Research Journal of Applied Sciences, Engineering and Technology 4(1): 33-40, 2012
ISSN: 2040-7467

Submitted: September 02, 2011 Accepted: October 07, 2011 Published: January0 1, 2012

Speaker Identification and Verification using Vector Quantization and Mel Frequency Cepstral Coefficients

A. Srinivasan
Department of ECE, Srinivasa Ramanujan Centre, SASTRA University,
Kumbakonam-612001, India

Abstract: In the study of speaker recognition, Mel Frequency Cepstral Coefficient (MFCC) method is the best and most popular which is used to feature extraction. Further vector quantization technique is used to minimize the amount of data to be handled in recent years. In the present study, the Speaker Recognition using Mel Frequency Cepstral coefficients and vector Quantization for the letter "Zha" (in Tamil language) is obtained. The experimental results are analyzed with the help of MATLAB in different situations and it is proved that the results are efficient in the noisy environment.

Key words: MATLAB, Mel frequency, speaker recognition, vector quantization (VQ)

INTRODUCTION

In speech recognition, HMMs have been used for modeling observed patterns from 1970s. Many researchers (Rabiner and Juang, 1986; Rabiner and Schafer, 1978; Russell and Moore, 1985 and Fu, 1980) published a large number of papers, which present HMM as tool for use on these practical problems. These studies are written by researchers interested in pattern recognition, often from a viewpoint in engineering or computer science, and they usually focus on algorithms and on results in practical situations. Speech recognition recognizes the words but speaker recognition identifies and verifies the speaker. It is a biometric modality that uses an individual's voice for recognition processes. In general speaker recognition is referred by two different subtasks viz, Speaker Identification (SI) and Speaker Verification (SV). Of which identification task is considered more difficult. When the number of speakers increases, the probability of an incorrect decision increases (Doddington, 1985; Furui, 1986, 1994, 2001; Prabhakar *et al*., 2003). The performance of the verification task is not, at least in theory, affected by the population size since only two speakers are compared.

In the auditory speaker recognition, it has been observed that there are considerable differences between individuals (Rose, 2002; Schmidt-Nielsen and Crystal, 2000). Moreover, human performance decreases as the time increases between listening the two voices (Kerstholt *et al*., 2003). Several studies have been conducted to compare human and machine performance in speaker recognition (Schmidt-Nielsen and Crystal, 2000; Liu *et al*., 1997; Sullivan and Pelecanos, 2001). Schmidt-Nielsen and Crystal have conducted a large-scale comparison in which nearly 50,000 listening judgments were performed by 65 listeners grouped in panels of 8 listeners. The results were compared with the state-of-the-art computer algorithms. It was observed that individual human listeners vary significantly in their ability to recognize speakers.

In recent years, the higher level cues have begun to interest more and more researchers in automatic speaker recognition (Campbell *et al*., 2003; Doddington, 2001; Reynolds *et al*., 2003; Xiang, 2003). For instance, recently automatic systems that use several low- and high-level speaker cues have been introduced (Campbell *et al.*, 2003; Reynolds *et al.*, 2003). Although many new techniques were invented and developed, there are still a number of practical limitations because of which widespread deployment of applications and services is not possible. Vector Quantization is an efficient data compression technique and has been used in various applications involving VQ-based encoding and VQ based recognition. Vector Quantization has been very popular in the field of speech recognition. Speech recognition of the letter "Zha" (in Tamil language) by using LPC (Srinivasan *et al*., 2009) and using HMM (Srinivasan, 2011) are recognized.

In the present study the Speaker Recognizer using Mel Frequency Cepstral coefficients and vector Quantization for the letter "Zha" (in Tamil language) is obtained. The experimental results are analyzed with the help of MATLAB**.**

METHODOLOGY

Modules of speaker recognition: A speaker recognition system is mainly composed of the following four modules:

- **Front-end processing:** It is the "signal processing" part, which converts the sampled speech signal into set of feature vectors, which characterize the properties of speech that can separate different speakers. Front-end processing is performed both in training- and recognition phases.
- **Speaker modeling:** This part performs a reduction of feature data by modelling the distributions of the feature vectors.
- **Speaker database:** The speaker models are stored here.
- **Decision logic:** It makes the final decision about the identity of the speaker by comparing unknown feature vectors to all models in the database and selecting the best matching model.

Mel frequency cepstral coefficients: Mel frequency Cepstral Coefficients are coefficients that represent audio based on perception. This coefficient has a great success in speaker recognition application. It is derived from the Fourier Transform of the audio clip. In this technique the frequency bands are positioned logarithmically, whereas in the Fourier Transform the frequency bands are not positioned logarithmically. As the frequency bands are positioned logarithmically in MFCC, it approximates the human system response more closely than any other system. These coefficients allow better processing of data. In the Mel Frequency Cepstral Coefficients the calculation of the Mel Cepstrum is same as the real Cepstrum except the Mel Cepstrum's frequency scale is warped to keep up a correspondence to the Mel scale.

Vector quantization: Vector quantization is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a code word. The collection of all code words is called a code book. Vector Quantization (VQ) is a lossy data compression method based on principle of block coding . It is a fixed-to-fixed length algorithm. VQ may be thought as an approximator.

The technique of VQ consists of extracting a small number of representative feature vectors as an efficient means of characterizing the speaker specific features. By means of VQ, storing every single vector that we generate from the training is impossible. Figure 1 shows a conceptual diagram to illustrate this recognition process in the figure, only two speakers and two dimensions of the acoustic space are shown. The circles refer to the acoustic vectors from the speaker 1 while the triangles are from the speaker 2. In the training phase, using the clustering algorithm described in a speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors. The result code words (centroids) are shown by black circles and black triangles for speaker 1 and 2, respectively.

The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of an unknown voice is "vector-quantized" using each trained codebook and the total VQ distortion is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified as the speaker of the input utterance. By using these training data features are clustered to form a codebook for each speaker. In the recognition stage, the data from the tested speaker is compared to the codebook of each speaker and measure the difference. These differences are then use to make the recognition decision.

Design problem: The VQ design problem can be stated as follows. Given a vector source with its statistical properties known, given a distortion measure, and given the number of code vectors, find a codebook (the set of all red stars) and a partition (the set of blue lines) which result in the smallest average distortion. We assume that there is atraining sequence consisting of M source vectors:

$$T = \{x_1, x_2, ..., x_M\}$$

This training sequence can be obtained from some large database. For example, if the source is a speech signal, then the training sequence can be obtained by recording several long telephone conversations. M is assumed to be sufficiently large so that all the statistical properties of the source are captured by the training sequence. We assume that the source vectors are k-dimensional, e.g,

$$x_m = (x_{m,1}, x_{m,2}, ..., x_{m,k}),\ m = 1, 2, ..., M$$

Let N be the number of code vectors and let:

$$C = \{c_1, c_2, ..., c_N\}$$

represents the codebook. Each code vector is K dimensional, e.g.,

$$c_n = (c_{n,1}, c_{n,2}, ..., c_{n,k}),\ n = 1, 2, ..., N$$

Let S_n be the encoding region associated with code vector C_n and let:

$$p = \{S_1, S_2, ..., S_N\}$$

Denote the partition of the space. If the source vector x_m is in the encoding region S_n, then its approximation (denoted by $Q(X_n)$ is Cn:

$$Q(x_m) = c_n, \text{ if } x_m \in s_n$$

Assuming a squared-error distortion measure, the average distortion is given by:

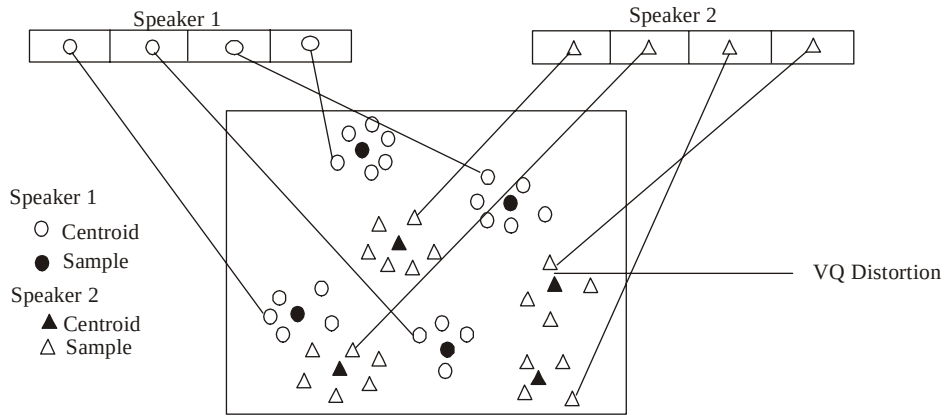

Fig. 1: VQ classification model

$$D_{ave} = \frac{1}{Mk} \sum_{m=1}^{M} \left\| x_m - Q(x_m) \right\|^2$$

The design problem can be succinctly stated as follows: Given T and N find C and P such that D is minimized.

Optimality criteria: If C and P are a solution to the above minimization problem, then it must satisfy the following two criteria.

Nearest neighbor condition:

$$s_n = \left\{ x : \left\| x - c_n \right\|^2 \le \left\| x - c_{n'} \right\|^2 \forall\, n' = 1, 2, \ldots, N \right\}$$

This condition says that the encoding region S_n should consists of all vectors that are closer to C_n than any of the other code vectors. For those vectors lying on the boundary (blue lines) tie-breaking procedure will do.

Centroid condition:

$$c_n = \frac{\sum_{x_m \in S_n} x_m}{\sum_{x_m \in S_n} 1} \qquad n = 1,2,\ldots,N$$

This condition says that the code vector C_n should be average of all those training vectors that are in encoding region S_n. In implementation, one should ensure that at least one training vector belongs to each encoding region (so that the denominator in the above equation is never 0).

Algorithm: After the enrolment session, the acoustic vectors extracted from input speech of each speaker provide a set of training vectors for that speaker. As described above, the next important step is to build a speaker-specific VQ codebook for each speaker using those training vectors. There is a well-know algorithm, namely LBG algorithm Linde *et al*., 1980, for clustering a set of *L* training vectors into a set of *M* codebook vectors. The algorithm is formally implemented by the following recursive procedure:

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

$$y_n^+ = y_n(1+\epsilon)$$
$$y_n^- = y_n(1-\epsilon)$$

  where, *n* varies from 1 to the current size of the codebook, and is a splitting parameter (we choose $\epsilon = 0.01$).
- Nearest-Neighbor 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).
- Centroid update: Update the codeword in each cell using the centroid of the training vectors assigned to that cell.
- Iteration 1: repeat steps 3 and 4 until the average distance falls below a preset threshold
- Iteration 2: repeat steps 2, 3 and 4 until a codebook size of *M* is designed

Intuitively, the LBG algorithm designs an *M*-vector codebook in stages. It starts first by designing a 1-vector codebook, then uses a splitting technique on the codewords to initialize the search for a 2-vector codebook, and continues the splitting process until the desired *M*-vector codebook is obtained.

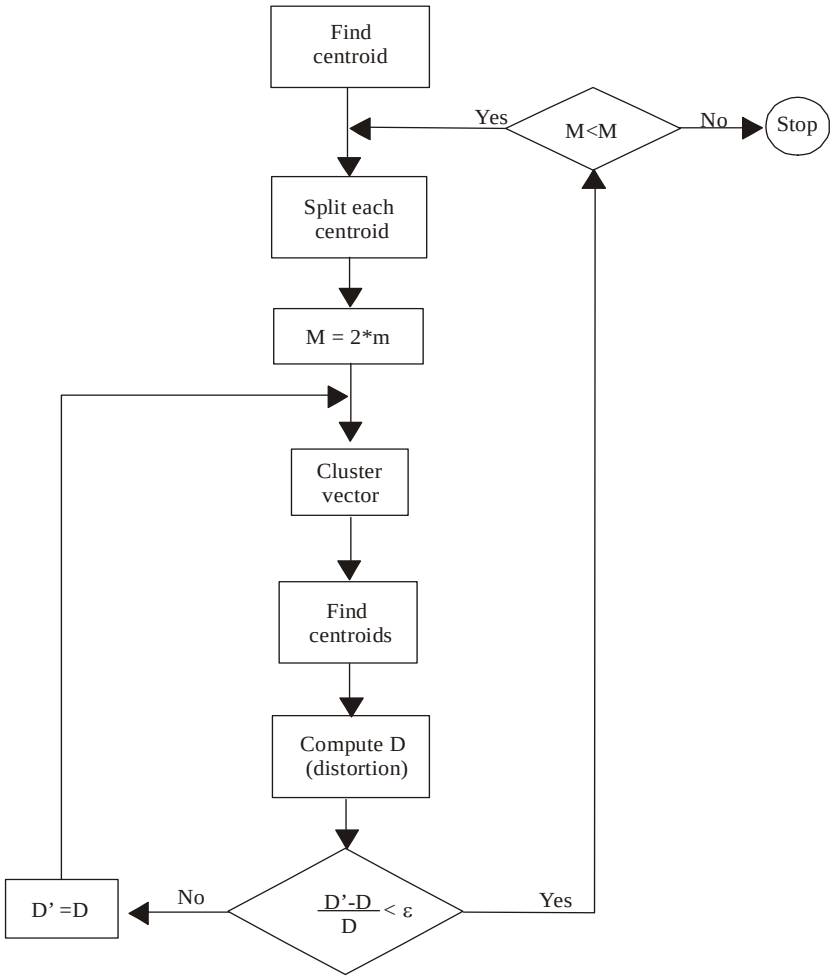

Fig. 2: VQ flow chart

LBG design algorithm: The LBG VQ design algorithm is an iterative algorithm which alternatively solves the above two optimality criteria. The algorithm requires an initial code book. This initial codebook is obtained by the *splitting* method. In this method, an initial code vector is set as the average of the entire training sequence. This code vector is then split into two. The iterative algorithm is run with these two vectors as the initial codebook. The final two code vectors are splitted into four and the process is repeated until the desired number of code vectors is obtained. Figure 2 shows the flow chart of the algorithm.

EXPERIMENTAL RESULTS

The most distinctive letter in Tamil language is "Zha" because of the deliberation, the articulation of the sound demands. Therefore the trained set of fifty speakers was selected to spell the letter "Zha" for the speaker recognition. Environmental noises during the recording process are overcome using Wavesurfer tool. It is a simple but powerful interface. The standard speech analysis such as waveform, Spectrogram, Pitch, and Power panes are analyzed. Magnitude and frequency comparison of 3 male and 3 female speakers is shown in Table 1.

Trained and test set of voice data are processed in MATLAB using vector quantization, subsequently Mel frequency cepstral coefficients are obtained and shown in Fig. 3. The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis.

In the speaker identification process it is identified that the speakers are matched with test and trained data (Fig. 4). During the verification process of the speaker recognition, positive and negative results have been arrived. If the speakers are matched (positive) with voice

Table 1: Magnitude and frequency comparison

S1.No.	Frequency (HZ)	*F1 dB	F2 dB	F3 dB	**M1 dB	M2 dB	M3 Db
1	15.625	- 20.47	- 20.98	- 19.67	- 21.02	- 20.86	- 20.71
2	140.625	- 32.68	- 32.73	- 32.01	- 32.94	- 33.13	- 32.75
3	390.625	- 36.63	- 36.74	- 36.13	- 36.94	- 37.14	- 37.01
4	640.625	- 41.36	34.99	- 40.97	- 41.48	- 41.92	- 40.99
5	1015.625	- 51.20	- 47.91	- 51.00	- 51.90	- 52.65	- 50.90

*: Female; **: Male

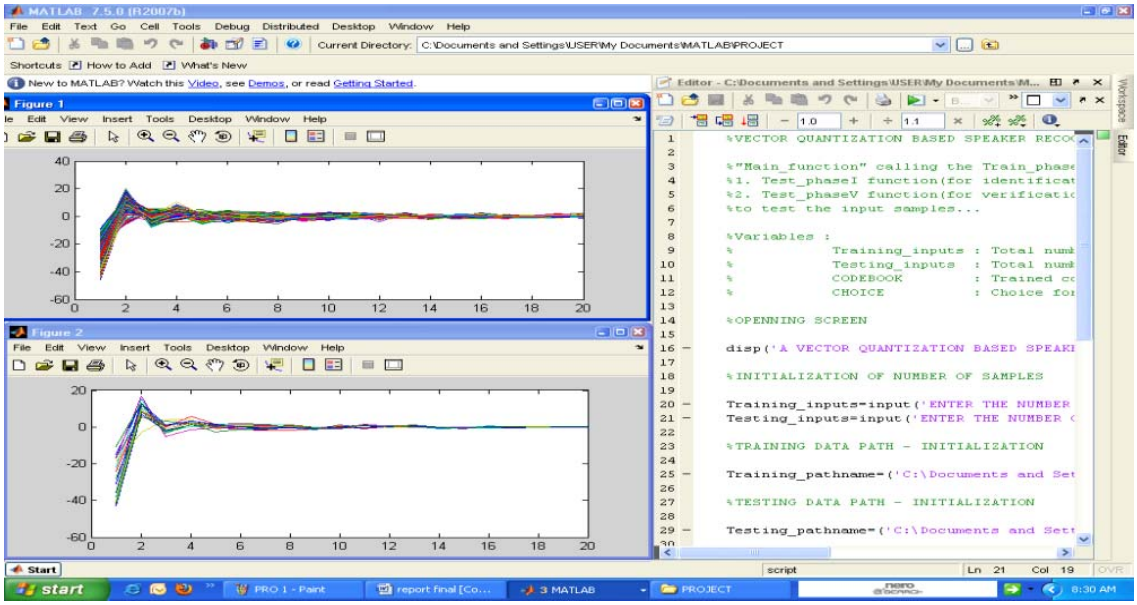

Fig. 3: MFCC vactors before and after VQ

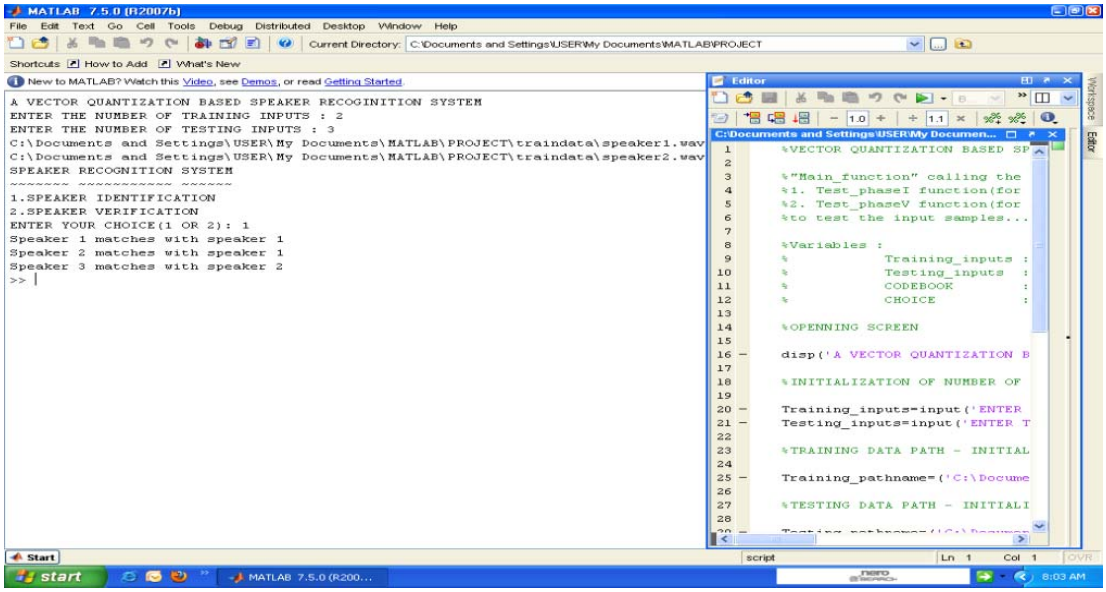

Fig. 4: Speaker identification

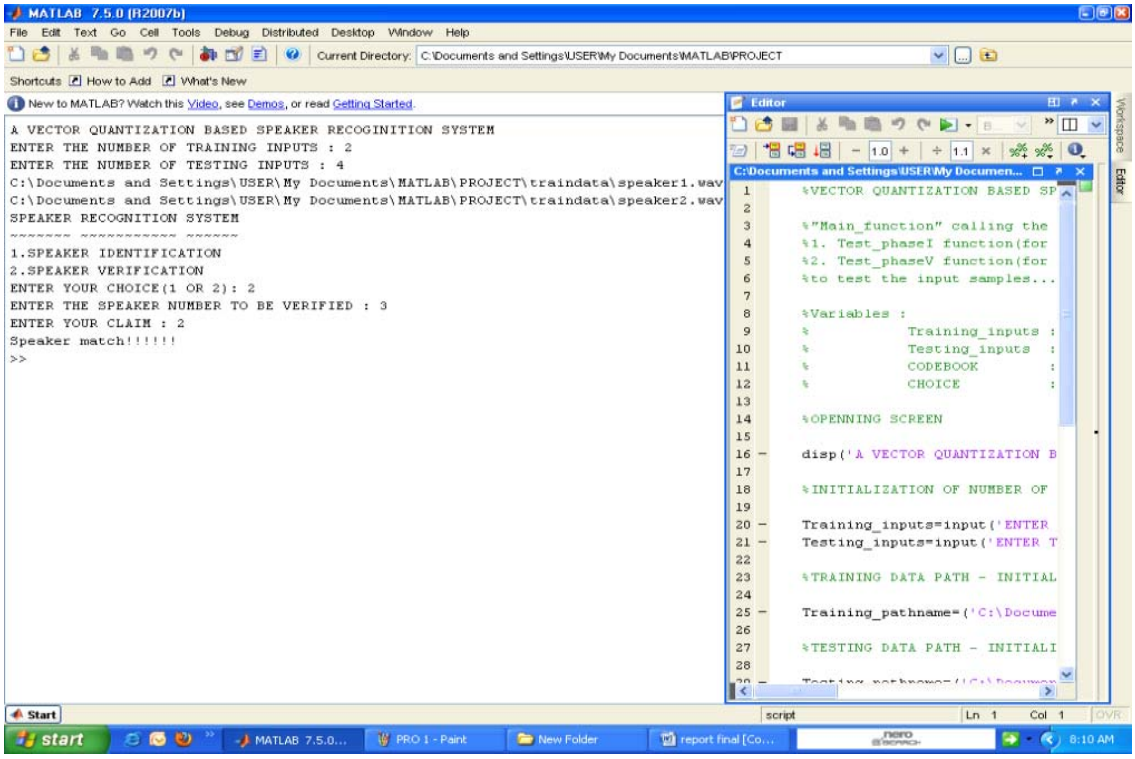

Fig. 5: Speaker verification (positive result)

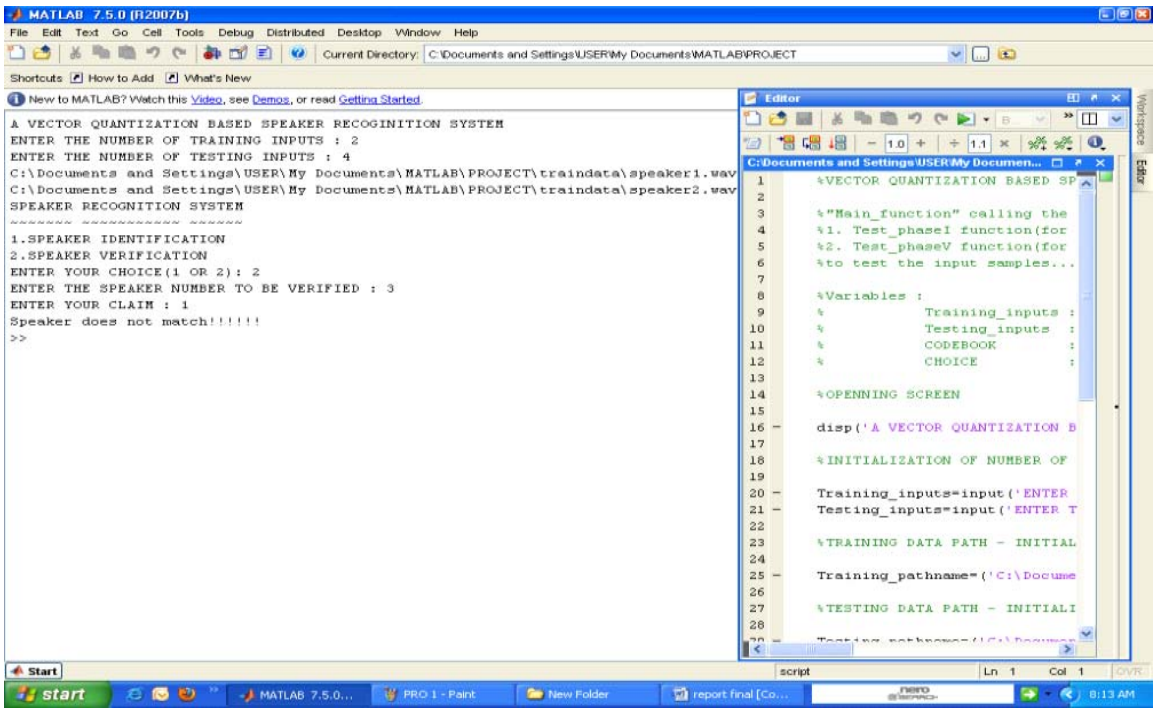

Fig. 6: Speaker verification (negative result)

the output will be matched otherwise does not matched. Fig. 5 (positive) and Fig. 6 (negative) show the verification of the speakers.

The performance of VQ is typically given in terms of the signal-to-distortion ratio (*SDR*):

$$SDR = 10\log_{10}\left(\frac{\sigma^2}{D_{ave}}\right)$$

where s^2 is the variance of the source and D_{ave} is the average squared-error distortion. The higher value of the *SDR* gives the better performance.

CONCLUSION

Speaker Recognition using Mel Frequency Cepstral coefficients and vector Quantization for the letter "Zha" (in Tamil language) is recognized. The experimental results are analyzed with the help of MATLAB and it is proved that the results are efficient. This process can be extended for n number of speakers. In future, Speaker Recognition process will receive the prime importance for voice based Automatic Teller Machine.

ACKNOWLEDGMENT

The Author would like to thank G. Rajaa Krishnamurthy and G. Raghavan, Tata consultancy services, Chennai, for their help during the process in Matlab and the speakers for spelling out the letter "Zha" during the process and the reviewer for his kind acceptance to review this article with valid suggestions to improve the manuscript.

REFERENCES

Campbell, J., D. Reynolds and R. Dunn, 2003. Fusing high-and low-level features for speaker recognition. Proceeding on 8th European Conference on Speech Communication and Technology (*Eurospeech*) (Geneva, Switzerland), pp: 2665-2668.

Doddington, G., 1985. Speaker recognition-identifying people by their voices. Proceedings IEEE, 73(11): 1651-1164.

Doddington, G., 2001. Speaker Recognition Based on Idiolectal Differences between Speakers, Proceeding. 7th European Conference on Speech Communication and Technology, (*Eurospeech*) (Aalborg, Denmark), pp: 2521-2524.

Fu, K.S., 1980. Statistical Pattern Classification Using Contextual Information. Research Studies Press.

Furui, S., 1986. Speaker independent isolated word recognition using dynamic features of speech spectrum. IEEE Transactions on Acoustic, Speech, Signal Processing, ASSP-34(1): 52-59.

Furui, S., 1994. An overview of speaker recognition technolog. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp: 1-9.

Furui, S., 2001. Digital Speech Processing, Synthesis, and Recognition, 2nd Edn., Marcel Dekker, Inc., New York.

Kerstholt, J., E. Jansen, A. van Amelsvoort and A. Broeders, 2003. Earwitness line-ups: Effects of speech duration, retention interval and acoustic environment on identification accuracy. Proceeding on. 8th European Conference on Speech Communication and Technology (*Eurospeech*) (Geneva, Switzerland), pp: 709-712.

Linde, Y., A. Buzo and R. Gray, 1980. An algorithm for vector quantizer design. IEEE Trans. Commun., 28: 84-95.

Liu, L., J. He and G. Palm, 1997. A Comparison of Human and Machine in Speaker Recognition. In Proc. 5th European Conference on Speech Communication and Technology (Eurospeech) (Rhodos, Greece), pp: 2327-2330.

Prabhakar, S., S. Pankanti and A. Jain, 2003. Biometric recognition: Security and privacy concerns. IEEE Security Privacy Magazine, 1: 33-42.

Rabiner, L.R. and B.H. Juang, 1986. An introduction to hidden Markov models. IEEE Acoustics, Speech Signal Processing Magazine, 3: 4-16.

Rabiner, L.R. and R.W. Schafer, 1978. Digital Processing of Speech Signals. Prentice-Hall, Englewood Cliffs, N.J.

Reynolds, D., W. Andrews, J. Campbell, J. Navratil, B.Peskin, A. Adami, Q. Jin, D. Klusacek, J.Abramson, R. Mihaescu, J. Godfrey, D. Jones and B. Xiang, 2003. The SuperSID project: exploiting high-level information for high-accuracy speaker recognition. In Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP) (Hong Kong), pp: 784-787.

Rose, P., 2002. Forensic Speaker Identification. Taylor & Francis, London.

Russell, M.J. and R.K. Moore, 1985. Explicit modeling of state occupancy in hidden Markov models for speech signals, Proceedings ICASSP-85 International Conference on Acoustics, Speech and Signal Processing, Institute of Electrical and Electronic Engineers, (New York), IEEE, pp: 5-8.

Schmidt-Nielsen, A. and T. Crystal, 2000. Speaker verification by human listeners: Experiments comparing human and machine performances using the NIST 1998 speaker evaluation data. Digital Signal Processing, 10: 249-266.

Srinivasan, A., K. Srinivasa Rao, D. Narasimhan and K. Kannan, 2009. Speech processing of the letter 'zha' in Tamil Language with LPC. Contemp. Eng. Sci., 2(10): 497-505.

Srinivasan, A., 2011. Speech recognition using hidden markov model. Appl. Mathe. Sci., 5(79): 3943-3948.

Sullivan, K. and J. Pelecanos, 2001. Revisiting Carl Bildt's Impostor: Would a Speaker Verification System Foil him? Proceeding on Audio and Video-Based Biometric Authentication (*AVBPA*) (Halmstad, Sweden), pp: 144-149.

Xiang, B., 2003. Text-independent speaker verification with dynamic trajectory model. IEEE Signal Proc. Lett., 10: 141-143