ASIAN JOURNAL OF PHARMACEUTICAL AND CLINICAL RESEARCH

INNOVARE ACADEMIC SCIENCES
Knowledge to Innovation

Vol 9, Suppl. 3, 2016

Online - 2455-3891
Print - 0974-2441

Research Article

VOICE RECOGNITION SECURITY SYSTEM USING MEL-FREQUENCY CEPSTRUM COEFFICIENTS

MAHALAKSHMI P[1]*, MURUGANANDAM M[2], SHARMILA A[1]

[1]School of Electrical Engineering, VIT University, Vellore - 632 014, Tamil Nadu, India. [2]Department of Electrical and Computer Engineering, Wollega University, Oromia, Ethiopia. Email: asharmila@vit.ac.in

Received: 23 June 2016, Revised and Accepted: 31 August 2016

ABSTRACT

Objective: Voice Recognition is a fascinating field spanning several areas of computer science and mathematics. Reliable speaker recognition is a hard problem, requiring a combination of many techniques; however modern methods have been able to achieve an impressive degree of accuracy. The objective of this work is to examine various speech and speaker recognition techniques and to apply them to build a simple voice recognition system.

Method: The project is implemented on software which uses different techniques such as Mel frequency Cepstrum Coefficient (MFCC), Vector Quantization (VQ) which are implemented using MATLAB.

Results: MFCC is used to extract the characteristics from the input speech signal with respect to a particular word uttered by a particular speaker. VQ codebook is generated by clustering the training feature vectors of each speaker and then stored in the speaker database.

Conclusion: Verification of the speaker is carried out using Euclidian Distance. For voice recognition we implement the MFCC approach using software platform MatlabR2013b.

Keywords: Mel-frequency cepstrum coefficient, Vector quantization, Voice recognition, Hidden Markov model, Euclidean distance.

 DOI: http://dx.doi.org/10.22159/ajpcr.2016.v9s3.13633

INTRODUCTION

The idea of human-machine interaction led to research in speech recognition. Automatic speech recognition (ASR) uses the process and related technology for converting speech signals into a sequence of words or other linguistic units by means of an algorithm implemented as a computer program. Speech understanding systems presently are capable of understanding speech input for vocabularies of thousands of words in operational environments. Speech signal conveys two important types of information: (a) Speech content and (b) the speaker identity. Speech recognizers aim to extract the lexical information from the speech signal independently of the speaker by reducing the inter-speaker variability. Speaker recognition is concerned with extracting the identity of the person. Speaker recognition has been seen an appealing research field for the last decades which still yields a number of unsolved problems. The main aim of this paper is speaker identification, which consists of comparing a speech signal from an unknown speaker to a database of known speaker.

MATERIALS AND METHODS

The performance levels of the most current speech recognizers degrade significantly when the environmental noise occurs during use. Noise has a greater effect on the higher modulating frequencies than the lower ones. Due to which pre-emphasis process is applied to the speech signal to artificially boost the higher frequencies to increase the signal to noise ratio (SNR). Gong's [1] survey paper has mentioned that the performance degradation is mainly caused by mismatches in training and operating environments. This paper surveys research results in the area of digital techniques for single microphone noisy speech recognition. The survey indicates that the essential points in noisy speech recognition consist of incorporating time and frequency correlations, giving more importance to high SNR portions of speech in decision-making and using class dependent processing and including auditory models in speech recognition.

Mel-frequency cepstrum coefficient (MFCC) and vector quantization (VQ)

The voice is a signal of infinite information. A direct analysis and synthesizing the complex voice signal is due to too much information contained in the signal A typical speech recognition system starts with a pre-processing stage, which takes a speech waveform as its input, and extracts from it feature vectors or observations which represent the information required to perform recognition. Muda *et al*. [2] have used the non-parametric method for modeling the human auditory perception system, MFCCs for extraction techniques and the non-linear sequence alignment known as dynamic time warping (DTW) introduced for features matching techniques. The DTW is for measuring the similarity between two-time series which may vary in time or speed. This technique also used to find the optimal alignment between two times series if one time series may be "warped" non-linearly by stretching or shrinking it along its time axis. DTW algorithm compares the parameters of an unknown spoken word with the parameters of one or more reference templates. The other method for pattern recognition is VQ as used in Kekre *et al*. [3]. Kekre *et al*. [3] described the development of an efficient speech recognition algorithm which used MFCC, VQ, and hidden Markov models (HMM) at different levels. In Swamy and Ramakrishnan's study [4], VQ is used in recognition phase for isolated word recognition resulting in maximum recognition rate of 85%. The main advantage of VQ over DTQ is its low computational burden. Therefore, we have used VQ in our project as it is easy to implement and takes less computational time.

HMM and other techniques

The second stage in a voice recognition system is speech recognition, or decoding, which is performed using a set of phoneme-level statistical models called HMMs. Rabiner [5] has focused on the statistical methods of Markov source or HMM and has put an attempt to carefully and methodically review the theoretical aspects of this type of statistical modeling and show how they have been applied to selected problems in machine recognition of speech. Bhupinder *et al.* [6] used HMM in recognition phase and explained its working. In the long history of speech recognition, both shallow and deep form (e.g. Recurrent nets) of artificial neural networks had been explored for many years during 80's, 90's and a few years into 2000. Most current speech recognition uses HMM to deal with the temporal variability of speech and Gaussian mixture models (GMMs) to determine how well each state of each HMM fits a frame or a short window of frames of coefficients that represent the acoustic input. Hinton *et al*. [7] mentioned the alternate way to evaluate

the fit which is to use a feed-forward neural network that takes several frames of coefficients as input and produces posterior probabilities over HMM states as output. Deep neural networks (DNN) that have many hidden layers and are trained using new methods have been shown to outperform GMMs on a variety of speech recognition benchmarks. Hinton *et al*. provided an overview of this progress and recent success in using DNNs for acoustic modeling in speech recognition. In spite of the advances accomplished throughout the last decades, ASR is still a challenging and difficult task. In particular, recognition systems based on HMM are effective under various circumstances, but they do suffer from major limitations which limit the applicability of ASR in a real-world environment. Various attempts were made to overcome these limitations using ANN but were unsuccessful in dealing with long time-sequences of the speech signals. The survey paper [8] reviews the significant use of hybrid models of ASR, which takes advantage from the properties of both HMMs and ANNs, resulting in improved flexibility and recognition performance.

IMPLEMENTATION

The following sections explain the working processes for the software simulation of the project. An explanation regarding most of the approaches has been given in brief.

Speech sample collection

Collecting speech samples is mostly concerned with recording speech samples of the different word by different speakers. However, Rabiner [6] has identified four main factors that must be considered while collecting speech samples, as they affect the training set of vectors which are used to train the VQ codebook. Those features include who the speakers are; the speaking conditions; the transducers and the transmission systems and the speech units.

- The first factor is the profile of the speakers. There were three different speakers whose speech samples were collected. Out of those three speakers, one was female and two were male.
- The second factor is the speaking condition in which the speech samples were collected, which basically refers to the environment where the samples are recorded. Here, the speech samples were collected in a noisy environment.
- The third factor is the transducers and transmission system. The samples were collected using a normal microphone.
- The fourth factor is speech units. The main speech units are specific words, for example, hello, one, two.

Table 1 gives the description of the parameters of the speech signal. A simple MATLAB function was used for recording the speech samples. However, this function requires defining certain parameters such as sampling rate in hertz and time duration in seconds. The time duration given for recording was three seconds because it was found that three seconds was more than enough for recording the speech samples. More than three seconds would result in a lot of silence time in the recorded speech sample. Speech samples were recorded and collected to be used. The collected samples were then passed through features extraction, features training, and testing stages.

Pre-emphasis

Noise has a greater effect on the higher modulating frequencies than the lower ones. Hence, the higher frequencies are artificially boosted to increase the SNR. Pre-emphasis process performs spectral flattening using a first order finite impulse response (FIR) filter.

The speech signal $s(n)$ is sent to a high-pass filter:

$$s_2(n)=s(n)-a*s(n-1) \quad (1)$$

where $s_2(n)$ is the output signal, and the value of a is usually between 0.9 and 1.0. The z-transform of the filter is:

$$H(z)=1-a*z^{-1} \quad (2)$$

Table 1: Parameters of the speech samples

Parameter	Defined value
Time length	3 seconds
Sampling rate	22050 Hz
Bits per sample	8
Frame size (N)	256
Overlap size (M)	100
Number of filters	20

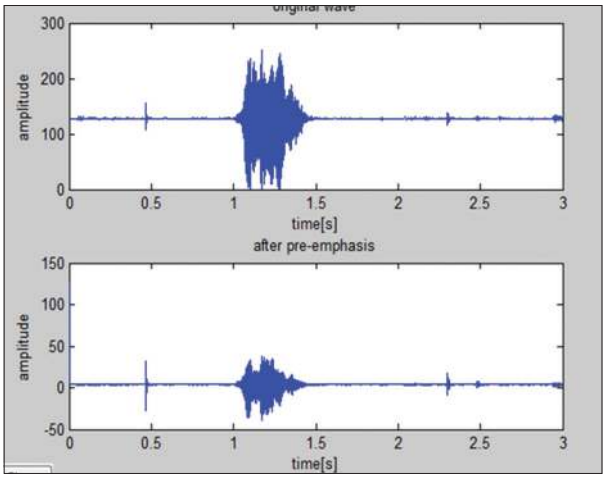

Fig. 1: Pre-emphasis of the original speech sample

Here are the graphs (Fig. 1) obtained for the original signal and the signal after pre-emphasis between time and amplitude.

Feature extraction

A typical speech recognition system starts with a pre-processing stage where, the speech waveform is taken as its input, and feature vectors are extracted from it which represents the information required to perform recognition. This stage is performed by software efficiently. A voice analysis is done on the speech input taken from the user. The design of the system involves manipulating the input voice signal. Different operations are performed on the signal such as pre-emphasis, framing, windowing, and Mel-cepstrum analysis. MFCC is used to extract features from the speech signal. It is based on the human peripheral auditory system. The human perception of the frequency contents of sounds of speech signals does not follow a linear scale. Thus, for each tone with a frequency f in Hz, a pitch is being measured on a scale called the "Mel scale." The Mel-frequency scale is a scale with a linear frequency spacing below 1000 Hz and logarithmic spacing above 1 kHz. As a reference point, 1000 Mels are defined as the pitch as 1 kHz. In the design of any voice recognition system, the selection and extraction of the best parametric representation of acoustic signals is a very important task as it significantly affects the performance of the system. A set of MFCCs will provide a compact graphical representation of the signal, which is obtained by the cosine transform of the real logarithm of the energy spectrum that is expressed on a Mel-frequency scale. The MFCCs are proved very efficient. Fig. 2 shows the steps to calculate the MFCC.

Framing and blocking

In this step, the continuous signal is blocked into small frames of N samples, with next frames separated by M samples (M<N) with this the adjacent frames are overlapped by N-M samples. The standard value after a lot of researches is taken for N=256 and M=100 so that the signal is divided into sufficient frames to get the required information. Because, if the frame size is smaller than the size taken then, the number of samples in the frames will not be enough to get the required information and with large size frames it can cause a frequent change in the information inside the frame.

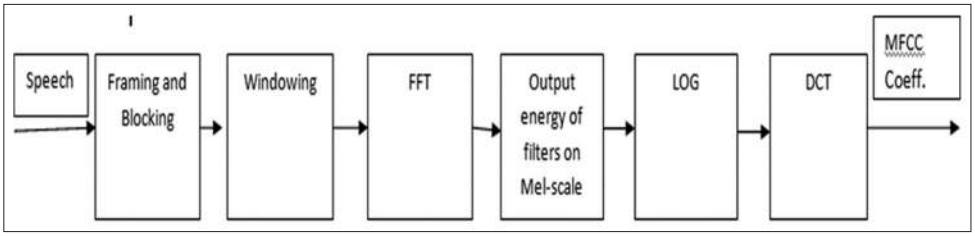

Fig. 2: Mel-frequency cepstrum coefficient calculation

As the sampling frequency is 22050 Hz so, the frame duration is:

Number of sample/sampling frequency (Hz)=256/22050=11.6 ms

And the frame rate is:

Sampling frequency/(N-M)=22050/(256-100)=141 frames per second

Hence, while working with MFCC, these parameters are very common in practice. This process of breaking up the signals into frames will continue until the whole 1D signal is broken down into small frames.

Windowing

Windowing of a signal is done to eliminate the discontinuities at the edges of the frames [9]. If, the windowing function is defined as w(n), 0<n<N-1 where, N is the number of samples in each frame, then the resulting signal will be; y(n)=x(n) w(n). Mathematically, framing is basically equivalent to multiplying the signal with a series of sliding rectangular windows. However, the use of rectangular windows may give rise to spectral leakage because the power contained in the side lobes is significantly higher. To avoid this, we have used a Hamming Window which has the form

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right); 0 \le n \le B-1 \quad (3)$$

According to Alan and Ronald [10], windowing is to be worked with short frames of the signal to select the portion of the speech signal which can be reasonably assumed to be a stationary speech signal. This is done to avoid any distortion in the spectrum or any unnatural discontinuities in the speech segment so that possible gaps between the frames are eliminated and all the parts of the speech signal are recovered.

Becchetti and Ricotti [11] have mentioned that hamming window is the most commonly used window shape in speech recognition technology, because a high resolution is not required, considering that the next block in the feature extraction processing integrates all the closest frequency lines.

The effect of windowing on the speech segment can be clearly seen in Figs. 3 and 4 where it seems to be a smooth transition towards the edge of the frame.

Fast Fourier transformation (FFT)

Note that, we use j here to denote the imaginary unit. In general, X_n's are complex numbers. The resulting sequence $\{X_n\}$ is interpreted as follows:

$$X_n = \sum_{k=0}^{N-1} x_k e^{-2\pi jkn/N}; n = 0,1,2,\ldots N-1 \quad (4)$$

The zero frequency corresponds to n=0, positive frequencies: 0<f<F_s/2 correspond to values 1≤n≤N/2-1 whereas negative frequencies – Fs/2<f<0 correspond to N/2+1≤n≤N-1.

Here, F_s denote the sampling frequency which is 22050 Hz here. The matrix obtained after performing this function contains the frames of the original speech signal filtered by hamming filter and transformed with FFT. The elements of the matrix are complex numbers and symmetrical because FFT was used to transform. By calculating DFT, we

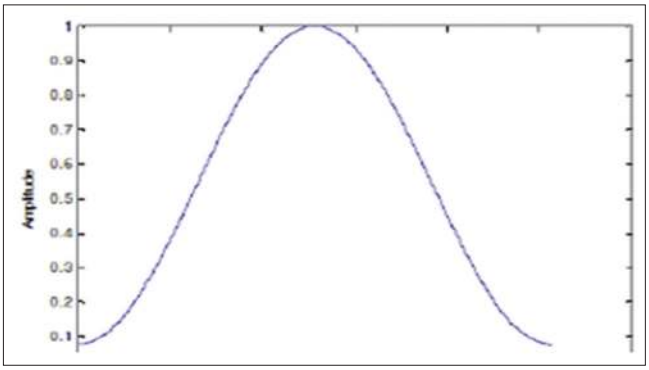

Fig. 3: Hamming window

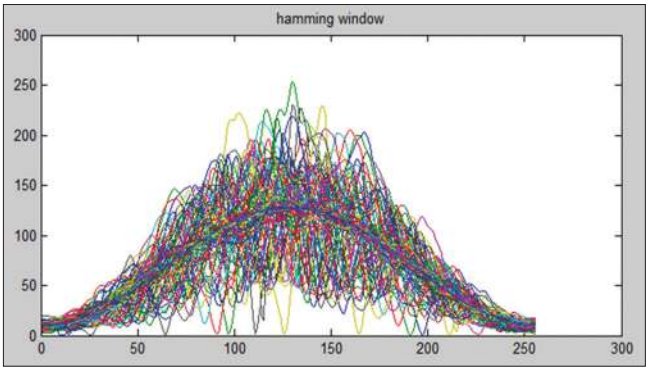

Fig. 4: Windowed speech signal

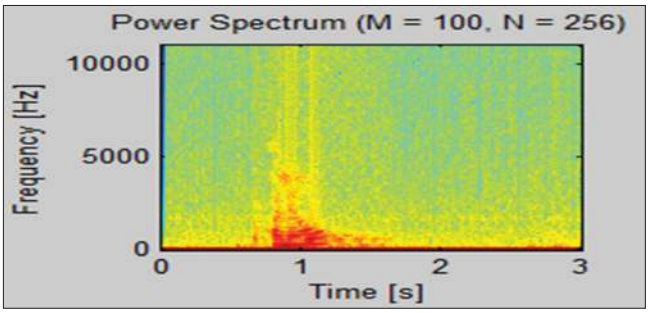

Fig. 5: Power spectrum of the speech signal

can obtain the magnitude spectrum. In the Spectrogram plot as shown in Fig. 5, the areas containing the highest level of energy are displayed in red. As we can see on the plot, the red area is located between 0 and 1 second. This plot shows that most of the energy is located in the lower frequencies (between 50 Hz and 1 kHz).

Mel scale

In this step, the magnitude spectrums calculated above are mapped on the Mel scale to know the approximation about the existing energy

at each spot with the use of Triangular overlapping window which is also known as triangular filterbank. This filter bank behaves like a succession of histograms on the spectrums. Each filter of the filter bank has a triangular frequency response. It quantifies the zone of the frequency spectrum. Therefore, a Mel scale mapping has to be done between the given real frequencies (Hz) and the required frequency scale (Mels). The filter bank is used to transform the spectrum of a signal into a representation which more closely reflects the behavior of the human ear. As the human ear favors low frequencies for analyzing speech, the filters are denser for the lower frequencies and get wider as the frequency gets higher. To mimic the human ear, the filters are linearly distributed at low frequencies (below 1 kHz). While at higher frequencies (above 1 kHz), the distribution of the filters is logarithmic as shown in Fig. 6.

Thus, with the help of filter bank and the proper spacing done by the Mel scaling, it becomes easy to get the estimation about the energies at each spot and once these energies are estimated then the log of these energies also known as Mel spectrum can be used for calculating the first 20 MFCCs using discrete cosine transformation (DCT).

Note that this filter bank is applied in the frequency domain; therefore it simply results to those triangle-shape windows on the spectrum as shown in Fig. 7. As most of the information is contained at low frequencies, therefore more number of filters can be seen at low frequency where they are closer to each other as compared to high frequencies. Each filter's magnitude frequency response is triangular in shape and equal to unity at the center frequency and decreases linearly to zero at center frequency of two adjacent filters. Then, each filter output is the sum of its filtered spectral components. After that the following equation is used to compute the Mel m_f for given frequency *f* hertz:

$$\mathrm{mf} = 2595 * \log 10\left(1 + \frac{\mathrm{f}}{700}\right) \quad (5)$$

Cepstrum

Cepstrum can be obtained by the cosine transformation of the log of the unwrapped phase of Fourier transform. Once we have filter energies, we have to take the logarithm of them. This step simply converts the multiplication of the magnitude of the Fourier transform into addition referred to as signal's logarithm Mel spectrum. This is motivated by human hearing: We do not hear loudness on a linear scale. Becchetti and Ricotti [11] have also mentioned that the magnitude and logarithm processing is performed by the human ear as well, where the magnitude discards the useless phase information while logarithm performs the

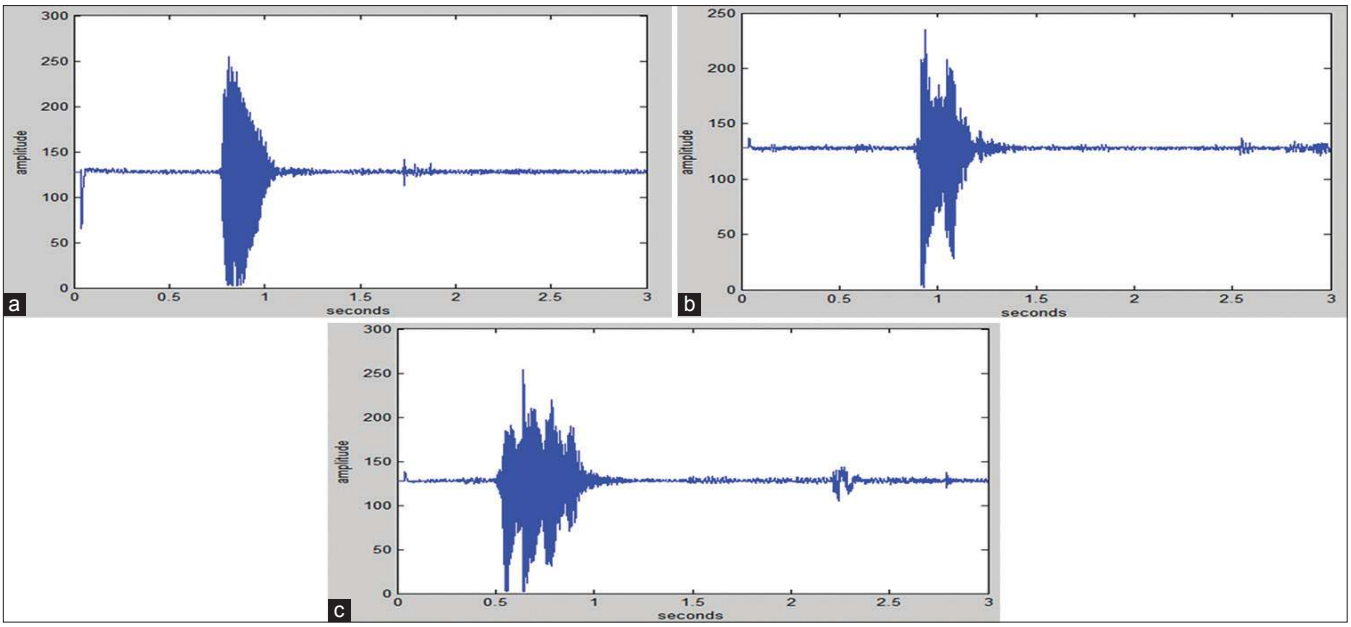

Fig. 10: (a-c) Input speech signal graphs

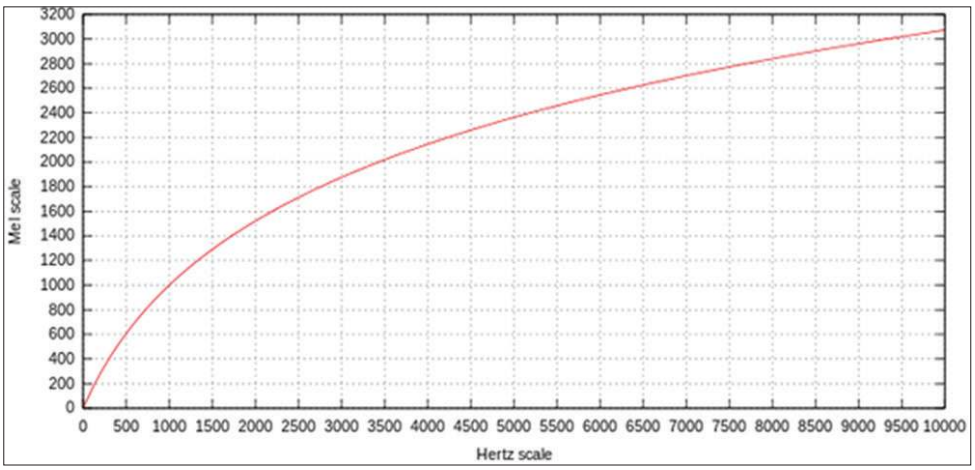

Fig. 6: Mel scale

dynamic compression to make feature extraction process less sensitive to various in dynamics. The result obtained after this step is referred as logarithm Mel spectrum.

After the log Mel spectrum is generated, DCT is used to convert the spectrum into the time domain. DCT gathers most of the information of the signal to its lower order coefficients, resulting in significant reduction in computational cost. The result of the conversion is called MFCC. The set of the coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector. In Fig. 8, there are 2 plots shown, first is the unmodified power spectrum obtained after windowing and FFT. In the first plot, most of the information is contained in the lower frequencies.

This information has been extracted and amplified in the second plot. The second plot, therefore, shows the main characteristics of the speech signal. Note that the transformation produces an acoustic vector of 20 dimensions.

VQ

A voice recognition system must be able to estimate the probability distribution of the computed feature vectors. It is impossible to store every single vector formed during the training mode since these probability distributions are defined over a high dimensional space. VQ is the most classical quantization technique from signal processing which allows the modeling of probability density functions by the distribution of prototype vectors. A large set of points is divided into smaller groups having a similar number of points closest to them. A centroid represents each group. The density matching property of VQ is very powerful, especially for identifying the density of large and high-dimensioned data. Since the data points are represented by the index of the centroid closest to them, commonly occurring data have low error while high error is seen in the rare data. Hence, VQ is also suitable for lossy data compression. Using these training, data features are clustered to form a codebook for each speaker and used to make the recognition decision. One speaker can be discriminated from another based of the location of their centroid. In the training phase, a speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors.

The distance from a vector to the closest code word of a codebook is called a VQ-distortion. In the recognition phase, the input signal of an unknown user is "vector-quantized" using the trained codebook and the total distortion is computed. The speaker corresponding to the smallest total distortion is identified. Fig. 9 displays the VQ codebook formation.

Distance measure

In the speaker recognition phase, an unknown speaker's voice is represented by a sequence of the feature vector and then, it is compared with the trained codebooks stored in the database. To identify the unknown speaker, the distortion distance between two vector sets of unknown speaker and the saved sample is measured based on minimizing the Euclidean distance. The formula used to calculate the Euclidean distance can is as following:

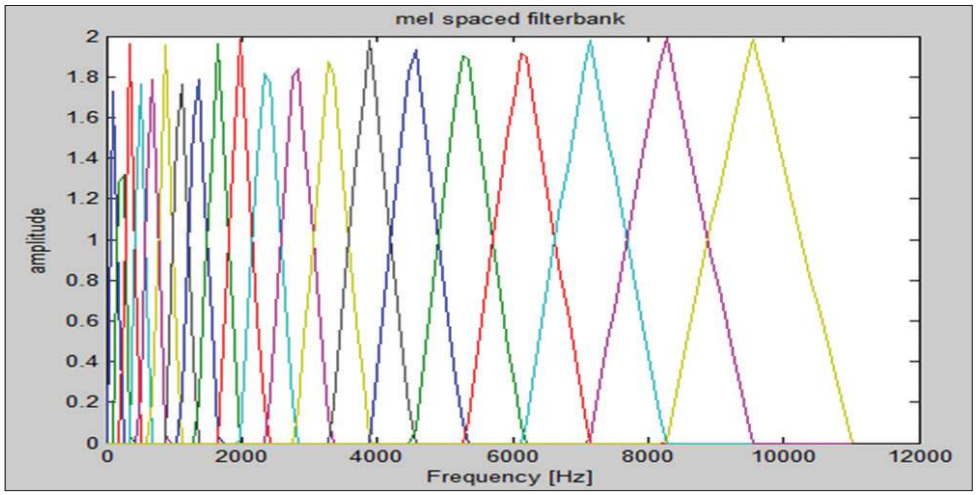

Fig. 7: Mel spaced filter bank

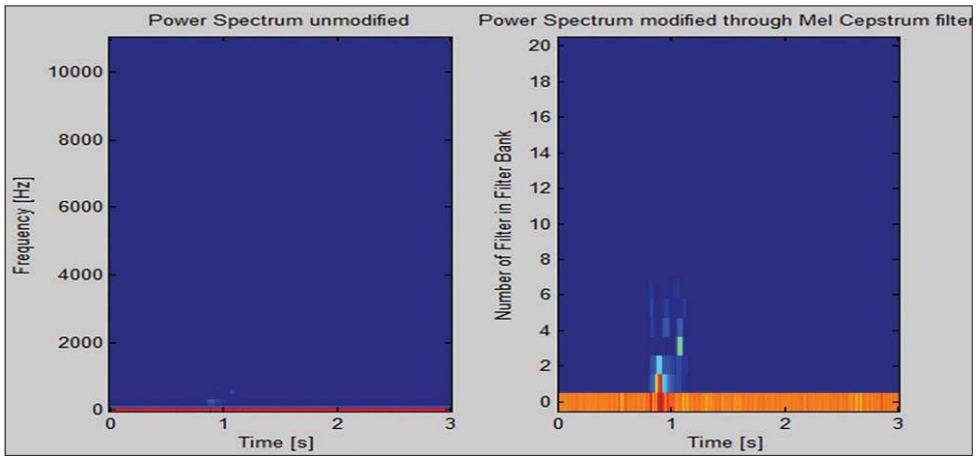

Fig. 8: Power spectrum of the speech sample (unmodified and modified through Mel cepstrum filter)

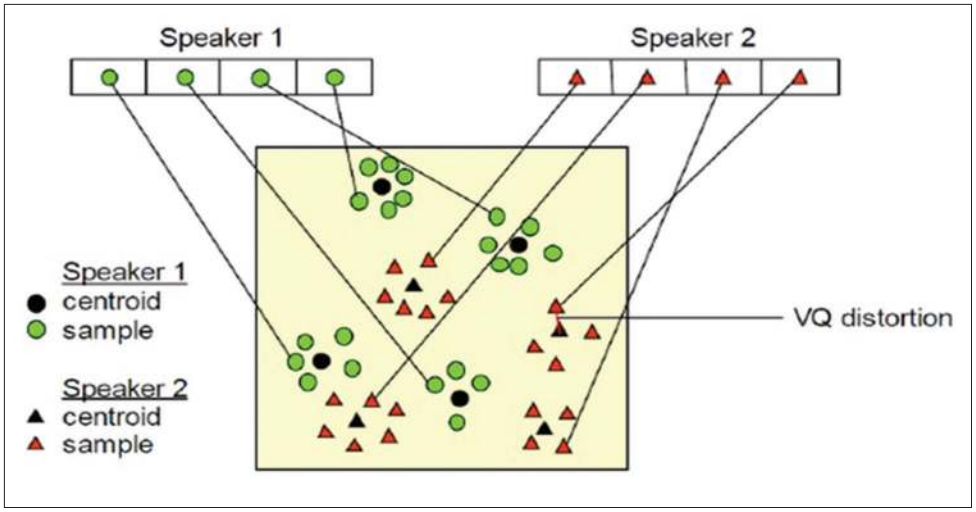

Fig. 9: Vector quantization codebook formation

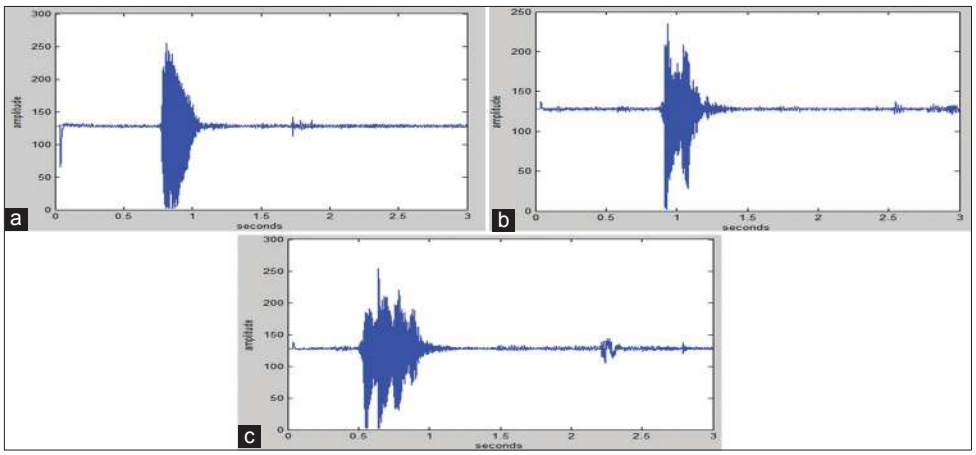

Fig. 10: (a-c) Input speech signal graphs

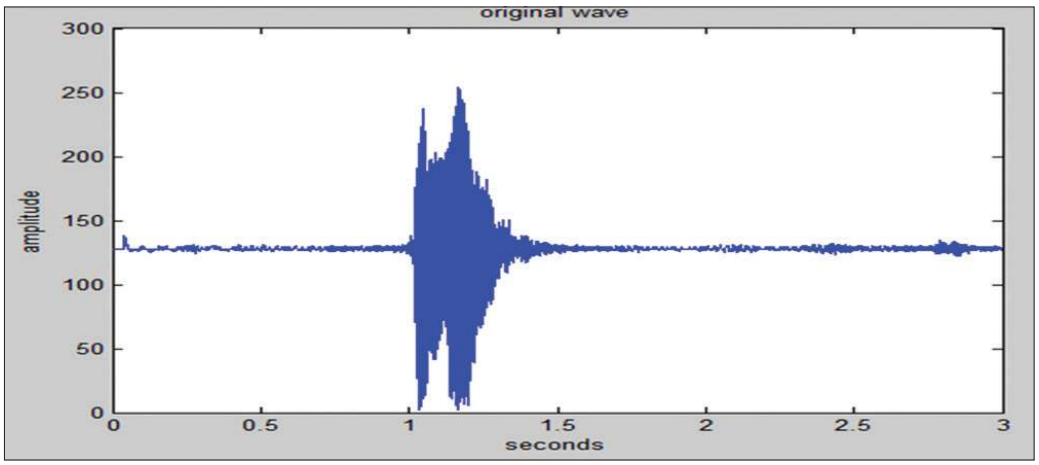

Fig. 11: Testing speech signal recorded by user 2

The Euclidean distance between two points $P = (p_1, p_2,..., p_n)$ and $Q=(q_1,.. q_2,., q_n)$

$$= \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (6)$$

The speaker with the lowest distortion distance is chosen to be identified as the unknown person.

RESULTS AND DISCUSSION

The goal of this project was to recognize an unknown speaker. The simulation has been done in MATLAB where the input speech signal is "Hello," given as

Fig: 12. The computed distortion distance

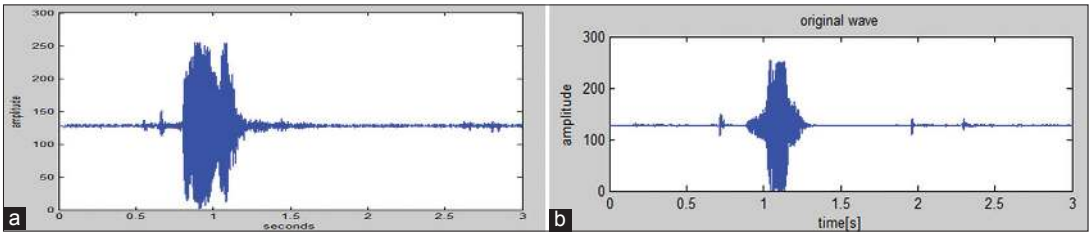

Fig. 13: Recorded signal (a) user says "one" (b) user says "hello"

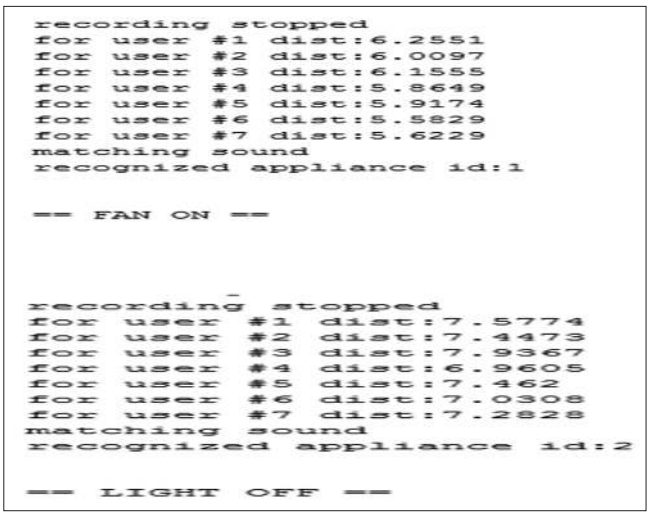

Fig. 14: Word recognition

input by three separate users of which two are male and one is female. Here are the recorded speech samples from the three separate users:

The above three graphs in Fig. 10, plotted amplitude versus time, show the recordings of the three users. The first graph Fig. 10a denotes the user 1 saying "hello." Similarly, the following two graphs Fig. 10b and c represent the users 2 and 3 saying "hello."

For testing the system, we have asked user 2 to say "hello" to find out if the system can recognize the speaker. The graph speech signal for the testing phase is shown in Fig. 11.

After completing the simulation, a distance comparison of all the different speech samples recorded with respect to the test speech sample, along with the matched user and the test speech sample graph. As seen in Fig. 12, the distortion distances between the speech sample of the saved samples and unknown speaker are given. The speaker with the lowest distortion distance is finally chosen as the unknown speaker by the system which is user 2. The system can also differentiate between two different words spoken by the same user.

Fig. 13 shows the graph of the recorded speech signal when the user 2 says "hello" and Fig. 13 shows the graphs of the speech signal when the user 2 says "one." As it can be seen that the first graph has 3 peaks as the word "hello" is spoken as "he-l-o." While the second graph has only 1 peak as the word "one" is spoken as "one." This word recognition can be used for home automation as after the user 2 was recognized. Hence, the user 2 trains the system with his voice giving application is as id 1 to "Fan" and id 2 to "Light" so, when the user 2 said "Fan," the system recognized the user input displayed "FANON" while when he said "light," the system recognized the user input as id5 and displayed "LIGHT ON." The system simulates successfully showing all the results giving 90% accuracy as the system does not respond properly when there is a lot of noise in the surrounding environment. This can be improved using HMM or DNN for pattern recognition. Fig. 14 shows the word recognition done by the system.

CONCLUSION

This paper discusses about the speaker recognition, applied to the speech of an unknown user which is done by extracting the features of the unknown speech and comparing to the already saved extracted features to identify the unknown speaker. The system simulates successfully showing all the results giving 85% accuracy as the system does not respond properly when there is a lot of noise in the surrounding environment. This can be improved using HMM or DNN for pattern recognition instead of VQ.

REFERENCES

1. Gong Y. Speech recognition in noisy environments: A Survey. Speech Commun 1995;9(3):261-91.
2. Muda L, Begam M, Elamvazuthi I. Voice recognition algorithm using Mel frequency Cepstral coefficient and dynamic time wraping techniques. J Comput 2010 2(3):138-43.
3. Kekre, HB. Athawala, AA, Sharma, GJ. Speech Recognition using Vector Quantization. International Conference & Workshop on Emerging Trends in Technology, 2011. p. 16-24.
4. Swamy S, Ramakrishnan KV. A review on speech recognition with hidden Markov model. Int J Comput Appl 2013;3(4):16-24.
5. Rabiner L. A tutorial on hidden Markov models and selected application in speech recognition. Proc IEEE 1989;77(2):257-86.
6. Bhupinder S, Neha K, Puneet K. A review on speech recognition with hidden Markov model. Int J Comput Appl 2012;2(3):16-24.
7. Hinton G, Li D, Yu D, Dahl GE. Deep neural networks for acoustic modeling in speech recognition. Signal Process Mag IEEE 2012;29(4):82-97.

8. Trentin E, Gori M. A Survey of hybrid ANN/HMM models for automatic speech recognition. Neuro Comput 2001;37(1-4):91-126.
9. Ghosh D, Debnath D, Bose S. A Comparitive study of performance of FPGA based Mel filter bank and Bark Filter Bank. Int J Artif Intell Appl 2012;3(3):37.
10. Alan V, Ronald SW, John RB. Discrete Signal Processing. New Jersey: Prentice Hall; 1999.
11. Bexhetti B, Ricotti A. Speech Recognition. New York: Wiley; 1999.