International Journal of Electrical and Electronics Engineers
Vol. No.7 Issue 02, July-December 2015
www.arresearchpublication.com
IJEEE
ISSN 2321 - 2055

ADVANCED VOICE CHARACTERIZATION USING MEL FILTERS AND LBG VECTOR QUANTIZATION

Arun kumar Choudhary[1], Jitendra Kumar Mishra[2]

[1]PG Student, [2]Assistant Professor, Dept. of ECE, Patel College of Science and Technology, Bhopal, (India)

ABSTRACT

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

***Keywords*: LBG, MFCC, Mel Frequency Wrapping, Voice Recognization, VQ,**

I. INTRODUCTION

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

II. SPEAKER AND SPEECH RECOGNITION

Speaker Recognition (SR) is a major topic which includes many different speaker specific tasks. According to Reynolds (2002)[1], the tasks can be sub categorized into text dependent (where speakers are expected to utter a certain piece of text) and text independent (where the speaker may speak anything they wish) tasks. Similarly,

depending on the information that a method is allowed to use and the output expected from the process; speaker recognition generally comprises of the listed tasks

2.1 Speaker Identification: A closed set of speakers is presented to the system along with the testing data. The system must decide who, among the available set of speakers resembles or matches the testing data. This is often referred to as Closed-Set Identification to avoid confusion with a verification task, or more conveniently speaker identification.

2.2 Speaker Verification: Two pieces of speech are presented to the system. The system must decide whether the same speaker spoke both segments or they were two different speakers. This is often referred to Open-Set Identification. Campbell (1997)[2] adds the following task under the SR umbrella:

2.3 Speaker Detection: One speaker's data (often called a target speaker) offered to the system along with many testing speeches. The system is expected to correctly flag the speeches of a target speaker. Other tasks are also related to Speaker Recognition, as they are considered of the same family of research (Kotti et al, 2008)[3]:

2.4 Speaker Segmentation: A large input stream, with more than one speaker present, is offered to a system. The system is expected to find the points where the speaker changes; i.e. turn points. If the knowledge about the speakers is available beforehand, then a system can build models for each of the speaker. Then the task is called the model-based speaker segmentation. Otherwise, it is called blind speaker segmentation, or metric-based speaker segmentation.

2.5 Speaker Clustering: A large number of test inputs are presented to the system. The system must correctly cluster them according to a speaker. This task is often done online, alongside another task, as to group segments of the same speaker together.

2.6 Speaker Diarization : A stream is presented to the system. The system is expected to decide who is the speaking at each period of the stream. This task is often thought of as segmentation of the stream followed by clustering. Similar to the segmentation task, if the knowledge is available a priori to the system then models can be the built (which helps in the online clustering as well) and task is called model-based speaker diarization.

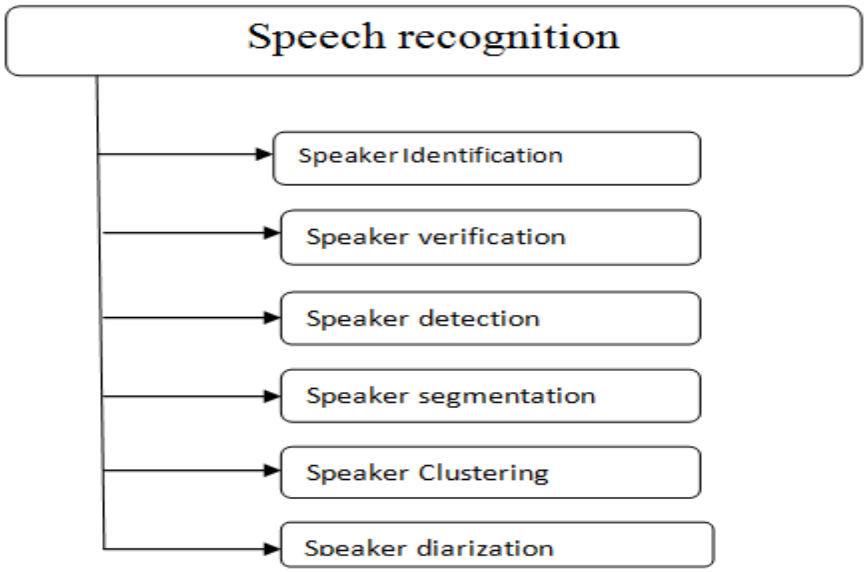

Fig 1. Speaker Recognition Tasks

III. SPEAKER RECOGNITION CLASSES

Speaker Recognition Classes is now possible using a range of different approaches each with costs and benefits. As 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:

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

IV. PROPOSED METHODOLOGY

The block diagram showing the computation of MFCC is shown in theFig2.

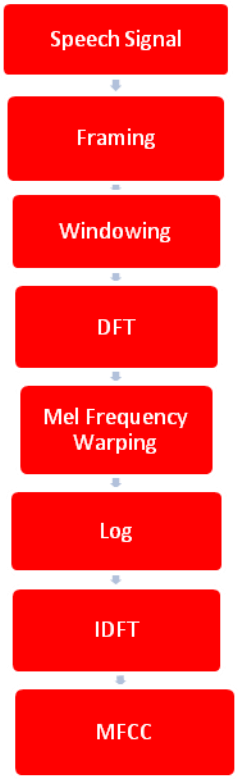

Fig 2 Computation of MFCC is Shown

MFCCs are obtained as the follows [4],[5]:

1. Take a Fourier transform of (a windowed excerpt of) a signal.
2. Map a powers of the spectrum obtained above onto the Mel scale, using triangular overlapping windows.
3. Take a logs of the powers at each of the Mel frequencies.
4. Take discrete cosine transform of the list of Mel log powers, as if it were a signal.
5. The MFCCs are a amplitudes of the resulting spectrum.

Speech signals are the normally pre-processed before features are extracted to enhance a accuracy and the efficiency of extraction processes. A Speech signal pre-processing covers digital filtering and speech signal detection. Filtering includes the pre-emphasis filter and filtering out any surrounding noise.

V. DESCRIPTION OF MFCC GENERATION STEPS

5.1 Pre-emphasis Filter

In general, A digitized speech waveform has a high dynamic range and suffers from the additive noise. In order to reduce the range, pre-emphasis is applied. This pre-emphasis is done by using the first-order FIR high-pass filter .In a time domain, with input x[n], a filter equation $y[n] = x[n] - a\, x[n-1]$ where $0.9 \leq a \leq 1.0$ and a transfer function of the FIR filter in z-domain is $H(Z) = 1 - \alpha z^{-1}$, $0.9 \leq \alpha \leq 1.0$ (3) where α is pre-emphasis parameter .The aim of this stage is to boost the amount of energy in high frequencies. Boosting a high frequency energy makes information from these higher formants available to the acoustic model. The pre-emphasis filter is applied on input signal before windowing.

5.2 Framing and Windowing

A first step is framing. The speech signal is split up into the frames typically with the length of 10 to 30 milliseconds. A frame length is important due to a trade off between time and frequency resolution. If it is too long it will not be able to capture local spectral properties and if it is too short the frequency resolution would degrade. The frames overlap each other typically by 25% to 70% of their own length.

A reason for this is because on each individual frame, we will also be applying the hamming window which will get rid of some of the information at beginning and end of each frame. Overlapping will then the reincorporate this information back into the our extracted features.

5.3 Windowing

A Windowing is performed to avoid unnatural discontinuities in speech segment and distortion in the underlying spectrum . A good window function has the narrow main lobe and low side lobe levels in their transfer function. The multiplication of a speech wave by the window function has the two effects:

- It gradually attenuates the amplitude at both ends of extraction interval to prevent an abrupt change at the endpoints.
- It produces convolution for the Fourier transform of the window function and the speech spectrum

  We have used a hamming window in our speaker recognition system. In speaker recognition, the most commonly used window shape is a hamming window . The hamming window $W_H(n)$, is defined as the

  $W_H(n) = 0.54 - 0.46 \cos(2n\pi/N-1)$ "equation 1"

5.4 Fast Fourier Transform

The third step is to 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 the hundred-folds[6]The resulting spectrum is then converted into mel scale.

5.5 Mel-scaled Filter Bank

One approach to simulating subjective spectrum is to use a filter bank, A one filter for each desired Mel frequency component. The filter bank has a triangular band pass frequency response. The spacing as well as bandwidth is determined by a constant Mel-frequency interval.

A information carried by low frequency components of a speech signal is more important compared to the high frequency components. In order to place more emphasis on a low frequency components, Mel scaling is performed. Mel filter banks are non-uniformly spaced on the frequency axis, so we have more filters in the low frequency regions and less number of filters in the high frequency regions. The Mel frequency warping is normally realized by triangular filter banks as shown in Fig. 4 with the centre frequency of the filter normally evenly spaced on a frequency axis[7]. The warped axis is implemented according to equation the (1) so as to mimic the human ears perception. A output of the ith filter is given by-

$$Y(i) = \sum_{j=1}^{N} s(j)\,\Omega_i(j)$$ "equation 2"

s(j) is the N-point magnitude of the spectrum (j =1:N) and Ωi(j) is the sampled magnitude response of a M-channel filter bank (i =1:M).

5.6 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). A cepstral representation of the speech spectrum provides a good representation of a local spectral properties of the signal for the given frame analysis. Because the Mel spectrum coefficients are the real numbers (and so are their logarithms), they may be converted to the time domain using a Discrete Cosine Transform (DCT).

The MFCC may be calculated using a equation-

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

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

VI. PATTERN RECOGNITION USING VECTOR QUANTIZATION ALGORITHM

A problem of speaker recognition belongs to a much broader topic in scientific and engineering so called the pattern recognition. A goal of pattern recognition is to classify the objects of interest into one of a number of categories or classes. A objects of interest are generically called the patterns and in our case are sequences of acoustic vectors that are extracted from the input speech using the techniques described in a previous section. The classes here refer to a 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 Linde, Buzo, and Gray is chosen. After taking the enormous number of feature vectors and approximating them with.

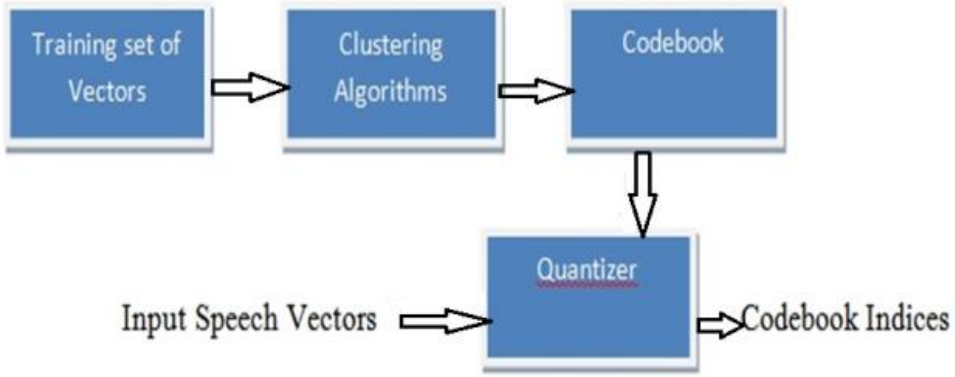

Fig 3 Block Diagram of the Basic VQ Training and Classification Structure

The smaller number of vectors, all of these vectors are the filed away into a codebook, which is referred to as codeword's. The result of the 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 the continuous. These can be mapped to the discrete vectors by quantizing. However, as vectors are quantized, this is termed as Vector Quantization. VQ is the potentially an extremely efficient representation of spectral information in the speech signal.

VII. SIMULATION RESULTS AND DISCUSSION

7.1 Testing Database

VidTIMIT Database

The Vidtimit database is comprised of video and corresponding audio recordings of the 43 volunteers (19 female and 24 male), reciting short sentences. it was recorded in 3 sessions, with the mean delay of 7 days between session 1 and 2, and 6 days between session 2 and 3. the 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 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 False Acceptance Rate (FAR) - The rate at which an imposter is accepted as a legitimate speaker,

2 True Acceptance Rate (TAR) - The rate at which a legitimate speaker is accepted

3 False Rejection Rate (FRR) - The rate at which a legitimate speaker is rejected (FRR=1-TAR)

International Journal of Electrical and Electronics Engineers
Vol. No.7 Issue 02, July-December 2015
www.arresearchpublication.com
IJEEE
ISSN 2321 - 2055

VIII. RESULTS

Amplitude Plots

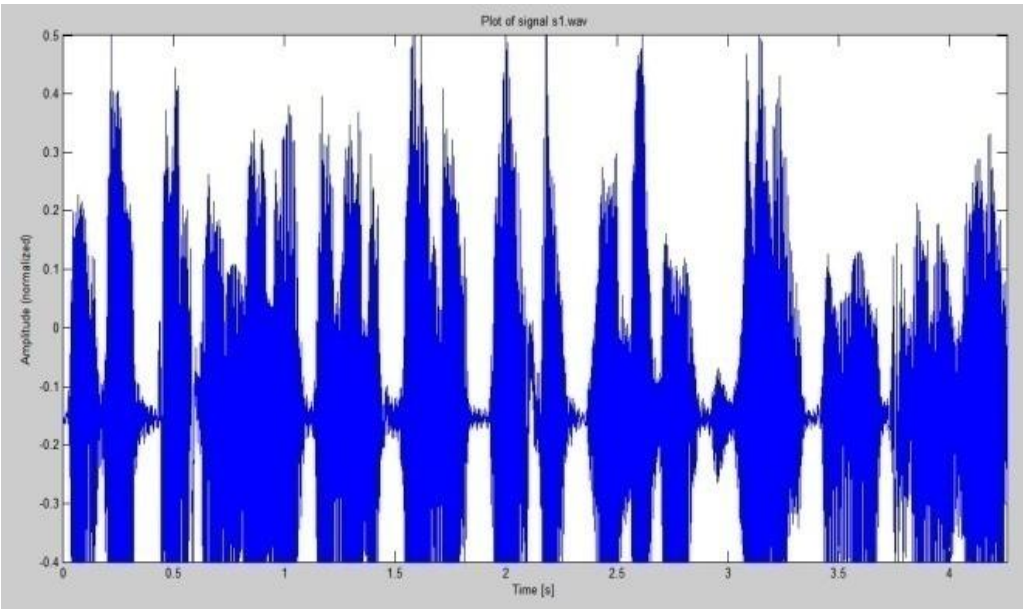

Fig. 4 Speech Signal 1 from testing database VID TIMIT

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

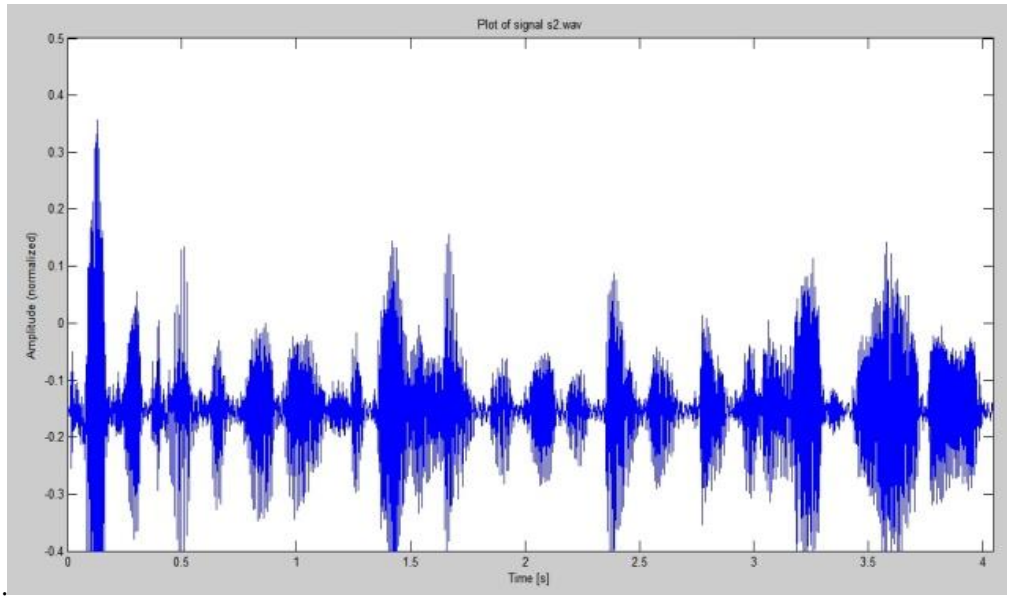

Fig.5 Speech Signal 2 from Testing Database VID TIMIT

8.1 Windowing and Fast Fourier Transform

After windowing, A next processing step is the Fast Fourier Transform, which converts each 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 6 (a) The figures are listed for M=100 and N=256. The logarithmic power spectrum has also been shown in fig 6(b)

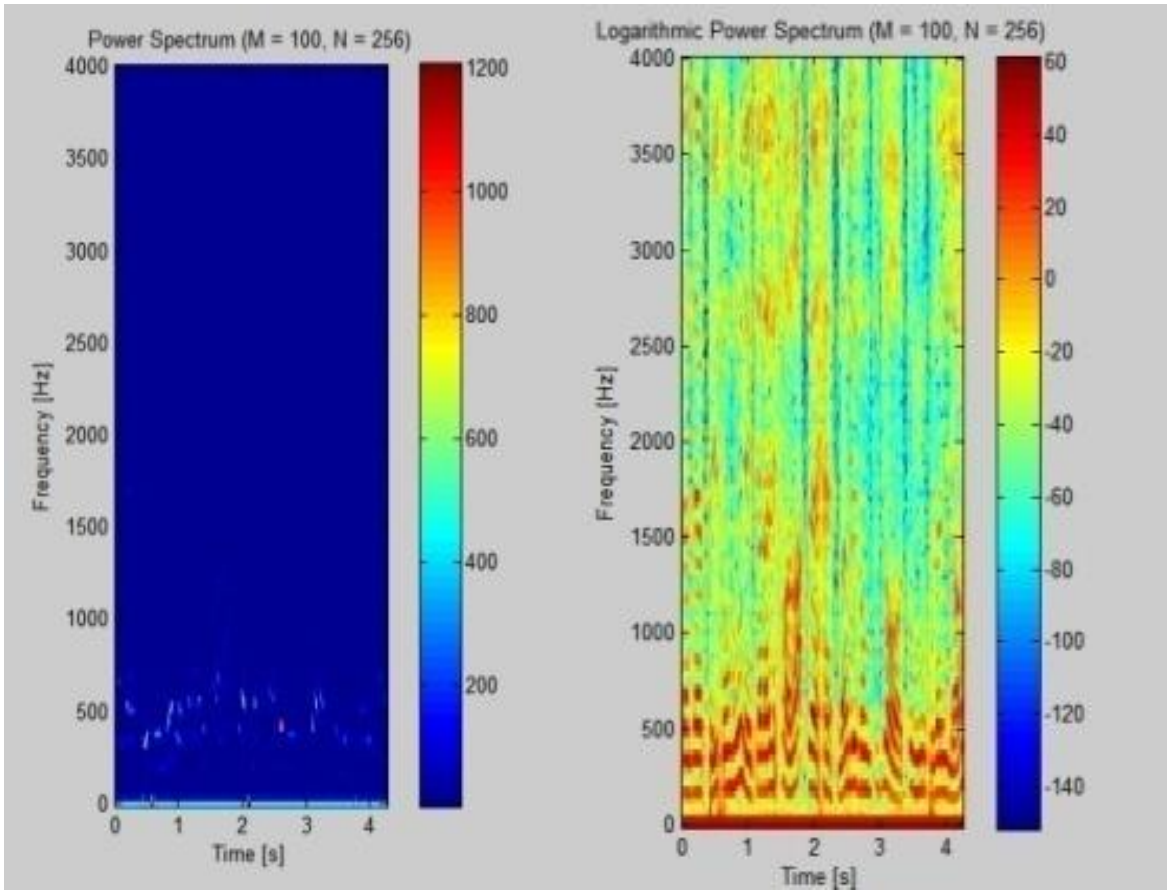

Fig. 6(a) Power Spectrum (b) Logarithmic Power Spectrum

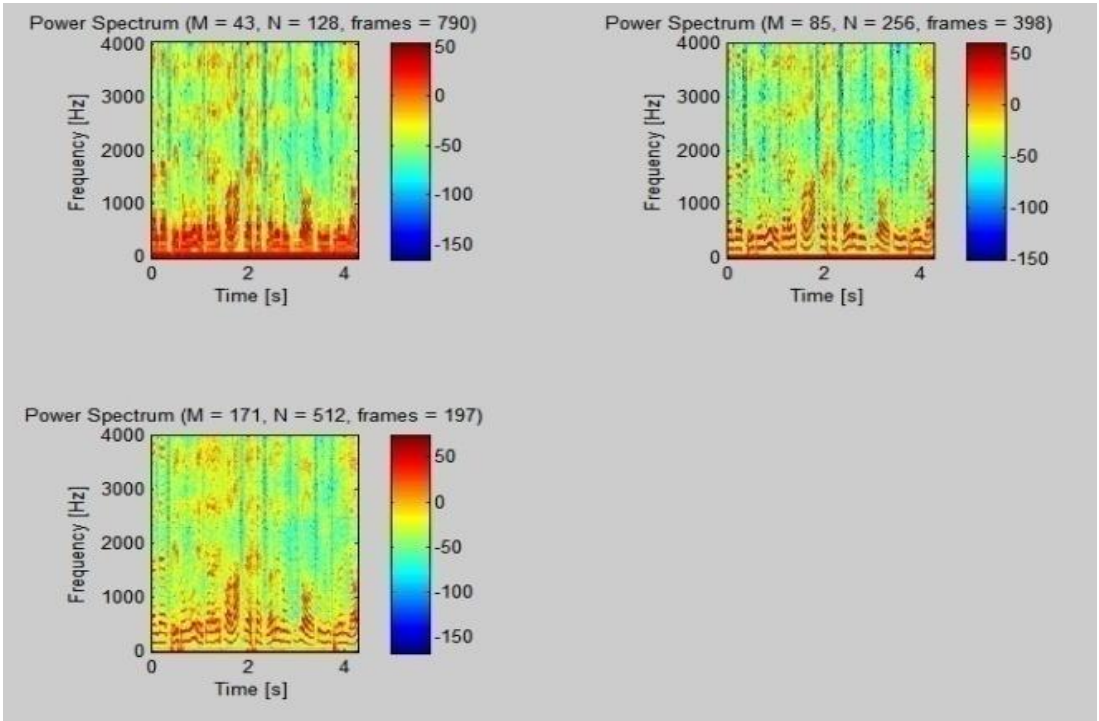

Fig 7 shows power spectrum for different values of M and N

International Journal of Electrical and Electronics Engineers
Vol. No.7 Issue 02, July-December 2015
www.arresearchpublication.com
IJEEE
ISSN 2321 - 2055

8.2 Calculation of MEL Frequency Coefficients

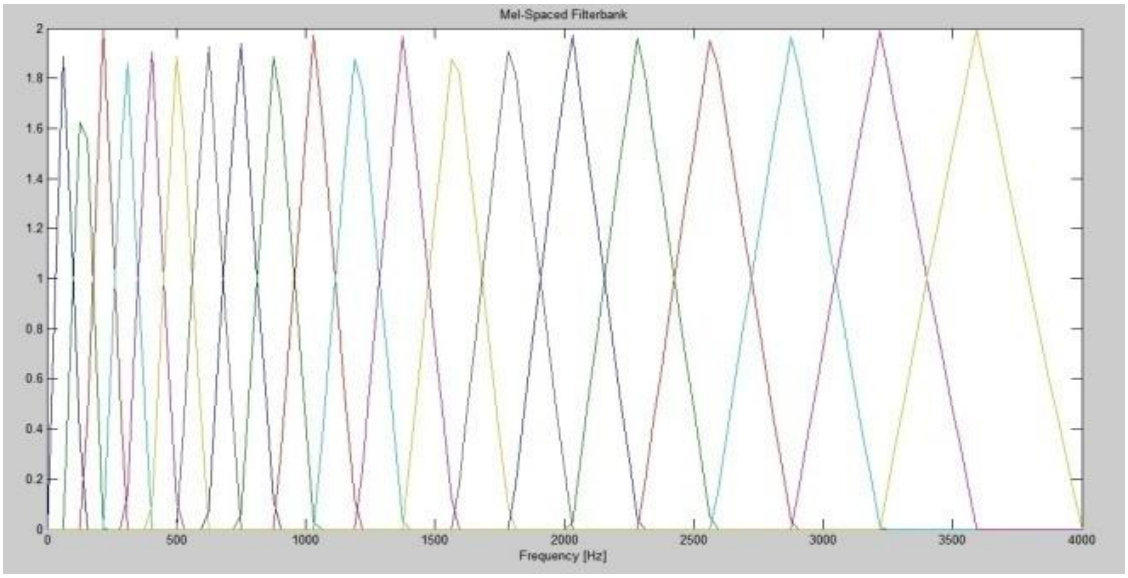

Fig. 8 Output from Mel filter bank

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. The number of Mel spectrum coefficients, K, is typically chosen as the 20.

The filter bank is applied in the frequency domain, therefore it simply amounts to taking those triangle-shape windows in the Figure(8) on the spectrum. A 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(9)shows the unmodified original spectrum and the spectrum which has been modified as a result of cepstrum analysis

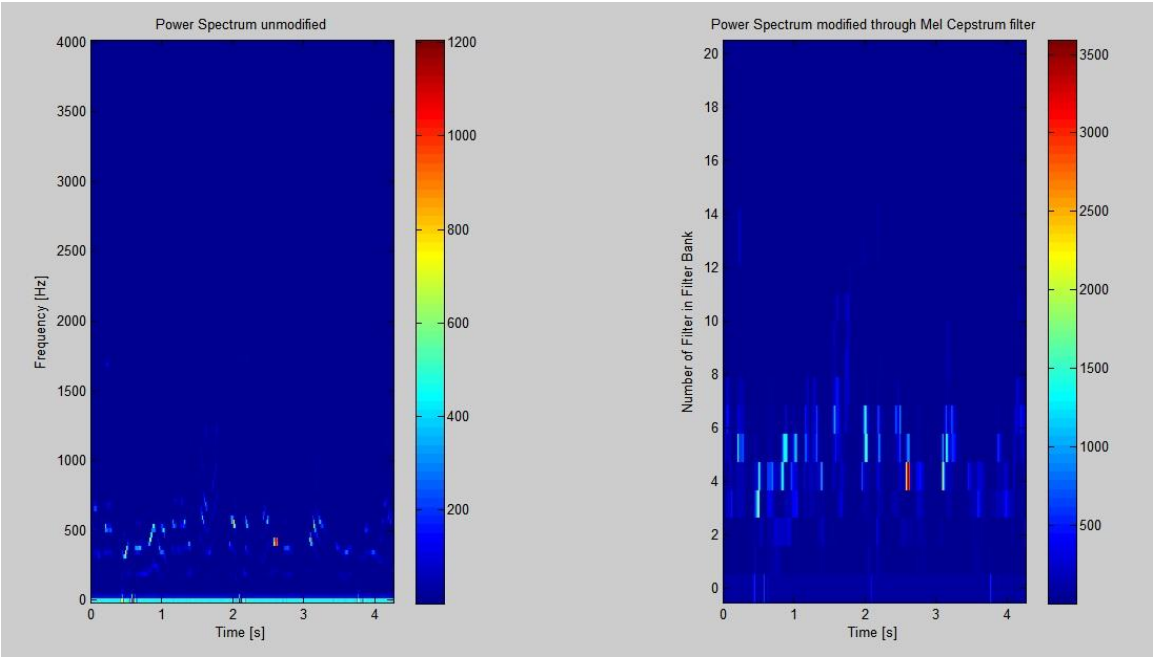

Fig .9.(a) Unmodified Spectrum(b) Power Spectrum after Mel Cepstrum filter

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

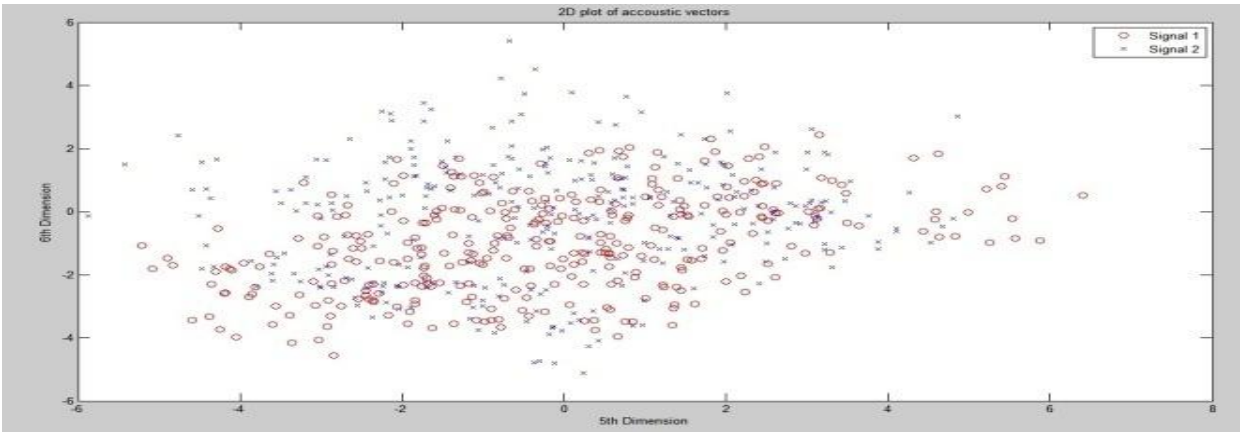

Fig. 10 Plot of 2D Acoustic Vectors

8.4 Vector Quantization using Linde Buzo Gray Algorithm

Finally on the application of VQ, we get a set of the 2D trained VQ code words. The figure(11) shows a 2D Plot of the 2D trained VQ codeword corresponding to the fifth and sixth dimension.

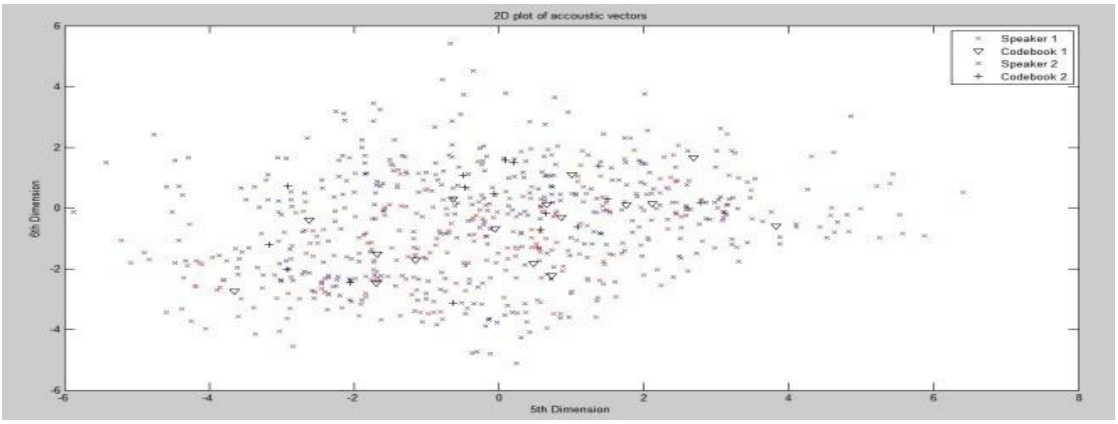

Fig. 11 2D plot of Trained VQ Code words

8.5 Performance Analysis Results

False Acceptance Rate (FAR) FAR Results have been shown in below fig(12) . As we can see that the False acceptance rate over a test set ranging from 44 to 100 users is roughly 9%.

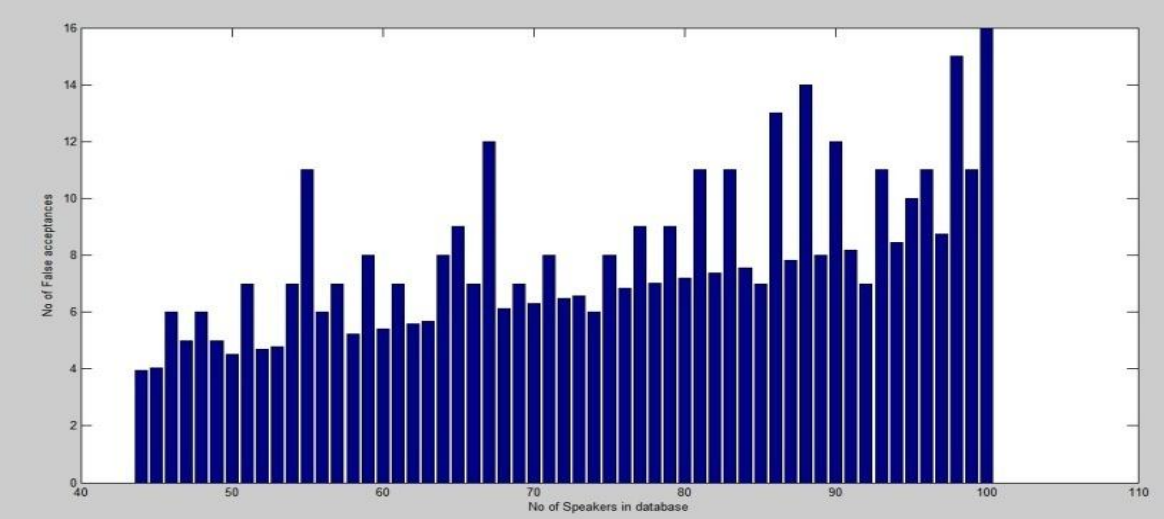

Fig 12 FAR for 44 to 100 user database

8.6 False Rejection Rate (FRR):-In below Fig(13) 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.

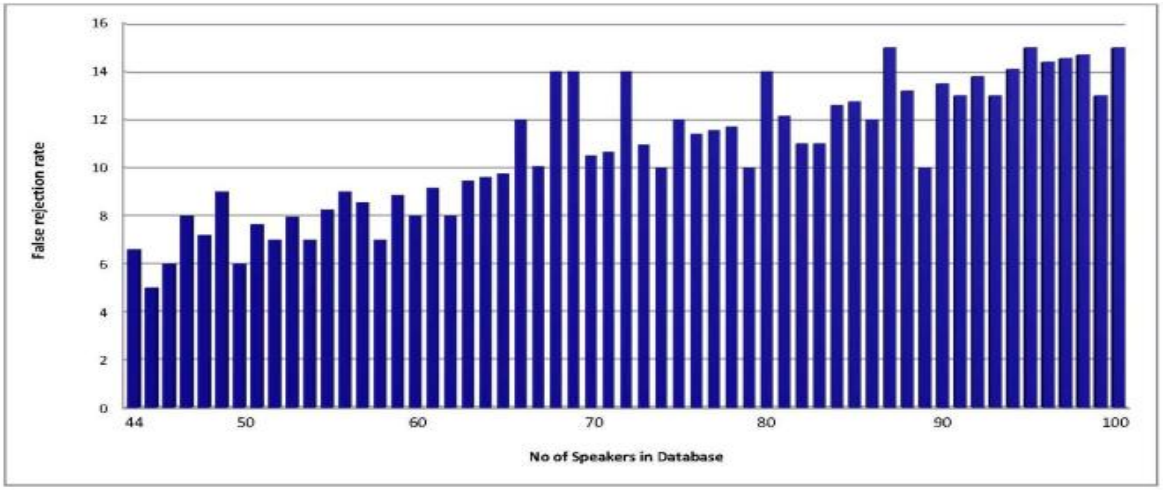

Fig 13 False Rejection Rate Results

8.7 True Acceptance Rate (TAR):- In below Fig(14) 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

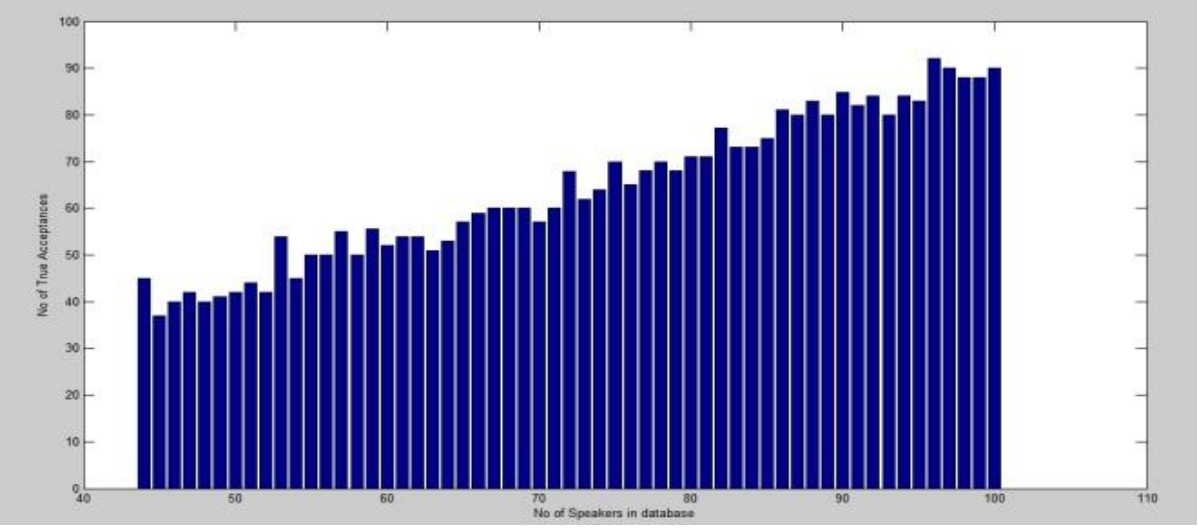

Fig14 True Acceptance Rate Results

8.8 Comparison Table

Table

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

IX. CONCLUSION

This work has presented an enhanced mechanism of Speaker Recognition using a combination of the well known MFCC algorithm as well as LBG algorithm for generating the vector code words. The 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 hum 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] Speaker segmentation and clustering (2008) M Kotti, V Moschou, C Kotropoulos., 2008"

[4] Todor Ganchev , Nikos Fakotakis , George Kokkinakis Comparative evaluation of various MFCC implementations on the speaker verification task (2005),

[5] 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

[6] From Frequency to Quefrency: A History of the Cepstrum Alan V. Oppenheim and Ronald W. Schafer, IEEE Signal Processing Magazine Reprinted 2004

[7] 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

[8]. R. Sambur A, . E. RosenbergL, . R. Rabinera, and C. A. McGonegal "On reducing the buzz in LPC synthesis", M Bell Laboratories

Under the Guidance of : Mr. Jitendra Kumar Mishra, Assistance Professor ,Dept of ECE PCST BHOPAL,