Scientific Journal of Impact Factor (SJIF): 4.14
e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 3, Issue 5, May -2016

User Authentication using Voice Identification and Verification

Dr Shiv Ratan Singh
Officer In-charge (Head) Department of ECE, Guru Nanak Dev Institute of Technology, Rohini, Delhi, India
Email: ssratan2014@gmail.com

Abstract—The use of biometric in user identification has increased tremendously in past few years. The systems based on biometrics are quite secure and safe. There are many biometric indicators that are used for user identification with their advantages and ids-advantages. In biometric system accuracy has been an issue and to improve on accuracy the proposed algorithms becoming very time consuming, and a person can't wait of 15-20 min. before he/she can be authenticated. The human voice has special features that are different in each individual. In this work basic algorithms of voice recognition systems are compared by using the mahanbolis distance criterion and accuracy is compared. The results presents in this paper clearly reveals that a fair accuracy is possible with basic algorithms and these algorithms in conjunction with other biometric identifier can provide very effective system which has a large degree of accuracy and computationally fast.

Index Terms— Face recognition, Fuzzy Logic

1. INTRODUCTION

In the recent years, Identity recognition is getting more and more useful. In order to make secure the access to lots of services or buildings the requirement of a reliable automatic user identification system is increasing. Biometric Identification [1] is the field that is establishes a link to the recognition of a person by the methodology of physiological features (voice, iris, fingerprints, face, and more). A recognition system of a person using biometric can be used for the identification or verification of a person. In the process of verification, a client guarantees a certain personality ("I am so & so "). The recognition system either agrees or disagreeswith the claim of identity (on the basis of the fact that the user is who he claims to be). At another side, for identification, there is no claim of identity. The system makes the decision on the user's identity. In this paper two voice recognition techniques are discussed and simulation results are presented in terms of percentage of recognition obtained through each of the methods.

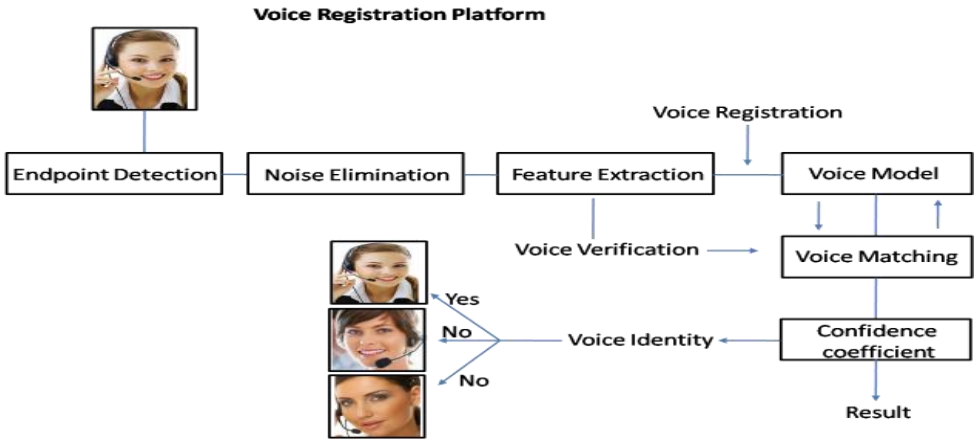


Fig. 1: Flow diagram of the speaker recognition system

2. SPEAKER RECOGNITION SYSTEM

These days the Automatic Analysis of speech is in progress, specifically in the fields of Speech Synthesis and Automatic Speech Recognition "ASR". The automatic recognition of speaker is depicted like a specific pattern recognition assignment. It associates the issues that are related to the speaker verification or identification utilizing data found as a part of the acoustic

signal: we need to recognize an individual by his voice. The use of ASR is in many fields such as military, domestic and jurisprudence applications [2-8].

2.1 Speaker Identification "SI"

The recognition a person among numerous speakers with the comparison of his vocal expression with known references is referred as the speaker identification. Figure 2 illustrates a diagrammatic point of view of a sequence of word given in entrance of the ASR system. A comparison of the arrangement of word is made with a speaker characteristic reference for each known speaker. The yield datum of the system (ASR) will be the **person's individuality** of the speaker whose reference is the closest to the sequence of word.

2.2 Speaker Verification "SV"

The testing for authentication of the speaker comprises in, after the declination of the speaker's identity, the adequacy check of its vocal message with the speaker acoustic reference to establish the claims (Figure 3). An estimation of similarity is computed between the above mentioned reference and the vocal message and then compared with a thresholdfixed by the designer according requirement. If we get a higher measurement of similarity than the threshold the speaker is accepted, else the speaker is rejected considering him as an impostor.

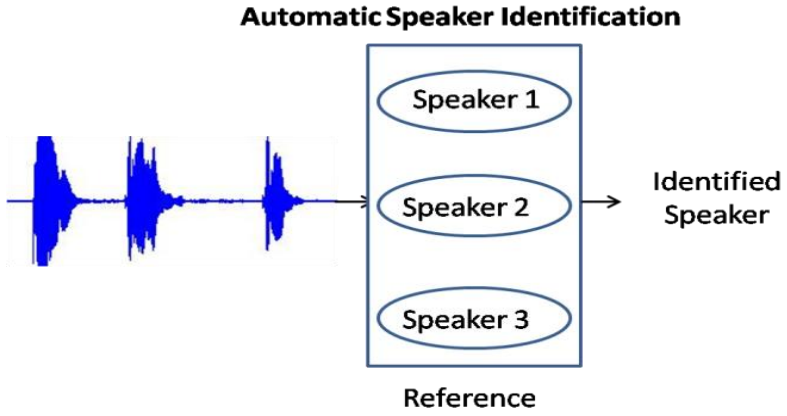


Fig. 2: Schematic of Speaker identification system

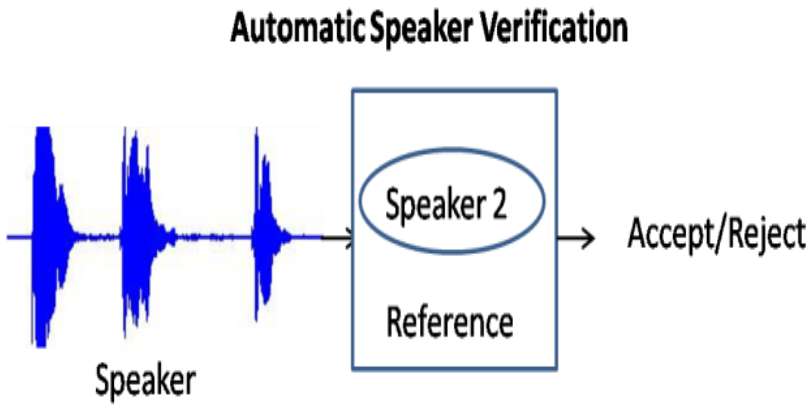


Fig. 3: Schematic of Speaker verification system

3. SPEAKER IDENTIFICATION

As each and every speaker has a unique voice feature, this unique feature of voice can be extract by various techniques. Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficient (MFCC) techniques are often used for this purpose [1, 3].

A. Linear Predictive Coding

It is a method where a speech signal by guessing the formants and then removing their effects from the speech signal, and finally intensity and frequency of the remaining buzz is estimated. The process namely inverse filteringis used to removing the formants, and the resulted signal is known as residue. In this system, all sample of the signal is represented as a linear combination of the previous samples, therefore known as linear predictive coding.The coefficients of the difference equation representthe formants. There are four following basic steps of LPC processor are as [3]:

1. ***Pre-emphasis*:**

This step of LPCdigitize the speech signal, *s*(*n*), is allowed to pass through a low order digital system, this process spectrally flatten thesignal and to make it less prone to finite precision effects later in the signal processing. The input and output *s*(*n*) of thepre-emphasizer network is related,by difference equation:

$$\tilde{s}(n) = s(n) - \tilde{a}s(n-1) \tag{1}$$

2**. *Frame Blocking*:**

In this step of LPC, the output of pre-emphasisstep $\tilde{s}(n)$ is divided into frames and each frame contains*N* samples, and adjacent frames are separated by *B*samples. If $x_l(n)$ is the l^{th} frame of speech, and entire speech signal has *I*frames, then

$$x_l(n) = \tilde{s}(BI + n) \tag{2}$$

*3. **Windowing*:**

In this step minimization of signal discontinuities at the beginning and end of each frametacks place, the process is named as windowing. If, window is definedas *w*(*n*), $0 \le n \le N-1$, then the result of windowing is the signal:

$$\tilde{x}_1(n) = x_1(n)w(n) \text{ where, } 0 \le n \le N-1. \tag{3}$$

***4. Autocorrelation Analysis*:** The next step is to auto correlate each frame of windowed signal. Where the value of '*p*'which provides the highest autocorrelation value, is the order of the LPC.

$$w(n) = 0.54 - 0.46\cos\left[\frac{2\pi n}{N-1}\right] \text{ where, } 0 \le n \le N-1 \tag{4}$$

$$r_l(m) = \sum_{n=0}^{N-1-m} \tilde{x}_l(n)\tilde{x}_l(n+m) \text{ , where } m = 0,1,.....p \tag{5}$$

***LPC Analysis*:**

Under thenext step of processing the signal is the LPC analysis, each frame of $p + 1$ autocorrelations converts into LPC parameter. The used algorithm for this purpose is known as Durbin's algorithm, which is mentioned as under

$$E^{(0)} = r(0)$$

$E^{(0)}$ energy of the speech signal.

$$k_i = \frac{r(i) - \sum_{j=1}^{t-1} \alpha_j^{i-1} r(|i-j|)}{E^{i-1}} \qquad 1 \le i \le p \tag{6}$$

$$\alpha_i^{(i)} = k_i \;\; \alpha_j^i = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \qquad 1 \le j \le i = 1 \tag{7}$$

$$E^{(i)} = \left(1 - k_i^2\right) E^{i-1} \tag{8}$$

The LPC coefficient, a_m is given as $a = a_m^{(p)}$ If we solv the above equations recursively for $i = 0,1,.....p$

$$c_m = \sum_{k=m-p}^{m-1} \left(\frac{k}{m}\right) . c_k . a_{m-k} \qquad m > p \tag{9}$$

B. Mel Frequency Cepstrum Coefficients

One of the most commonly used feature extraction method in speech recognition is Mel Frequency Cepstral Coefficients (MFCC). A technique using which feature vectors are extracted from the frequency spectra of the windowed speech frames is called FFT [2,4,8].The Mel frequency filter bank contains a series of triangular bandpass filters. The filter bank is based on a non-linear frequency scale called the mel-scale. A mathematical relationship between the Mel scale and the linear frequency scale is as under [4]

$$f_{Mel} = 1127.01 \ln\left(\frac{f}{700} + 1\right) \tag{10}$$

The Mel frequency filter bank consist of triangular bandpass filters in such a way that Frequency rages are overlap with previous and next frequency range of filtersin the pattern asupper boundary of one filter issituated at the center frequency of the nextfilter and the lower boundary situated in the center frequency of the previous filter. Corresponding to a logarithmic scaling of therepetition frequency, a fixed frequency resolution in the Mel scale is computed using the formula mentioned under

$\Delta f_{mel} = \frac{m\left(f_{H(Mel)} + f_{L(Mel)}\right)}{M+1}$ where $f_{H(Mel)}$ is represented the highest frequency of the filter bank on the Mel scale, computed from using equation given above, $f_{L(Mel)}$ is represented the lowest frequency in Mel scale, and *M* is the number of filter bank. The center frequencies on the Mel scale are given by:

$f_{cm(Mei)} = f_{L(Mel)} + \frac{m\left(f_{H(Mel)} + f_{L(Mel)}\right)}{M+1}$, Where, $1 \le m \le M$.

The center frequencies in Hertz, is given by $f_{cm} = 700\left(e^{\frac{f_{cm(Mel)}}{1127.01}} - 1\right)$

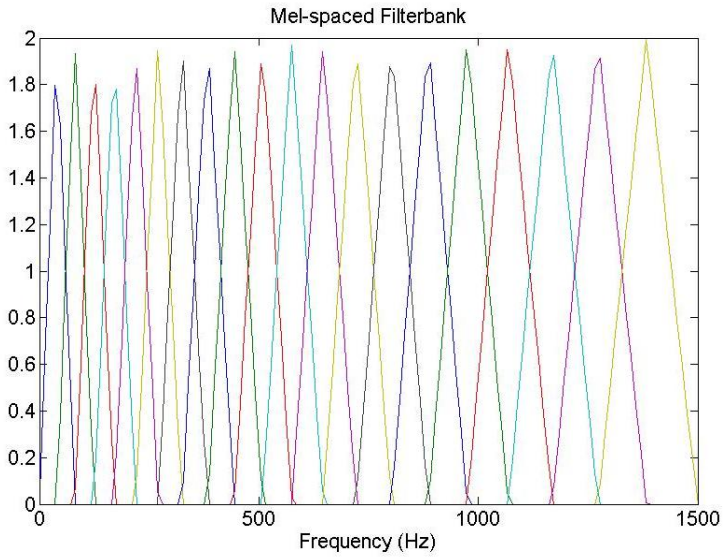


Fig. 4: Mel Frequency Cepstrum filter

Above Equation is inserted into equation of fmel to give the Mel filter bank. Finally, the MFCCs are computed using the discrete cosine transforms $c(l)=\sum_{m-1}^{M} X^{t}(m)\cos\left(l\frac{\pi}{M}\left(m-\frac{1}{2}\right)\right)$.

For $l=1,2,3.....\mathrm{M}$,where $c(l)$ is the l^{th} MFCC.

The time derivative is approximated by a linear regression coefficient over a finite window, which is defined as

$$\square c_t(l)=\left[\sum_{K-2}^{2} k\ c_{t-k}(m)\right].G,\quad 1\le l\le M \tag{11}$$

Where is the l^{th} cepstral coefficient at time *t* and *G*is a constant used to make the variances of the derivative terms equal to those with the original cepstral coefficients.

4. SPEAKER RECOGNITION

In the speaker recognition, voice of unknown speaker is represented as a sequence of feature vector $\{x_1, x_2 ... x_i\}$ This feature vector is compared with the codebooks from the database. On the basis of the distortion distance of two vector sets an unknown speaker can be identified. This is based on minimizing the Euclidean distance. The Euclidean distance is the "ordinary" distance between the two points that can be measure with a ruler.

Using Vector Quantization algorithm the extracted feature of a speaker are vector quantized.In both, training and testing processes VectorQuantization (VQ) is used for feature extraction. VQ provides thegood picture of spectral information in the speech signal by mapping the vectors from large vector space to a finite number of regions in the space called clusters[6], [8]. Each cluster can be represented by its center called a codeword.

After feature extraction, to identify the unknown speakerthe matching process takes place. In this processextracted features is compared with the databasefeature. Vector quantization is based on K-means algorithm, thus it only based on the mean values. A more generalized approach known as Gaussian mixture with rely on both mean and variance provide more accurate cluster forming.

A. The GMM Formulation

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities [9]. Commonly used parametric model of the probability distribution of continuous measurements or

features in a biometric system are GMM. Features in biometric system like vocal-tract related spectral features in a speaker recognition systemare in used.

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. Using the iterative Expectation-Maximization (EM) algorithm or Maximum *A Posteriori* (MAP) estimation,theGMM parameters are estimated from training data from well-trained prior model. The Gaussian mixture model for speech representation assumes that a M component mixture model contains windowing function weights $p(\omega_i)$ and the mixture components in the input voice samplecontains Gaussian components given by,

$$p(x/\omega_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp - \left(\frac{x-\mu_i}{\sqrt{2\sigma_i^2}} \right)^2 \quad (12)$$

Here, for the component, $p(\omega_i)$ the mean and the standard deviation represented by μ_i and σ_i^2 respectively.

The iterative Expectation-Minimization algorithm (EM) is used for training as well as matching purposes.
The E-step comprised of,

(1) The total likelihood that g is the Gaussian PDF

$$t_j = \sum_{i=1}^{N} p_t g(x_i, \mu_i, \Sigma_i) \text{ where } j = 1 \text{ to } N \quad (13)$$

(2) The normalized likelihoods

$$n_{ij} = p_t \frac{g(x_i, \mu_i, \Sigma_i)}{t_j} \text{ where } i = 1 \text{ to } j\text{-}1 \quad (14)$$

(3)The notional counts $C_i = \sum_{j=1}^{N} n_{ij}$

(4)The notional means

$$\hat{x}_i = \sum_{j=1}^{N} (n_{ij} x_i) \frac{1}{c_i} \quad (15)$$

(5) The notional sums of squares

$$SS_i^{pq} = \sum_{j=1}^{N} (n_{ij} x_i^p x_i^q) \frac{1}{c_i} ; p, q = 1,2,3.. \quad (16)$$

For multicomponent Gaussian
The M-step is an updating step wherein,

(1) $p_i = \frac{C_i}{S_i}$

(2) $\mu_i = \hat{x}_i$

(3) $\sum_i^{pq} = SS_i^{pq} - x_i^p x_i^q$, $i = 1$ to N

Thus, the following underlying probabilistic measure became the underlying model for the input utterances during training which then can be stored in the system as a reference for matching when the system was ported into real-time environment,

Mixture weight: $\bar{w}_i = \frac{1}{T}\sum_{t=1}^{T} P(i / x_t)$ (17)

Mean: $\mu_i = \frac{\sum_{t=1}^{T} P(i / x_t, \lambda) x_t}{\sum_{t=1}^{T} P(i / x_t, \lambda)}$ (18)

Diagonal co-variance: $\sigma_i^2 = \frac{\sum_{t=1}^{T} P(i / x_t, \lambda) x_t^2}{\sum_{t=1}^{T} P(i / x_t, \lambda)} - \mu_i^2$ (19)

And a posteriori probability for component is given by:

$$P(x_t, \lambda) = \frac{w_i P(x_t / \mu_i, \sigma_i^2)}{\sum_{k=1}^{M} w_k P(x_t / \mu_i, \sigma_i^2)} \quad (20)$$

B. Mahanalobis Distance

A modified version of the Euclidean distance is known as Mahanalobis distance.Mahalanobis distance is unitless and scale-invariant, and takes into account the correlations of the data set.Mahanalobis distance, each dimension is given a weight which is inversely proportional to its variance. The covariance matrices of the random variables are taking into consideration during the distance computation. The Mahanalobis distance is expressed as:

$$Q_{MH} = \left(\mu_x - \mu_y\right)^T \Sigma^{-1} \left(\mu_x - \mu_y\right) \quad (21)$$

Here Σ is the covariance matrix of the two random variables combined.

5. SIMULATION AND RESULTS

In the case of speaker identification, there are two techniques have been used (1) MFCC (2) Distance Minimum techniques. These two techniques provided more efficient for speaker identification system. The most efficient algorithm uses for speechrecognition is HMM Algorithm. It is found thatthe efficiency of speech recognition scores improves by using speaker recognition module. The MATLAb has been used for the coding of all the techniques mentioned above.
In the test system voice of three persons is recorded using the MATLAB code. Each person speak six words: ‘A’, ‘B’, ‘C’, ‘Five’, ‘Point’ and ‘V’. Each person recorded 25 voice samples of each word.

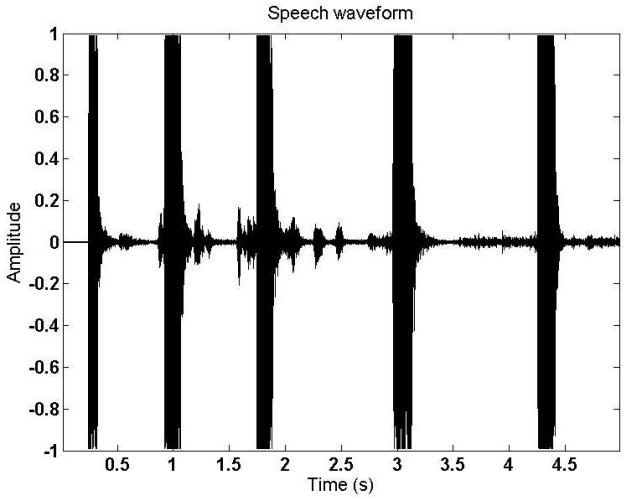


Fig. 5: Speech Signal for uttered word '3'.

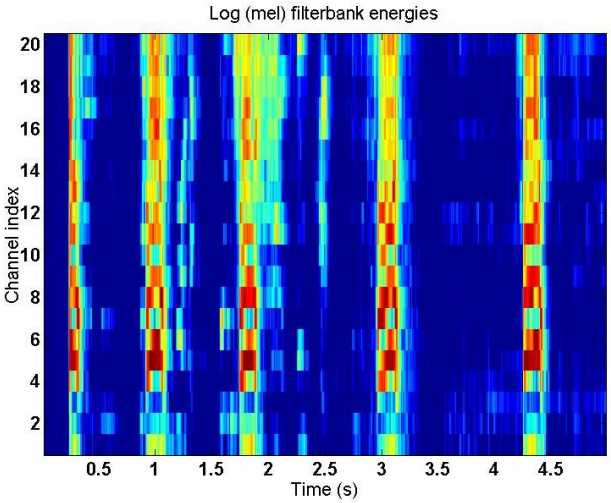


Fig. 6: Log(mel) filter bank energies

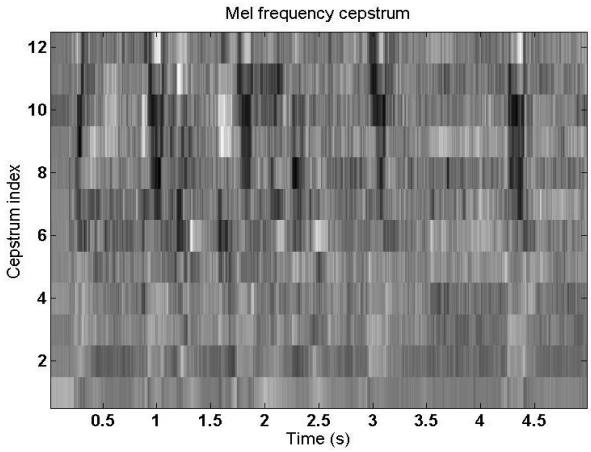


Fig. 7: Mel frequency cepstrum

In figure 5 to 7, MFCC results are presented. In figure 5, speech waveform for the uttered word '3' is shown. Log(mel) filter bank energies is shown in figure 6, and figure 7, Mel frequency cepstrum analysis is plotted.
In figure 8 and 9, windowed and autocorrelation speech signal with varying number of LPC order is shown.

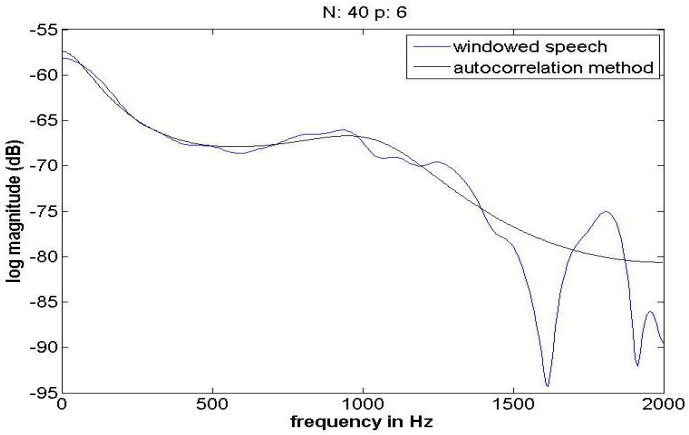


Fig. 8: Windowed and autocorrelation speech signal with LPC order 6.

In figure 8, the LPC order is 6 while number of samples as 40. It is clear that with low order LPC the windowed and speech signal are far way from each other, but as the filter order is increased to 64, they exactly match (Fig.9).

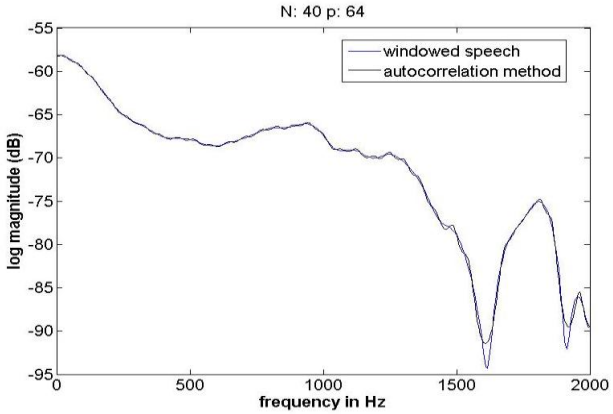


Fig. 9: Windowed and autocorrelation speech signal with LPC order 64.

Figure 10 to 12 are presented using GMModel. In these figure two Gaussian pdf with different mean and co-variance matrix are presented. It is clear from these figures that GMM has very good capability to isolate data by dividing then into separate non-overlapping clusters, thus enables the separation of speech.

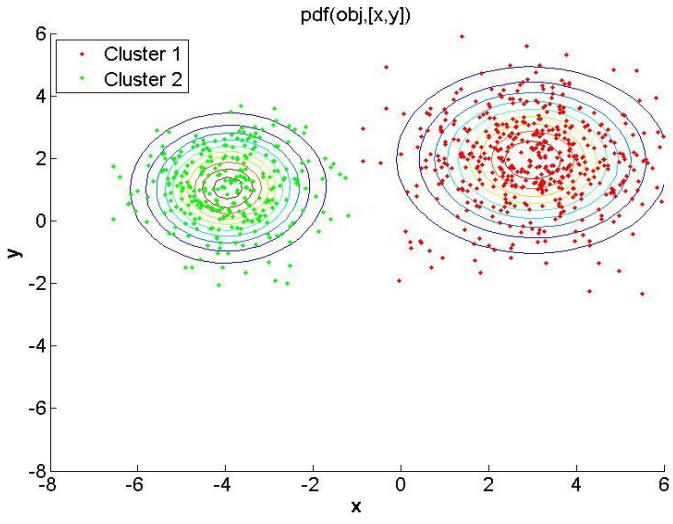


Fig. 10: GMM as separate cluster and contour

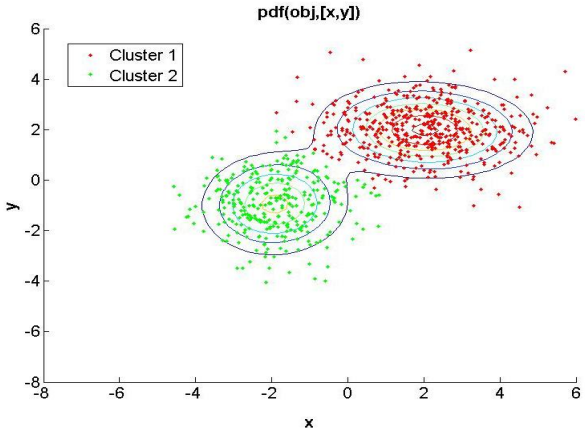


Fig. 11: GMM as common counter and separate cluster

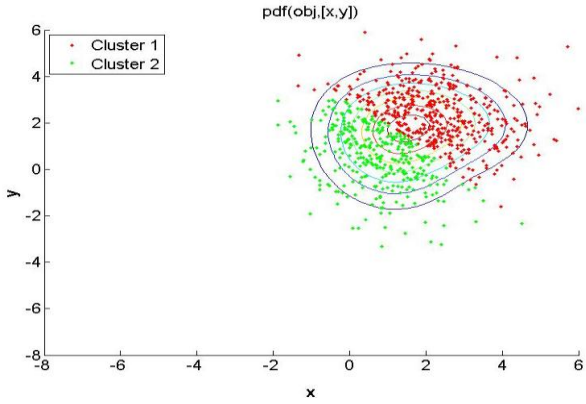
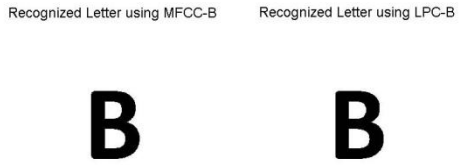


Fig. 12: GMM as common counter and common cluster

In figure 13, the obtained results for MFCC and LPC is shown.



(a)

(b)

Fig. 13: MFCC and LPC recognition

In the results it is shown that letter 'B' is recognised by both the MFCC and LPC. However, 'Five' is recognised by LPC only.

Table 1 : Comparison of MFCC and LPC

	Speaker 1	Speaker 2	Speaker 3
MFCC	69.06%	71.09%	73.04%
LPC	98.41%	98.89%	99.01%

6. CONCLUSIONS

There are several solutionsto recognition of human speech, but still having a scope as none of the current methods are fast and precise enough to be comparable with recognition capability of human beings. Although there are various methods but among all these methods, few of themare used in real ARS.This paper is focused in the direction ofan effective method for feature extraction and isolated word recognition system based on Mel-Frequency Cepstral Coefficient (MFCC) and LPC as recognition method. It also compared the recognition systems of LPC and MFCC features. It has been found that the efficiency of the PLC methods is nearly 99%. The 99% of recognition is fairly good in most of the application, but in some application even more accuracy is desirable. In such application where nearly 99.9% accuracy is desirable multimodal biometrics such as combination of face, finger and voice can be combined together. In the future work face, finger and voice can recognition will be combined together to obtain very high degree of rank one recognition.

7. REFERENCES

[1] LindasalwaMuda, MumtajBegam and Elamvazuthi.,"Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and DTW Techniques ",Journal of Computing, Volume 2, Issue 3, March 2010.
[2] Mahdi Shaneh and AzizollahTaheri ,"Voice Command Recognition System based on MFCC and VQ Algorithms", World Academy of Science, Engineering and Technology Journal , 2009.
[3] MostafaHydari, Mohammad Reza Karami, EhsanNadernejad, Speech Signals Enhancement Using LPC Analysis based on Inverse Fourier Methods,Contemporary Engineering Sciences, Vol. 2, 2009,no. 1, 1 - 15
[4] RemziSerdarKurcan, "Isolated word recognition from in-ear microphone data using hidden markov models (hmm)", Master's Thesis, 2006.
[5] Nikolai Shokhirev ,"Hidden Markov Models ", 2010.
[6] L.R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in SpeechRecognition", Proceedings of the IEEE Journal, Feb 1989, Vol 77, Issue: 2.
[7] Suma Swamy, Manasa S, Mani Sharma, Nithya A.S, Roopa K.S and K.V Ramakrishnan, "An Improved Speech Recognition System", LNICST Springer Journal, 2013.
[8] Kevin M. Indrebo, Richard J. Povinelli, Michael T. Johnson, IEEE Minimum Mean-Squared Error Estimation of Mel-Frequency Cepstral Coefficients Using a Novel Distortion Model IEEE Transactionson audio, speech, and language processing, vol. 16, no. 8, November 2008
[9]Matthew Nicholas Stuttle, "A Gaussian Mixture Model Spectral Representation for SpeechRecognition". Hughes Hall and Cambridge University Engineering Department. July 2003.

