Research Article

Literati Journals
Empowering Researchers

User Authentication Using Text - Prompted Technique

Laxmi Doddappagol,[1*] Geetha B[1]

Abstract: The authentication of the user through voice is one of the emerging technologies in today's internet based world. Recognition of the person is process to identify the individual based on characteristic features available in the speech signals. The methods of speaker recognition can be divided into text-dependent (e.g. fixed phrase/word), text-independent (e.g. No specific phrase/word) and text-prompted. Text-prompted is one where the text on the screen is prompted and the imposter cannot know in advance the uttered phrase/word. The system gives access to the genuine users and accepts the input utterance else rejects the imposters. Hence, this technique alleviates the drawbacks of other two techniques. This project is implemented based on the techniques of text prompted. The database of 25 users is being collected. Each user utters the digits from 0-9 in English language. Various preprocessing techniques such as hamming window, normalization, end point detection and speech separation, pre emphasis are performed. The features such a Mel frequency cepstral coefficients, pitch and formant features are extracted. Support Vector Machine (SVM) is used as a classifier.The work is carried in order to provide the liveness detection of users through user voice. The accuracy of 92% to 88.7% is achieved.

Asian Journal of Engineering and Technology Innovation
Volume 4, Issue 7
Published on: 7/05/2016

Cite this article as: Laxmi Doddappagol, Geetha B. User Authentication Using Text - Prompted Technique. Asian Journal of Engineering and Technology Innovation, Vol 4(7): 15-21, 2016.

INTRODUCTION

The majority of the current customary authentication systems utilized for human PC interface is based on the secret key and client name for confirmation. Today's reality is web based, henceforth all utilization of the e-business applications consequently enhancing the heartiness of the framework and give a block to burglary is a critical errand. Since the conventional methods utilized secret word, Personal Identification Number (PIN) is in a verge to dissipate. Thus, the Biometric systems have been emerged. There are numerous sorts of Biometrics. Few among them are palm print, finger-print, face recognition, iris, hand geometry and voice. The heartiness given by these biometrics is not overwhelming due to the spoofing attacks. The voice biometrics becomes base for different speech systems such as speech coding, speech synthesis and speaker recognition etc. Speaker recognition is a standout amongst the most valuable and mainstream biometric acknowledgment methods on the planet particularly identified with zones in which security is a note worthy concern. The capacity of perceiving voice of those natural to us is a key piece of oral correspondence between people. Speaker acknowledgment is one such innovation that makes new administrations which parallel serves to lead more secure life. Another critical use of speaker recognition is for forensic purposes. Speaker recognition (identification/verification) may be categorized into closed set and open set. Speaker recognition task can be classified into Speaker Identification and Speaker verification i.e, to distinguish a specific individual or to confirm the individual's guaranteed character. In Speaker Identification, the system chooses who the individual is, the thing that gathering the individual is an individual from, or (in the open-set case) that the individual is unknown. Speaker Verification is characterized as choosing if a speaker is who he claims to be.

All speaker acknowledgment frameworks need to serve two recognized stages. The first is alluded to the enrolment or preparing stage, while the second one is alluded to as the operational or testing stage. In the preparation stage, each enrolled speaker needs to give tests of their discourse so that the framework can fabricate or prepare a reference model for that speaker. In the testing stage, the data discourse is coordinated with put away reference model(s) and an acknowledgment choice is made.

In this paper we have presented a robust approach for text prompted voice recognition. The recognition system involves MFCC (Mel Frequency Cepstrum Co-efficients), Pitch and Formant technique for extracting features. These features are utilized for training the classifier in the training stage. In the testing stage the database serves as an input for SVM classifier which recognizes the speaker based on his or her voice.

RELATED WORKS

[1]Reva Institute of Technology and Management, Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Near Border Security Bustop, Bengaluru, Karnataka-560064, India.
E-mail: ashwin@revainstitution.org
*Corresponding author

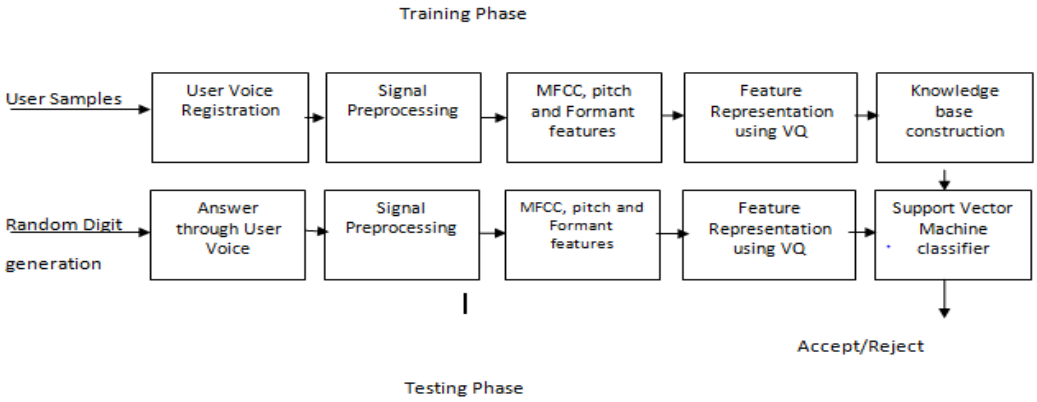

Figure 1: Proposed Methodology

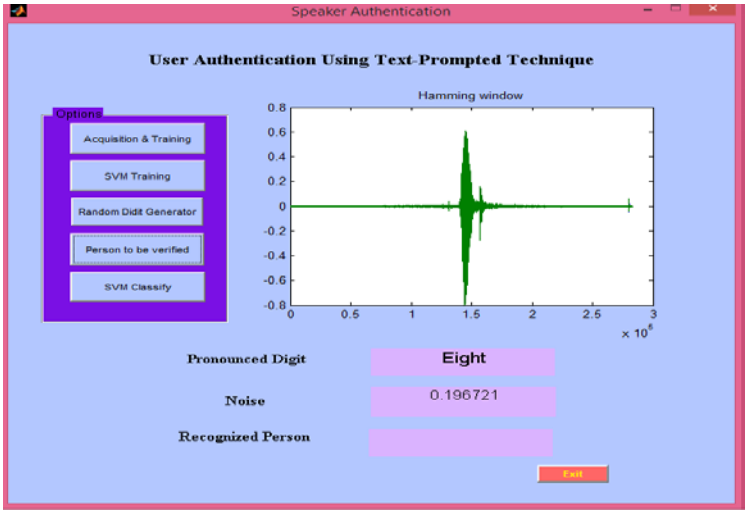

Figure 2: Front end results

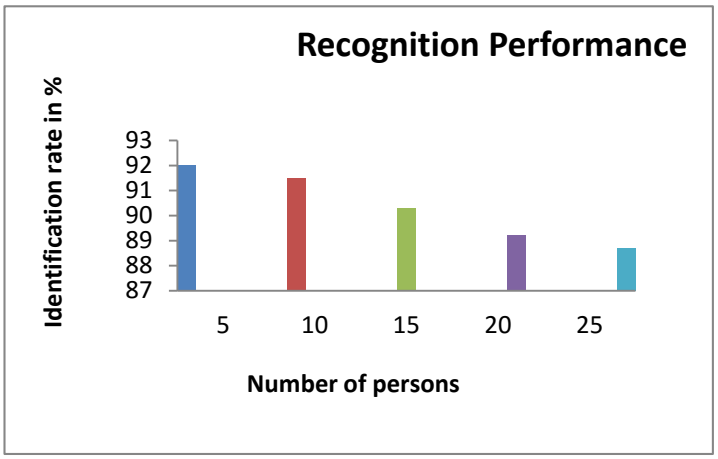

Figure 3: Recognition Performance of different set of users

Many more methods have been presented for text prompted speaker authentication. Every method employs different strategy. Review of some prominent solutions is presented.

Speaker identification system using two different approaches are described in [1]. The first approach is based on vector quantization and the LBG algorithm, while the second approach uses self-organizing maps (SOM). The preprocessing steps like pre-emphasis and framing are carried to the speech signals. The features such as MFC and LFC of 13 counts are extracted. Polynomial classifiers are used. In [2] the preprocessing of the speech signals is done using pre-emphasis at 44.1 kHz. The features like Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Cepstral Coefficients (LPCCs) and Short-Time Fourier Transforms (STFT) are

extracted. CHAINS corpus is used as the database. In [3] the pre-processing such as pre emphasizing, DC removal and signal normalization are performed. The robust speaker verification system is constructed successfully by implementing Multitaper MFCC feature extraction method & GMM classifier by reducing variance of the extracted MFCC.

In [4] author describes the preprocessing techniques such as pre-emphasis, framing and windowing are performed. The speaker features are extracted by using LPC and MFC methods and these features are trained and tested by using back propagation neural network (BPNN). Further, the preprocessing steps like framing, windowing, fast fourier transform (FFT) and Mel-frequency wrapping are carried. Extracting the feature of speech by using MFCC and compare those with the stored speakers extracted features. The speaker is modeled using Vector Quantization (VQ) due to high accuracy and K-means algorithm is used as a classifier in [5].

In [6] the preprocessing techniques such as FFT, Pre-emphasis, discrete cosine transform (DCT) are used. The features such as LPCC, MFCC and MACV are extracted by applying from spectrum. Gaussian mixture model (GMM) is used as a classifier. [7] Describes VQ which reduces the relative number of features during feature extraction in both the training and testing phases. The codebook is generated for each user using Kekre's Fast Codebook Generation (KFCG) algorithm. In [8] the Gaussian mixture model is used as a classifier. In order to improve the efficiency MFCC and LPC are extracted. In [9] the Voice Activity Detection is used as a preprocessing step to improve the performance. MFCC, LPCC and LPC methods are applied. Artificial neural network (ANN) is used as a classifier. [10] Describes a text dependent recognition model. Here the users are asked to speak fixed phrase i.e. “Basaveshwar Engineering College”. Automatic Speaker recognition System recognizes or identifies a person from a spoken phrase, which includes the characteristics of voice which differs from person to person. MFCC’s are extracted and Vector Quantization (VQ) technique is applied. ANN is used as a classifier. The experiment was carried out with 50 users with 10 phrases from each user and is followed by two phases. The accuracy of 92% to 72% is achieved.

PROPOSED METHODOLOGY

Speaker recognition (i.e. Identification/verification) systems contain pre-processing, feature extraction and matching modules. Pre-processing is one where the silence detection, noise removal, pre-emphasis tasks are performed. Feature extraction is the process that extracts a data from the voice signal that can later be used to represent each speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing extracted features from his/her voice input with the ones from a set of known speakers. The random digit generator is used to prompt the random digits during testing. The block diagram for our work is shown in Figure 1.

Voice Sample Acquisition

Voice sample is the action of retrieving voice from a user using digital voice recorder. The voice corpus i.e. database is consisting of the 1800 voice samples collected from 25 persons, where voice samples of the each person is consists of the 100 files per person with different sessions. Every user is uttered the text dependent voice samples in English from number 0 to 9. In order to check the efficiency of the user. The recording was carried out in two different sessions. In the first session 5 samples of each digit of particular users were collected and in the second session the next five samples of the digits were collected. Hence a total of 10 samples for each digit/user is used. Among them 6 voice samples are used for training and 4 samples are being used for testing. The voice samples are stored in different folders. The user voice signal is recorded from the Sony audio recorder ICD-UX 533f. It is PC compatible with Microsoft Windows and has 288 hours of recording time (LP mode). The storage capacity of 1 GB Flash Memory and Stereo recording with external microphone with correct dictation in playback and has playback format MP3/AAC/WMA/WAV, low cut filter, digital pitch control, track mark, sync recording, recording frequency is 44.1kHz with a sampling rate of 16-bits,mono recording channel.

Signal Pre-Processing

Pre-processing is a technique to enhance raw voice samples by removing noise by filtering and concatenating the bit rate of the voice sample and segmentation into frames. The speech signal is considered to be a slowly varying signal with respect to time. The voice sample i.e. input speech needs to undergo various signal conditioning steps before being subjected to the feature extraction methods.

1. Hamming Window: The process of windowing helps to smooth the data. Hamming window is selected for this purpose because it greatly reduces the sort of "spectral leakage" problem and reduces the signal in the beginning and at the end of each frame to zero. In order to perform Fast Fourier transform, the voice signal end points are required. The energy of the frame is computed using the short-term log energy equation in order to detect the word boundary as shown in equation 1.

$$〖E〗_((i))=10 \log \sum_(t=u1)^(n1+N-1)▒S(t) \qquad (1)$$

Where…. E_((i))is energy, S(t) is voice sample.

The window is defined as w(n), 0≤N -1, where N is the frame length. The equation 2 shows the windowing is the signal Y (n).

$$Y(n)=x(n)w(n),0\leq n\leq N-1 \quad (2)$$

The hamming window can be represented using the equation 3 as follows:

$$w(n)=0.54-0.46\cos⁡〖(2\pi n/N),0\leq n\leq N〗 \quad (3)$$

2. Normalization: It is the process to make the power of the speech samples to unity. Since the extracted samples have different intensities due to the speaker loudness, speaker distance from the microphone and recording level. The normalization is done by dividing each sample by the square root of the sum of squares of all the samples in the segment. The samples of speech have diverse intensities because of distance of the user from microphone, loudness and recording level. By dividing and calculating the square root of the sum of the squares of all of all samples the normalization is performed. It can be as shown in equation 4.

$$s[n] = s[n]/\sqrt{\sum_{n=0}^{N-1} s\text{^}2 \mid n \mid} \quad (4)$$

Where s[n] is the speech sample, N is the number of samples.

3. Pre Emphasis: Pre-emphasis works with the aim of "Improving the signal to noise ratio by increasing the magnitude of higher frequency signals with respect to lower frequency signals". In speech processing, the original signal usually has too much lower frequency energy, and processing the signal to emphasize higher frequency energy is necessary. To perform pre-emphasis, value of α is selected between 0.9 and 1. Then each value in the signal is re-evaluated using equation 5:

$$Y[n] = X[n]- \ *X\ [n-1] \quad (5)$$

This is first order high pass filter.

Feature Extraction

The general procedure of sound/voice order includes extracting components from the sound information and nourishing them to classifier. The motivation behind this module is to convert signal waveform into a set of features or rather feature vectors (at an impressively lower data rate) for further examination. This is frequently considered as the signal processing front end. The features such as MFCC, pitch and formants are extracted.

1. MFCC Algorithm: In speech processing the Mel Frequency Cepstral Coefficient (MFCC) were first introduced and applied. Based on the choice of the number, shape, bandwidth and spacing of filters there differs in the MFCC variants. The differing qualities in the MFCC executions was additionally brought on by the progression made in psychoacoustics which progressively gave more refined models of the human sound-related observation. Human perception of frequency contents of sounds for speech signal does not follow a linear scale. Various estimates were built up in view of how the pitch discernment is identified with the human sound-related framework. Human view of frequency content of sounds for speech signal does not take a linear scale. Along these lines for every tone with a actual frequency f measured in Hz, a subjective pitch is measured on a scale called the "mel" scale. The Mel frequency scale is a linear frequency dividing underneath 1000Hz and a logarithmic separating over 1000Hz.

As a kind of perspective point, the pitch of a 1 KHz tone, 40dB over the perceptual listening to threshold, is characterized as 1000 Mels. The formula to calculate the number of Mels for a frequency f in Hz is given by equation 6.

$$Mel(f)=2595\times\log⁡10(1+f/700) \quad (6)$$

The MFCC feature extracted from fixed length signal frames effectively capture the characteristics of the speakers. It was also reported that the MFCC performs well for the task of speaker verification if the frame size ranging from 20 ms to 50 ms, and the frame step ranging from 1/6 to 1/3 of the frame size is used to analyze the speech. Thus keeping in view these recommendations, the MFCC based feature extraction method was implemented on short-time signal (frame by frame basis) using frames of length 20ms with 10 ms of overlap between adjacent frames.

2. Cepstrum: The cepstral representation of speech spectrum gives a representation that is good for the local spectral properties of the signal frame analysis. Because the Mel's are real numbers and can be converted into time domain using discrete cosine transform (DCT). This last step log Mel range is changed over back to time, and is called Mel Frequency Cepstrum Coefficients (MFCC).The discrete cosine transfer is done changing the Mel coefficients back to time domain. It can be represented using the equation 7.

$$C_n = \sum(\log S_k)\cos\left\{n(k-0.5)*\frac{n}{k}\right\},\ n = 1,2,3\ldots k \quad (7)$$

Where as S_k, K = 1, 2, … K are the outputs of last step.

3. Pitch-Pitch is considered to be the fundamental frequency of speech signal. An area of stress evaluation in the speech signal is major characteristics of pitch. The assessment of pitch frequency, its mean, variance, and distribution are considered. The pitch is generated due to the vibration of vocal cords and depend on vocal fold's tension and sub glottal air pressure during speech generation. There are various methods for the extraction of pitch. In this project work 'Autocorrelation based' pitch extraction has been implemented.

Autocorrelation is based on the center-clipping method. Initially the speech is low-passed filtered to 900 Hz.The first stage of processing is the computation of a clipping level. The autocorrelation function for the center clipped section is computed over a range of frequency from 60 to 320 Hz (the normal range of human pitch frequency) and is given by equation 8.

$$R(j) = \sum_{n=0}^{N-1-j} x(n)x(n-j) \quad (8)$$

4. Formants: Formants are characterized as the spectral peaks of sound range, of the voice, of a person. A formant frequency is the acoustic reverberation of the human vocal tract. They are frequently measured as amplitude peaks in the recurrence range of the sound wave. We have considered the initial 3 formants f1, f2, f3 for investigation of client voice. For diverse vowels, the scope of f1 lies between 270 to 730 Hz while the scope of f2 and f3 lie between 840 to 2290 and 1690 to 3010 Hz individually. Formant frequencies are all that much essential in the examination of the emotional condition of a person. The Linear prescient coding system (LPC) has been utilized for estimation of the formant frequencies. The analog signal is converted in .wav digital format. The signal is transformed to frequency domain using FFT and the power spectrum is further calculated. The frame of speech signal to be analyzed is given by N length sequence s(n). The sample of speech is hamming windowed to form the number of frames. The LPC is calculated out of the signal frames. The calculation of the LPC coefficients guarantees to obtain all members of bandwidth and formant frequencies.

5. Vector Quantization

The VQ was initially intended to be utilized as an information pressure strategy where an extensive arrangement of vectors in a multidimensional space could be replaced by a smaller set of representative vectors distribution matching the distribution of original data.

A typical VQ algorithm divides a large set of vectors into clusters having number of points. Each cluster is represented by its central point. A regular VQ calculation partitions an expansive arrangement of vectors into groups having number of points. Every bunch/cluster is given by the central point. As indicated by the Shannon's rate distortion hypothesis, the central point of each bunch /cluster is calculated as a centre of gravity and the bunch/cluster individuals should be ideally selected such that, for every bunch part, the bunch centroid is the closest centroid.

A VQ method envelops two basic assignments: An encoding procedure which includes a nearest-neighbor (NN) inquiry, allotting the closed codeword to a given vector. A codebook era process which finds an ideal, small arrangement of vectors (codebook) speaking to a given large arrangement of vectors. The components of codebook are known as the codewords.

In the proposed model, once the features are extracted from the user voice sample by using MFCC, Pitch and formant methods and the output of this module is fed as input to VQ module. The VQ is used for feature representation and it also compresses the set of feature.

Support Vector Machine (SVM) Classifier

Identification of speaker is the process of figuring out which speaker attributes from the speakers known framework best matches the unknown voice test. Speaker Identification requires different choice options and to execute speaker recognizable system utilizing SVM procedures requires multiclass SVM classifier. Speaker model is implemented Based on SVM system. Here, rather than bunching the speakers, a SVM classifier which isolates a group of speakers.

Hence, each hyper line built from the SVM separates the speakers until there is only one speaker in the sub group. There are two methods for SVM:

- One-isolates-all.
- One isolates- one.

In One-isolates-rest, the feature vectors of one speaker separates from the rest of other speakers. But it is found to b less efficient. In One-isolates-one, a total of S(S-1) /2 (S, the speaker population) binary SVM classifiers are formed, where each one is formed using data vectors from a pair of speakers. There exists 4 kernal functions like linear, polynomial, RBN and sigmoid.

IMPLEMENTATION RESULTS AND PERFORMANCE ANALYSIS

The work is carried out in order to achieve the liveness detection of users through their voice samples. Total of 64 MFCC, pitch and formant features are extracted for every user through VQ method. Hence, for 25 users total of 1,15,200(1800 samples × 64 features) feature values are used for training. Figure 2 shows the results displayed for text prompted speaker authentication using front end.

Further, 4 voice samples are used for testing the accuracy of the system. While testing, from each testing voice sample 64 features are extracted and compared with trained multi class SVM. If the user is imposter then multi class SVM classifier will reject the user which further results in rejecting the access to the system.

In-order to analyze the performance of the developed system, it is tested for 5,10,15,20 and 25 users separately. The accuracy of 92% is achieved for 5 users, 91.5% is achieved for 10 users, 90.3% is achieved for 15 users, 89.2% is achieved for 20 users and last of all 88.7% is achieved for 25 users.

Performance Analysis

The performance of biometrics system is measured using correct identification rate (CIR) or correct recognition rate (CRR), false acceptance rate (FAR) and false rejection rate (FRR). The CIR is the ratio of the number of authorized users accepted by the biometric system to the total number of identification attempts made. It is stated as follows in equation 9.

$$CIR = \frac{Number\ of\ correctly\ identifiable\ claims}{Total\ number\ of\ claims} * 100 \quad (9)$$

The FAR or 'type 2 error' is the ratio of the number of unauthorized users accepted by the biometric system to the total number of identification attempts made. It is stated as follows in equation 10.

$$FAR = \frac{Number\ of\ false\ acceptance}{Number\ of\ identification\ attempts} * 100 \quad (10)$$

The FRR or 'type 1 error' is the ratio of the number of authorized users rejected by the biometric system to the total number of attempts made. It is stated as follows in equation 11.

$$FRR = \frac{Number\ of\ false\ rejection}{Number\ of\ identification\ attempts} * 100 \quad (11)$$

The performance of the proposed system is tabulated in Table 1.

S. No.	Number of Users	Number of Tested Voice Samples	CIR in %	FAR in %	FRR in %
1	05	200	92	8	8
2	10	400	91.5	9	8.5
3	15	600	90.3	8	9.7
4	20	800	89.2	8.2	10.8
5	25	1000	88.7	8.7	11.3

The Figure 3 shows the recognition performance for a set of 5, 10, 15, 20 and 25 users.

Hence, by using these parameters the identification accuracy of around 92% to 88.7% is achieved for 5 to 25 users and false rejection rate of 8% to 11.3% is obtained for the 5 to 25 users.

CONCLUSION

User authentication using Text prompted Technique is proposed in this project work in-order to perform liveness detection. The MFCC, Pitch and Formant features are extracted and represented using vector quantization to form feature vector. The user will utter the prompted digit for authentication. The multi class support vector machine is used for the recognition. The identification accuracy of around 92% to 88.7% and false rejection rate of 8% to 11.3% is obtained for the 5 to 25 users.

Acknowledgment

I express my deep sense of gratitude to Dr Sunilkumar S Manvi, principle and HOD, Reva ITM for providing facilities and encouragement to carry out the project work.

I am highly indebted to my guide Prof. Geetha B and project coordinator Prof. Laxmi B Ranannavare for their motivational supportive encouragement, guidance and constant supervision as well as for proving necessary information regarding the project and also for their support in completing the project.

REFERENCES AND NOTES

1. Norbert Adam, "A Speech Analysis System Based on Vector Quantization Using the LBG Algorithm and Self-Organizing Maps", International Journal of Computer and Information Technology, Volume 3 – Issue 5, September 2014.
2. Shanthini Pandiaraj and K.R. ShankarKumar, "Speaker Identification Using DiscreteWavelet Transform",Journal of Computer Science Vol.11 No.1,pp 53-56, July 2014.
3. Rupali G. Shintri, S. K. Bhatia, "Review: GMM based Speaker Verification using MFCCFeature", International Journal of Electronics Communication and Computer Engineering, Volume 5, Issue 4, July 2014.
4. PPS Subhashini, TurimerlaPratap, "Text-Independent Speaker Recognition Using Combined LPC and MFC Coefficients", International Journal of Research in Engineering and Technology, June 2014.
5. Kavita Yadav, Moresh Mukhedkar, "MFCC Based Speaker Recognition using Matlab", International Journal of VLSI and Embedded Systems Volume 05, May 2014.
6. Apurva Adikane, Minal Moon, Pooja Dehankar, ShraddhaBorkar, SandipDesai,"Speaker Recognition Using MFCC and GMM with EM", International Journal of Engineering Research and Applications and International Conference On Industrial Automation and Computing , April 2014.
7. Niragi Shah, Omkar Tapkire, PriyankaRamakrishnan, Pratik Mutha,"Recognition Of Speaker Using Vector Quantization ByKekre's Fast Codebook Generation Algorithm", IRF International Conference, February 2014.
8. Om Prakash Prabhakar, Navneet Kumar Sahu "Performance Improvement of Human Voice Recognition System using Gaussian Mixture Model", International Journal of Advanced Research in Computer and Communication Engineering Volume 3, January 2014.

9. Sundeep Sivan, Gopakumar C ,"An MFCC Based Speaker Recognition using ANN with Improved Recognition Rate" , International Association of Scientific Innovation and Research,2014.

10. S.S. Wali , S.M.Hatture and S. Nandyal , "MFCC Based Text-Dependent Speaker Identification Using BPNN",International Journal of Signal Processing Systems Volume 3,2014.