Real Time Speaker Recognition System using MFCC and Vector Quantization Technique

Roma Bharti
Mtech,
Manav rachna international university
Faridabad

Priyanka Bansal
Mtech
Manav rachna international university
Faridabad

ABSTRACT
This paper represents a very strong mathematical algorithm for Automatic Speaker Recognition (ASR) system using MFCC and vector quantization technique in the digital world. MFCC and vector quantization techniques are the most preferable and promising these days so as to support a technological aspect and motivation of the significant progress in the area of voice recognition. Our goal is to develop a real-time speaker recognition system that has been trained for a particular speaker and verifies the speaker. ASR is a type of biometric that uses an individual's voice for recognition processes. Speaker-vocal discriminative parameters exist in speech signals and due to dissimilar resonances of different speakers speaker recognition system verifies the speaker. These different characteristics can be accomplished by extracting features in vector form like Mel-Frequency Cepstral Coefficient (MFCCs) from the audio signal. The Vector Quantization (VQ) technique maps vectors from a large vector space to a limited number of regions in the same multidimensional space. LBG (Linde, Buzo and Gray) algorithm is mostly used and preferred for clustering a set of L acoustic vectors into a set of M codebook vectors in speaker recognition.

Keywords
Mel frequency cepstrum coefficient (MFCC), speaker recognition, speaker verification, vector quantization (VQ).

1. INTRODUCTION
The main motive of our project is to develop a real-time speaker recognition system which automatically recognizes the speech of enrolled speakers depending upon the vocal characteristics of the speakers. Today security is the major requirement, which creates essentiality of the development of biometrics, one of them is speaker recognition and it is beneficial for authentication of remote users. However, some confusion arises regarding speaker recognition and speech recognition, but actually both the systems are different as speech recognition is referred for recognition of words whether speaker recognition approaches speaker identification and verification. The objective of the speaker recognition system is to convert the acoustic audio signal into computer readable form. The human speech is then processed by the machine depending upon the two factors (i) feature extraction and (ii) feature matching. In our project, MFCC (Mel frequency cepstrum coefficient) is used for feature extraction as it is well known and most preferable technique, based on the known variation of the human ear's critical bandwidths with frequency.[8][11] Historically due to advancement of scientific researches, many feature matching techniques (BTW, GMM, HMM and VQ) have come in existence having a different probability of success rates and most of them are based on pattern recognition technique. Vector quantization technique for feature extraction is used in our project and it is based on pattern recognition and LBG (Linde, Buzo and Gray) algorithm and it provides better result even in the presence of noise. A complete speaker recognition system is generally categories as speaker identification and speaker verification and both of them can be text dependent or text independent depending on the applications.

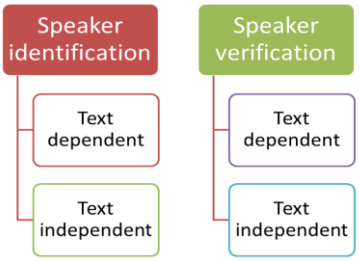

Fig 1: Scope of speaker recognition

A typical speaker recognition system has some basic functionalities like feature extraction, storing database, providing speaker IDs and feature matching. The modules of speaker recognition system defining these functionalities are defined by following steps[5]:-

- **Front-end processing:** In this part the sampled form of the continuous audio signal is converted into a set of feature vectors, which defines the relative parameters of speech of different speakers. In both training and recognition phases, front-end processing is performed.
- **Speaker modeling:** Modelling performs a reduction of feature data of the distributions of feature vectors.
- **Speaker database:** The speaker's reference database along with the speaker IDs are stored in this section.
- **Decision logic:** The final decision of the identity claim by the individual speaker is performed in this section by evaluating test audio feature vectors to all enrolled models in the database and selects the best matching model.

2. Feature extraction: MFCC (MEL FREQUENCY CEPSTRUM COEFFICIENT)
The Mel-Frequency Cepstrum (MFC) is a representation of short-period power spectrum of sound wave, and the collection of coefficients of MFC is referred as MFCC (Mel frequency cepstrum coefficient) which is based on auditory characteristics of human.[7]. According to psychological studies it has been proven that a human can only recognize the sound below 1000Hz means the critical bandwidth is limited to 1000Hz for the human ear.

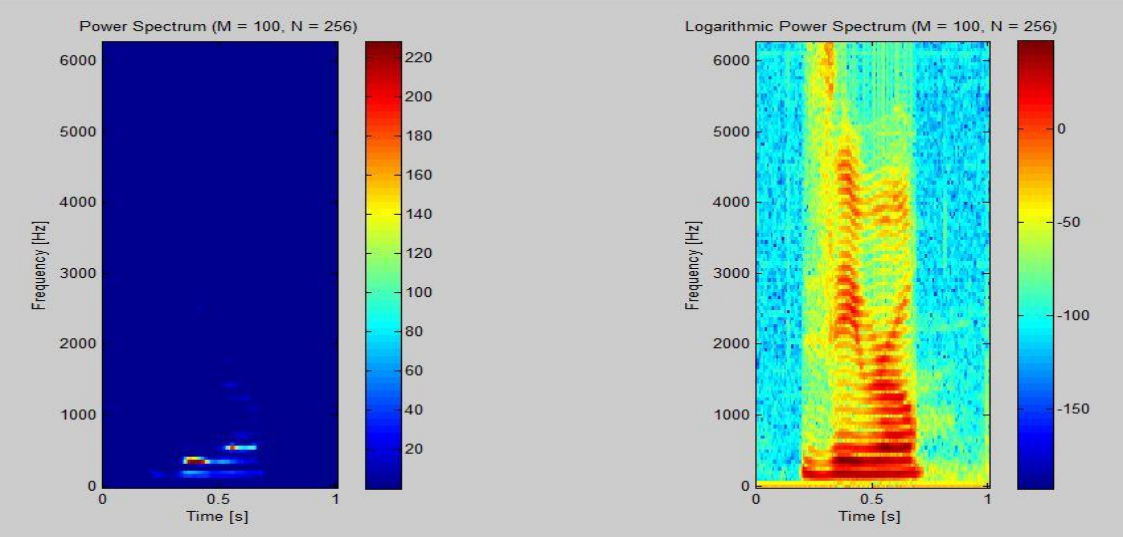

Fig 2: Power spectrum on linear and logarithmic scale

MFCC provides very much similar response as it is linearly spaced frequency below 1000Hz and logarithmically above 1kHz[5][7]. As shown in the fig 2 that maximum information is available at the lower frequencies. Hence, to make the information properly understandable power spectrum is replaced with a logarithmic scale.

2.1 Frame Blocking

In many kinds of research and experiments, it was found that the characteristics of speech signal remains stationary for short period of time(10-30msec) and highly changing for long period of time. Thus we convert the speech signal into a number of small frames having frame size N and separated from adjacent frame by M(M<N)for further processing (also called short time analysis)[3]. Fig 3 provides the graphical representation of power spectra for different frame size N and frame separation M. N=128 provides very high time resolution whereas N=512 provides very low resolution and N=256 provides a compromise between them.

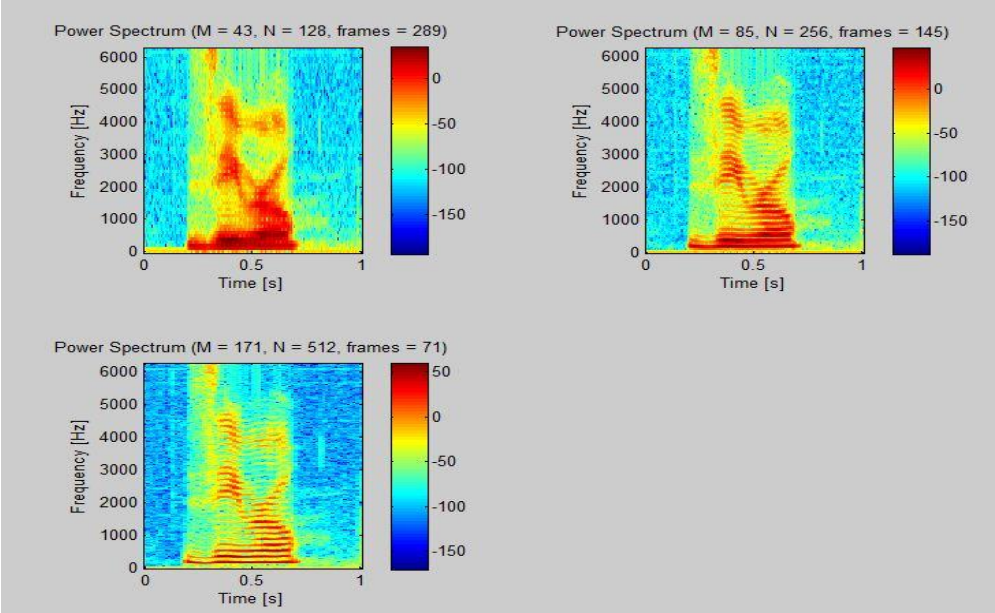

Fig 3: Power spectrum with different number of frames

2.2 Windowing

When frame windowing of each individual frame is modified so that there are no sudden discontinuities at the beginning and end, thus the corners of any frame do not overlap with the corner of another one. The hamming window and triangular window are mostly used for a speaker recognition system. The Hamming window is given by the equation as:

$Y[n]=X[n]*w[n]$

$W[n]=0.54-0.46\cos[2*\Pi*n/(N-1)]$

Where $0 \leq n \leq N-1$

Where, X[n] represents the input signal,

W[n] is the hamming window,

N is number of samples in each frame and

Y[n]=output signal

2.3 Fast Fourier Transform

The main function of this step is the conversion of N samples from the time domain into the frequency domain.[7]This step is used to eliminate the redundant mathematical calculations and enable analyzing the spectral properties of a signal.

$$\boldsymbol{X_n = \sum_{k=0}^{N-1} X_{ke^{-2k\Pi n/N}}} \quad n=0,1,2,3\ldots\ldots\ldots N-1$$

In case of complex only absolute value is considered. Where frequency range $0 \leq f \leq F/2$ corresponds to $0 \leq n \leq N/2-1$ and frequency range $-F/2 \leq f < 0$ corresponds to $N/2+1 \leq n \leq N-1$. Where F denotes the sampling frequency.

2.4 Mel Frequency Wrapping

Human perception of the frequency contents of acoustic signals is not found to follow linear scale instead it usually follows the Mel scale. Mel is a unit of measure of perceived pitch or frequency of the tone[3]. Thus, each tone having an actual frequency in Hz, a subjective pitch is represented in linear scale (linear spacing) below 1000Hz and logarithmically above to emphasize the result. Therefore, the approximate mathematical formula for Mel conversion is given by

$$\text{mel(f)}=2595*log_{10}(1+\frac{f}{700})$$

Where f(Hz) denotes frequency on the linear scale. In general filter banks are used for simulating the subjective spectrum (uniformly spaced on the Mel-scale). Each filter bank consists triangular band pass frequency response [3] which is applied in the frequency domain to get an efficient result. Overlapping histogram bins are usually implemented to provide representation of Mel wrapping in the frequency domain.

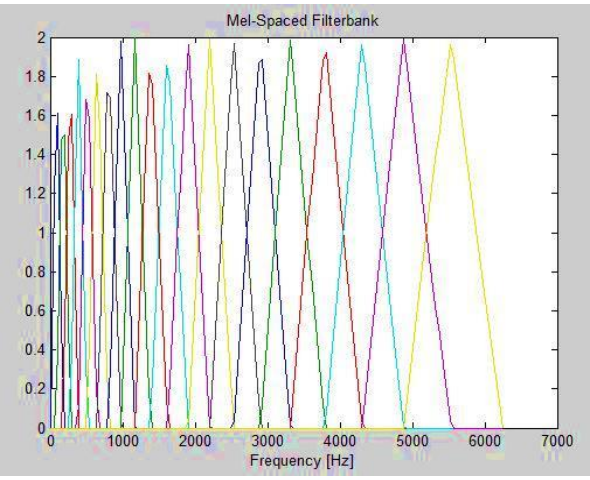

Fig 4: Mel spaced filterbank

2.5 Cepstrum

The final stage is the conversion of Mel spectrum back to time domain representation which provide a better representation of the spectral properties for a given time analysis. The result of the last step is is $S_k, k = 0,2, \ldots \ldots K-1$. Now we can calculate the cepstrum coefficient (MFCC) as: [11]

$$C_s \sum_{k=1}^{k}(logS_k)\cos[\frac{n\left(k-\frac{1}{2}\right)\Pi}{K}] \quad n=0,1,2\ldots\ldots K-1$$

Where k is a positive integer value.

3. FEATURE MATCHING USING VECTOR QUANTIZATION

Vector quantization maps the vectors from a large set into clusters (set of small region vectors) and these clusters are named as codebook. For every enrolled speaker codebook is generated, and during testing the encludian distance between the acoustic vector of test input signal and the mapped codebook is calculated. The speaker having the smallest encludian distance (also known as VQ distortion) is selected.

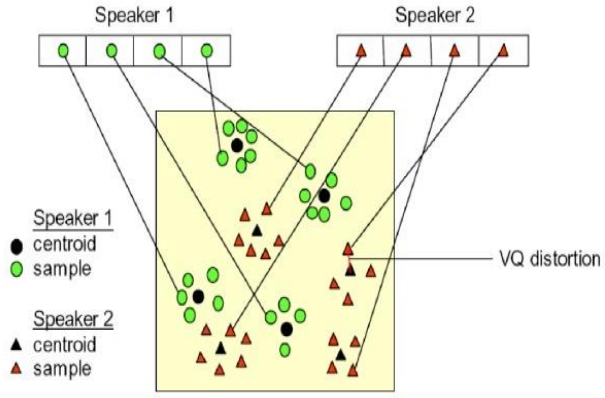

Fig 5: VQ approach

Vector quantization is based on the behavior of block coding and also referred as lossy compression of data. Fig shows the basic approach of VQ techniques, in diagram only 2 speaker's acoustic vector and the codebook is shown. The acoustic vector of speaker 1 is shown by circles while the acoustic vector of speaker 2 is shown by triangles. Codebooks of both the speakers are shown with black color.

VQ distortion illustrates the distance from the nearest codebook, calculated while testing phase of speaker recognition system. The speaker corresponds to minimum VQ distortion is selected and verified.

3.1 LBG Algorithm

Now the task is to build a codebook from a set of training vector for this purpose LBG algorithm is used. LBG ([Linde, Buzo and Gray,1980) is an algorithm used for clustering of L training vectors into a set of M codebook vectors. The following recursive procedure implemented in this algorithm[5].

1. **Design a 1-vector codebook :-** This is the centroid of the entire set of training vectors (hence, no iteration is required here).
2. Double the size of the codebook by splitting each current codebook y according to the rule

$$y_n^+ = y_n(1+e)$$

$$y_n^- = y_n(1+e)$$

Where, n varies from 1 to the current size of the codebook, and e is a splitting parameter.

3. **Nearest-Neighbour Search**: for each training vector, find the codeword in the current codebook that is closest (in terms of similarity measurement), and assign that vector to the corresponding cell (associated with the closest codeword).
4. **Centroid update**: Update the codeword in each cell using the centroid of the training vectors assigned to that cell.

5. **Iteration 1**: repeat steps 3 and 4 until the average distance falls below a preset threshold
6. **Iteration 2**: repeat steps 2, 3 and 4 until a codebook of size M is designed.

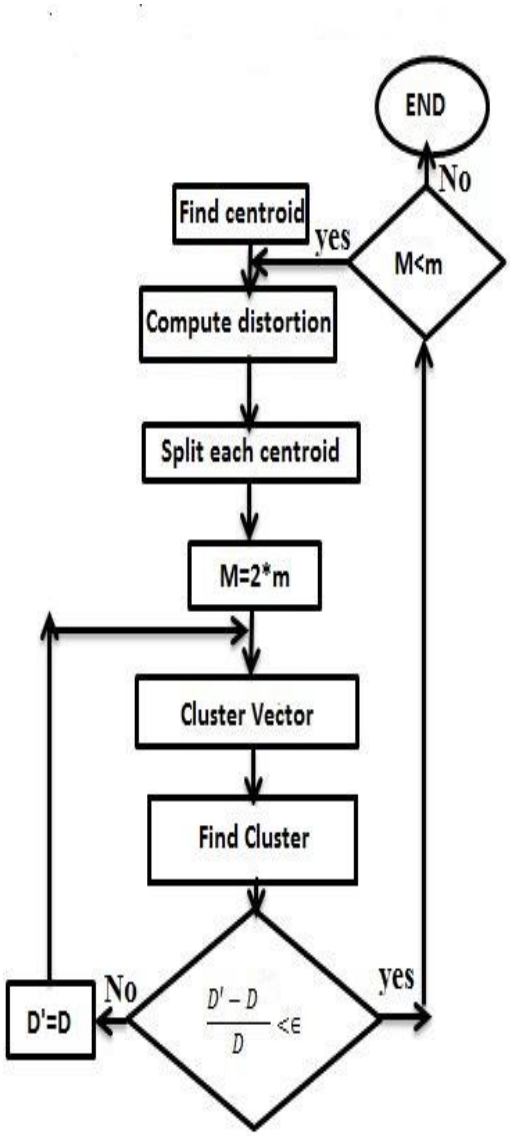

Fig 6: LGB algorithm

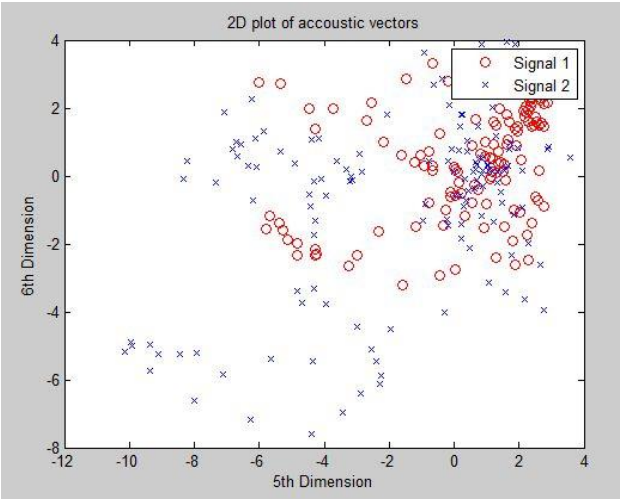

Fig 7: Acoustic vector after feature extraction

Fig 7 represents the plot of MFCC vector(acoustic space) in any two dimensional (say x and y) space. Plot represents the vectors of 2 speakers with red and blue color correspondingly.

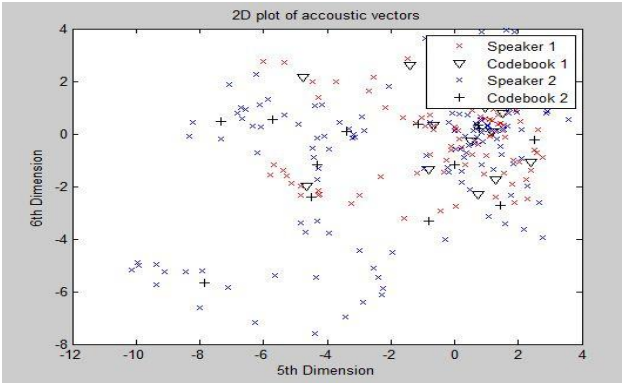

Fig 8: Acoustic vectors after training

Fig 8 plots the codebook for both the speakers (say s1 and s2) after applying VQ and LBG clustering on the MFCC features. These codebooks are used to calculate VQ distortion while the testing period.

4. EXPERIMENTAL RESULTS

Real-time speaker recognition system is developed using MATLAB. The results of the system are represented by the screenshots. In our project vector quantization method consist 2 phases: enrollment and testing phase. In the enrollment phase, we create a database of the speaker and stored it as a reference. Enrollment phase is shown in fig 7 while in the testing phase we verify the speaker's identity. Fig 8 shows the positive result and Fig 9 shows the negative result of the real-time speaker recognition system. Table 1 shows VQ distortion of feature vector some enrolled speakers (e=0.003). The threshold value of VQ distortion is selected as 7e, means if testing speech has VQ distortion less than 7e than only user is verified otherwise rejected.

Table 1: VQ distortion of feature vector

	Speakers	VQ distortion calculated in training phase
1	Speaker 1	2.211223e+000
2	Speaker 2	3.893212e+000
3	Speaker 3	3.732433e+000
4	Speaker 4	6.899947e+000
5	Speaker 5	5.456941e+000

5. CONCLUSION AND FUTURE SCOPE

A real-ime speaker recognition system using MFCC and vector quantization has been achieved and the experimental result has been analyzed using MATLAB. The result is concluded using 120 speakers with TIDIGIT database and provide 91% of accuracy in normal environmental condition at 20dB SNR. This system recognizes real-time speaker with the help of stored database and can be extended with more number of users. This system is applicable in security systems. In future it can be extended to more number of users and can also be used with other matching techniques (HMM) and biometrics (face recognition). Delta spectral cepstrum

coefficient can also be used in future to improve the robustness to the noisy condition.

6. ACKNOWLEDGEMENT

The author would really like thank Mrs. Deepali , Manav Rachna International University, Faridabad for their help and support during the development of MATLAB code , the speakers for creating database and the reviewers for the kind acceptance and important suggestion to improve this article.

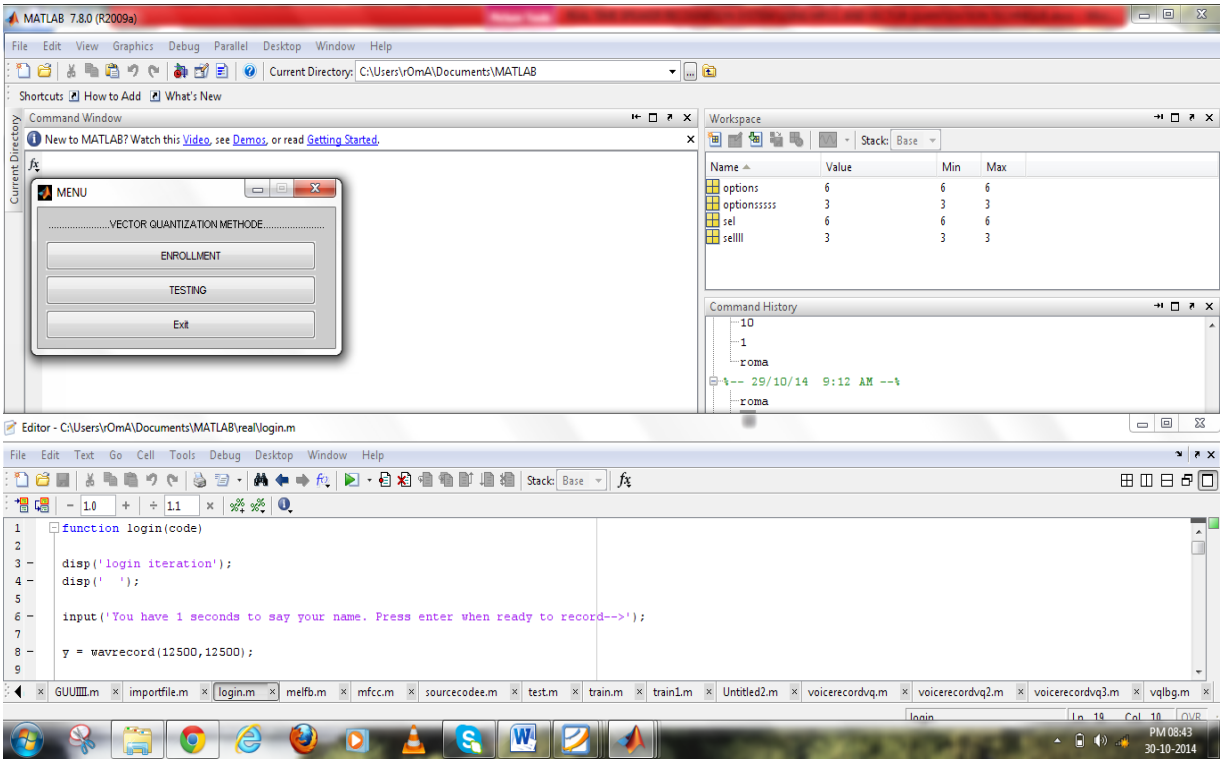

Fig 9: Speaker recognition phase

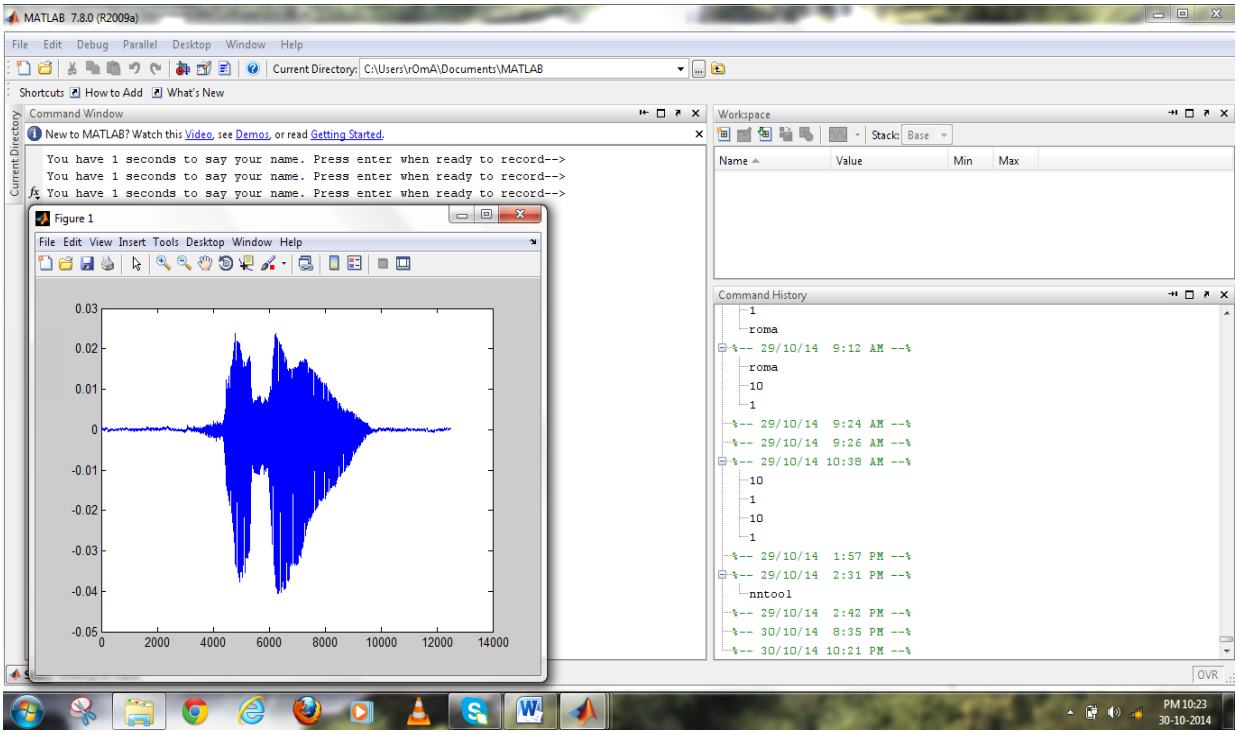

Fig 10: Enrolment phase

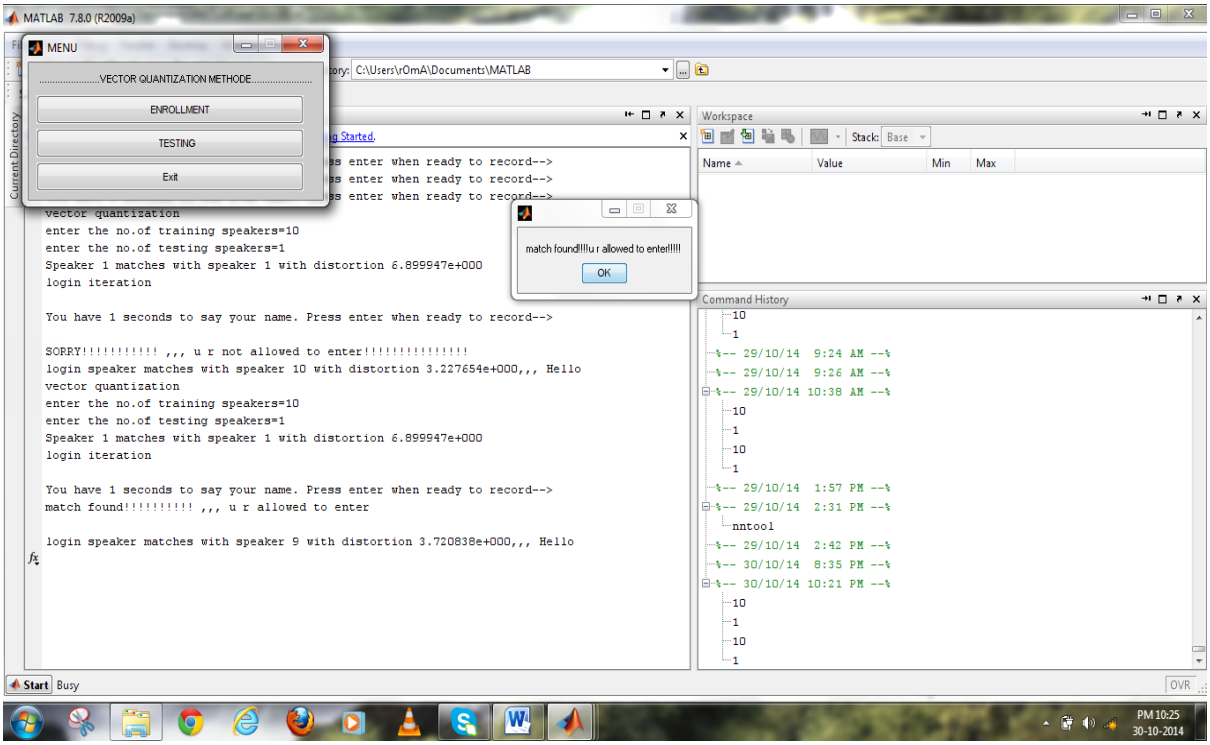

Fig 11: Speaker verification (positive result)

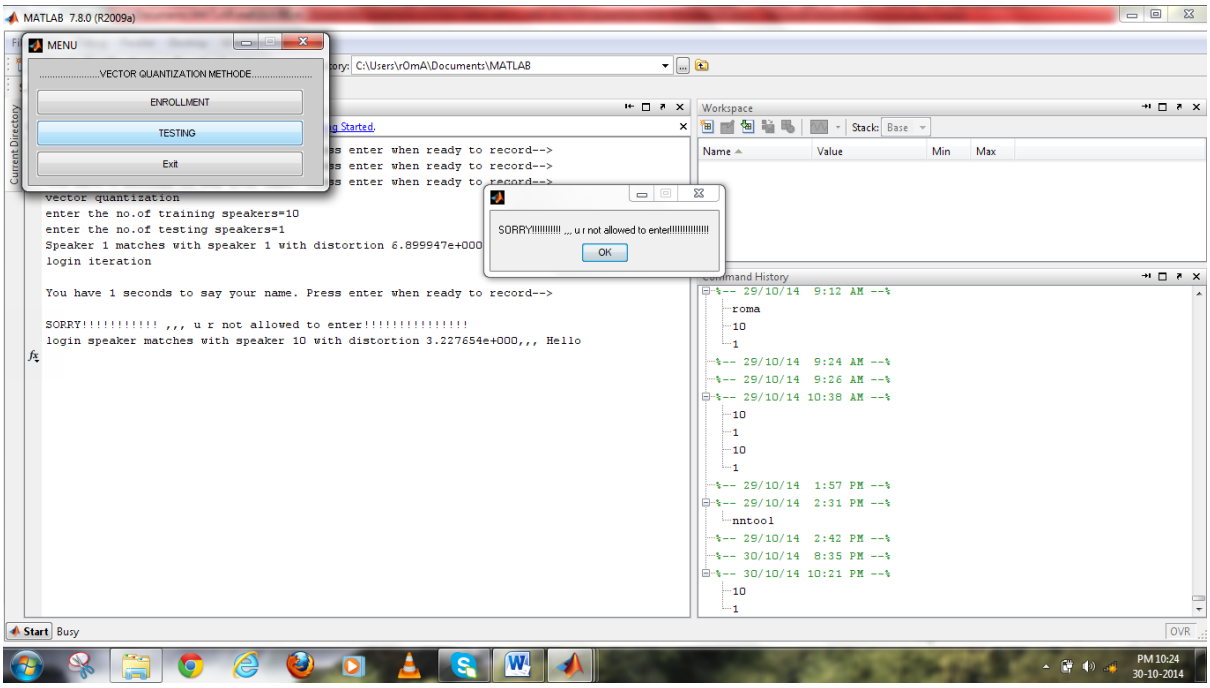

Fig 12: Speaker verification (negative result)

7. REFRENCES

[1] Rishiraj Mukherjee, Tanmoy Islam, and Ravi Sankar **"**text dependent speaker recognition using shifted mfcc" IEEE, 2013,.

[2] Hemlata Eknath Kamale, Dr.R. S. Kawitkar "Vector Quantization Approach for Speaker Recognition" International Journal of Computer Technology and Electronics Engineering (IJCTEE), Volume 3, , March-April 2013.

[3] Anjali Jain, O.P. Sharma "A Vector Quantization Approach for Voice Recognition Using Mel Frequency Cepstral Coeicient (MFCC): A Review" nternational Journal of electronics & communication technology, April - June 2013.

[4] Priyanka Mishra, Suyash Agrawal " Recognition of Speaker Using Mel Frequency Cepstral Coefficient & Vector Quantization" International Journal of Science, Engineering and Technology Research (IJSETR) ,December 2012

[5] A. Srinivasan , MAY 2012, "Speaker Identification and Verification using Vector Quantization and Mel Frequency Cepstral Coefficients" Research Journal of Applied Sciences, Engineering and Technology 4(1): 33-40, 2012

[6] Dipmoy Gupta, Radha Mounima C. Navya Manjunath, Manoj PB "Isolated Word Speech Recognition Using Vector Quantization (VQ)" International Journal of Advanced Research in Computer Science and Software Engineering , May 2012.

[7] Nitisha, Anshu Bansal "Speaker Recognition Using MFCC Front End Analysis and VQ Modeling Technique for Hindi Words using MATLAB" International Journal of Computer Applications (0975 8887) Volume 45 No.24, May 2012.

[8] Prof. Ch.Srinivasa Kumar, Dr. P. Mallikarjuna Rao "Design Of An Automatic Speaker Recognition System Using MFCC, Vector Quantization And LBG Algorithm" international journal of computer science and Engineering (IJSCE); Aug 2011

[9] HarisBC, GPradhan, AMisra, SShukla, RSinha and SRMP rasanna" Multi-Variability Speech Database for Robust SpeakerRecognition" IEEE 2011.

[10] A. Stolcke, E.Shriberg, L. Ferrer, S. Kajarekar, K. Sonmez, G. Tur "speech recognition as feature extraction for speaker recognition" Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA.

[11] Vibha Tiwari "MFCC and its applications in speaker recognition" International Journal on Emerging Technologies, 2010.

[12] Jeng-Shyang Pan, Zhe-Ming Lu, and Sheng-He Sun "An Efficient Encoding Algorithm for Vector Quantization Based on Subvector Technique", IEEE VOL. 12, NO. 3, March 2003.

[13] Linde, Y., A. Buzo and R. Gray "An algorithm for vector quantizer design". IEEE Trans. Commun., 2884-95,1980.