,An Overview of Automatic Speech Recognition Automatic Speech and Speaker Recognition, The Kluwer International Series in Engineering and Computer Science Volume 355, 1996, pp 1-30
4. Ezzaidi, H., Rouat, J., and O'Shaughnessy, D. Towards combining pitch and MFCC for speaker identication systems. In Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001) (Aalborg, Denmark, September 2001), pp. 2825– 2828
5. Todor Ganchev , Nikos Fakotakis , George Kokkinakis Comparative evaluation of various MFCC implementations on the speaker verification task (2005),
6. R. Sambur A, . E. RosenbergL, . R. Rabinera, and C. A. McGonegal "On reducing the buzz in LPC synthesis", M Bell Laboratories
7. Speaker segmentation and clustering (2008) M Kotti, V Moschou, C Kotropoulos., 2008"
8. From Frequency to Quefrency: A History of the Cepstrum Alan V. Oppenheim and Ronald W. Schafer, IEEE Signal Processing Magazine Reprinted 2004

	Mr. Jitendra Kumar Mishra, Assistance Professor ,Dept of ECE PCST BHOPAL
	Mr. Arun Kumar Choudhary PG Student(Digital Communication ,Pcst Bhopal)