I.J. Information Engineering and Electronic Business, **2014, 2, 35-40**
Published Online April 2014 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2014.02.04

Real Time Speaker Recognition System for Hindi Words

Geeta Nijhawan
Research Scholar,Faculty of Engineering and Technology, Manav Rachna International University, Faridabad, India
e-mail: geeta.fet@mriu.edu.in

Dr. M.K Soni
ED & Dean,Faculty of Engineering and Technology, Manav Rachna International University, Faridabad, India
ed.fet@mriu.edu.in

Abstract—Real time speaker recognition is needed for various voice controlled applications. Background noise influences the overall efficiency of speaker recognition system and is still considered a challenge in Speaker Recognition System (SRS). In this paper MFCC feature is used along with VQLBG algorithm for designing SRS. A new approach for designing a Voice Activity Detector (VAD) has been proposed which can discriminate between silence and voice activity and this can significantly improve the performance of SRS under noisy conditions. MFCC feature is extracted from the input speech and then vector quantization of the extracted MFCC features is done using VQLBG algorithm. Speaker identification is done by comparing the features of a newly recorded voice with the database under a specific threshold using Euclidean distance approach. The entire processing is done using MATLAB tool.The experimental result shows that the proposed method gives good results.

Index Terms—Hindi, Mel frequency cepstral coefficients**,** voice activity detector, MATLAB, Vector Quantization, LBG Algorithm.

I. Introduction

In India, the most widely spoken languages is Hindi.It is also spoken in countries like Fiji,Singapore, Mauritius, UAE, etc. Hindi speech recognition systems would be helpful in acquiring information from the masses. However, there is need for doing a lot in developing robust systems that can correctly recognize Hindi words[13],[20]. In this paper, we propose a speaker recognition system based on isolated digit word.

Speaker recognition is the process of recognizing the speaker from the database based on characteristics in the speech wave. Most of the speaker recognition systems contain two phases.

In the first phase feature extraction is done. The unique features from the voice signal are extracted which are used latter for identifying the speaker. The second phase is feature matching in which we compare the extracted voice features with the database of known speakers. The overall efficiency of the system depends on how efficiently the features of the voice are extracted and the procedures used to compare the real time voice sample features with the database.

A general block diagram of speaker recognition system is shown in Fig 1[1].

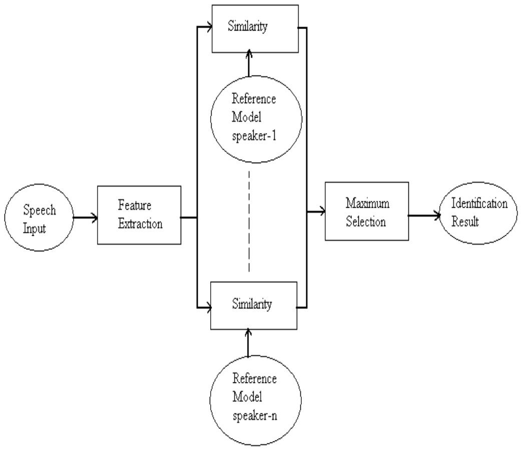


Fig 1: Speaker Recognition System

We can say,speaker recognition is a 1: N match where one unknown speaker’s extracted features are matched to all the templates in the reference model for finding the closest match. The speaker feature with maximum similarity is selected.

For more than fifty years, development on speaker recognition techniques has been an active area of research. Many methods [2] like simple template matching, dynamic time-warping approaches, and statistical pattern recognition approaches, such as neural networks and Hidden Markov Models (HMMs) [Siohan, 1998] have been used in the past. Gaussian Mixture Modeling (GMM) [Reynolds, 1995], multi-layer perceptrons [Altosaar and Meister, 1995], Radial Basis Functions [Finan et al., 1996] and genetic algorithms [Hannah et al., 1993] have also been used for speaker recognition. We have used VQ approach as it is easy to implement and gives accurate results [3].

The remainder of this paper is organized as follows. In section II, MFCC used for feature extraction from the input voice is presented in detail. Section III gives the details about the vector quantisation method used for feature matching.Section IV gives the experimental results and finally in section V conclusions are drawn.

II. Feature extraction

Speaker recognition is the process of automatically recognizing the identity of the speaker on the basis of individual information included in speech waves**.** Speaker recognition system consists of two important phases. The first phase is training phase in which a database is created which acts as a reference for the second phase of testing. Testing phase consists of recognizing a particular speaker. Speaker recognition systems contain three main modules [4]:

(1) Pre processing
(2) Features extraction
(3) Feature matching

These processes are explained in detail in subsequent sections.

2.1. ACOUSTIC PROCESSING

Acoustic processing is the process of converting analog signal from a speaker into digital signal for digital processing. Human speech frequency usually lies in between 300Hz-8000 kHz [5].Therefore according to Nyquist rule of sampling, twice the frequency of the original signal i.e. 16 kHz sampling rate is to be taken for recording [6]. The output of acoustic processing is discrete time voice signal which contains meaningful information.

2.2 FEATURE EXTRACTION

The purpose of Feature Extraction module is to give the acoustic feature vectors which are used to characterize the spectral properties of the time varying speech signal [14].These feature vectors are used for recognition of speaker.

2.3 Voice Activity Detector

Voice Activity Detector (VAD) [7] has been used to primarily distinguish speech signal from silence. VAD compares the extracted features from the input speech signal with some predefined threshold. Voice activity exists if the measured feature values exceed the threshold limit, otherwise silence is assumed to be present. Block diagram of the basic voice activity detector used in this work is shown in Fig. 2.

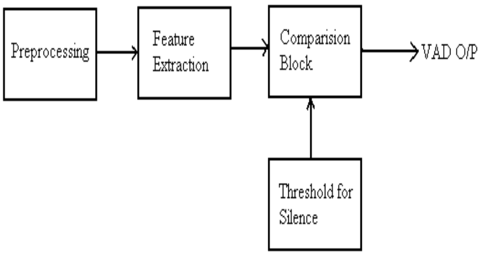


Fig 2: VAD block diagram

The effectiveness of the VAD block relies heavily on the preset values of the threshold for detection of voice activity. The VAD proposed here works well when the energy of the speech signal is higher than the background noise and the background noise is relatively stationary. The amplitude of the speech signal samples are compared with the threshold value which is being decided by analyzing the performance of the system under different noisy environments.

2.4 MFCC Extraction

Mel frequency cepstral coefficients (MFCC) is probably the best known and most widely used for both speech and speaker recognition [8],[9]. A mel is a unit of measure based on human ear's perceived frequency. The mel scale is approximately linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. The approximation of mel from frequency can be expressed as

$$\text{mel}(f) = 2595*\log(1+f/700) \tag{1}$$

where f denotes the real frequency and mel(f) denotes the perceived frequency. The block diagram showing the computation of MFCC is shown in Fig. 3.

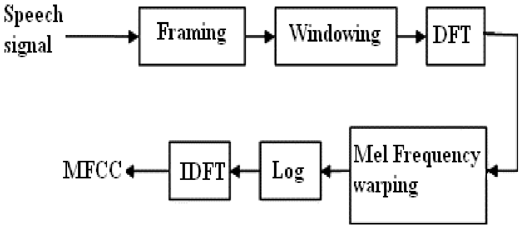


Fig 3: MFCC Extraction

In the first stage speech signal is divided into frames with the length of 20 to 40 ms and an overlap of 50% to 75%. In the second stage windowing of each frame with some window function is done. In time domain window is point wise multiplication of the framed signal and the window function. If we take the window as w (n), $0 \le n \le N\text{-}1$where N is the frame length, then the result of windowing is the signal

$$y(n)=x(n)w(n),\ 0 \le n \le N-1 \tag{2}$$

In our work hamming window is used to perform windowing function, which has the form:

$$w(n)=0.54-0.46\cos\frac{2\Pi n}{N-1} \tag{3}$$

A good window function should have a narrow main lobe and low side lobe levels in their transfer function. In the third stage DFT is done.It changes each frame from time domain to frequency domain [10]. Then mel frequency warping is done to transfer the real frequency scale to human perceived frequency scale called the mel-frequency scale. The new scale spaces linearly below 1000Hz and logarithmically above 1000Hz. The mel frequency warping is normally realized by triangular filter banks with the center frequency of the filter normally evenly spaced on the frequency axis. The warped axis is implemented according to Equation 1 so as to mimic the human ears perception. The output of the ith filter is given by-

$$y(i)=\sum_{j=1}^{N} s(j)\Omega_i(j) \qquad (4)$$

S(j) is the N-point magnitude spectrum (j =1:N) and Ω_i(j) is the sampled magnitude response of an M-channel filter bank (i =1:M). In the fifth stage,log of the filter bank output is computed and finally DCT (Discrete Cosine Transform) is computed. The MFCC may be calculated using the equation-

$$C_s(n,m)=\sum_{i=1}^{M}(\log Y(i))\cos[i\frac{2\pi}{N'}n] \qquad (5)$$

where N’ is the number of points used to compute standard DFT.

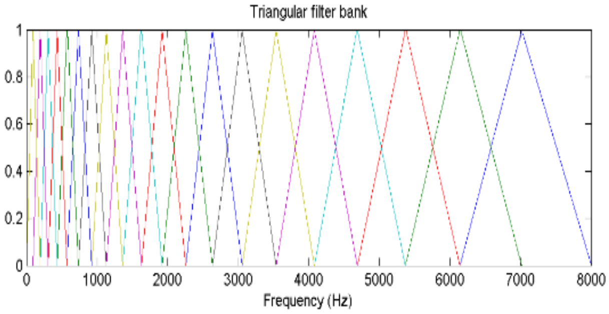
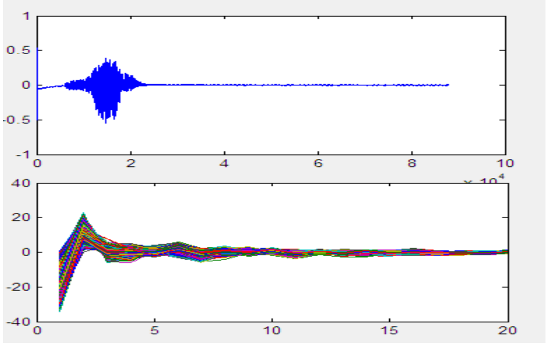


Fig 4: Triangular filter bank

Fig.5 shows screen shot of GUI developed using MATLAB of the input speech, MFCC, pitch and power plots.



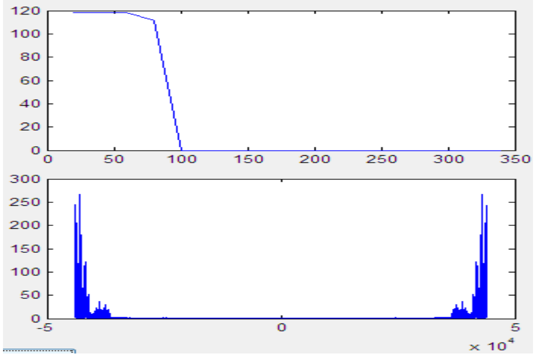


Fig 5: GUI waveforms showing input speech, MFCC, pitch and power plots.

III. Feature Matching

The aim of VQ is to compress data wherein we select the more effective features instead of using the whole feature vector. By clustering the speaker’s feature vector into a known cluster numbers the speaker models are formed. Each cluster is called centroid and is represented by a code vector .The code vectors constitute a codebook[11].

The recognition process is shown in Fig. 6 where we have shown only two speakers and two dimensions of the feature space [17], [18]. The circles are the feature vectors from the speaker 1 whereas the triangle refers to speaker 2. In the training phase, a VQ codebook is generated for each known speaker by clustering his training feature vectors. The resultant codewords are called centroids and are shown in Fig. 5 by black circles and black triangles for speaker 1 and 2, respectively. VQ-distortion is the distance of a vector from the closest code word of a codebook. During recognition process, an input speech of an unknown speaker is “vector-quantized” using each trained codebook and then the *total VQ distortion* is calculated [12]. A sequence of feature vectors {x_1, x_2,….,x_n}for unknown speaker are extracted. Each feature vector of the input is then compared with all the other codebooks. The codebook which gives the minimum distance is selected as the best. The formula used to calculate the Euclidean distance is given below:
Let us take two points A = (a_1, a_2…a_n) and B= (b_1, b_2 , ...b_n). The Euclidean distance between them is given by-

$$\sqrt{(a_1-b_1)^2+(a_2-b_2)^2+......(a_n-b_n)^2}=\sqrt{\sum_{i=1}^{n}(a_i-b_i)^2} \qquad (6)$$

The speaker corresponding to the VQ codebook with smallest total distortion is identified as the speaker of the input speech.

The steps involved can be summarized as:
Training Phase:

1. Calculate MFCC
2. Execute VQ
3. Find nearest neighbor
4. Compute centroids and create codebook

Testing Phase:

1.Calculate MFCC
2.Find nearest neighbor
3.Find minimum distance
4.Decision

Different speakers can be differentiated from one another on the basis of the location of centroids (Adapted from Song et al., 1987) [16].

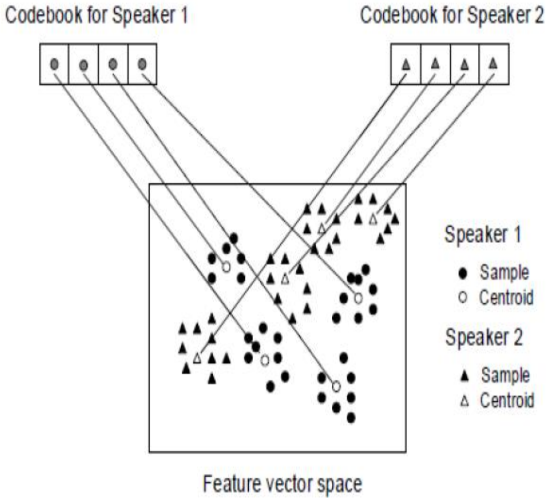


Fig 6: Vector quantization codebook

3.1 Clustering the Training Vectors

One of the most widely used vector quantization algorithms is Linde, Buzo and Gray (LBG) algorithm[1980]. A set of *L* training vectors are clustered into a set of *M* codebook vectors .Then *M*-vector codebook is designed in stages. Initially a 1-vector codebook is designed.The search for a 2-vector codebook is initialized by using splitting technique on the codewords.This splitting process continues until the required *M*-vector codebook is obtained.

The steps involved in implementing this procedure is given below [19]:

1. A 1-vector codebook is designed; this becomes the centroid of the complete set of training vectors (no iteration is required).

2. The size of the codebook is doubled by splitting codebook $\mathbf{y}_n$ according to the rule

$$y_n^{+} = y_n(1+ \in)$$
$$y_n^{-} = y_n(1- \in)$$

where *n* can take values from 1 to the current size of the codebook, and ε is c splitting parameter (we choose ε =0.01).

3. Nearest-Neighbor Search: for each training vector we find the codeword in the current codebook that is closest .This vector is allocated to the corresponding cell.

4. Centroid Update: The codeword in each cell are updated using the centroid of the training vectors allocated to that cell.

5. Iteration 1: steps 3 and 4 are repeated until we get the average distance below a preset threshold.

6. Iteration 2: steps 2, 3 and 4are repeated until a codebook size of *M* is designed.

Fig. 7 [14], [15] gives the steps of the LBG algorithm. “*Cluster vectors*” represents the nearest-neighbor search procedure. “*Find centroids*” is the centroid update procedure. “*Compute D (distortion)*” sums the distances of all training vectors in the nearest-neighbor search and tells whether the procedure has converged.

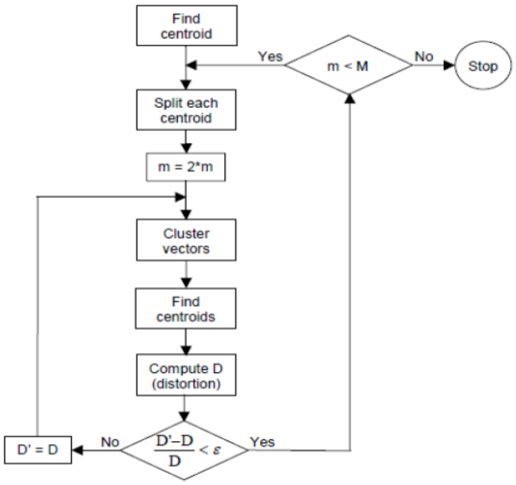


Fig 7: Flow diagram of the LBG algorithm (Adapted from Rabiner and Juang, 1993)

IV. Results

A database of five speakers is created .The feature extraction was done by using MFCC (Mel Frequency Cepstral Coefficients). The speakers were modeled using Vector Quantization (VQ). A VQ codebook is generated for each speaker and then stored in the speaker database by clustering the training feature vectors. In the proposed method, we have used the LBG algorithm for clustering purpose. In the recognition stage, VQ distortion based on Euclidean distance is calculated to match an unknown speaker with the speaker database. VQ based clustering approach gives faster speaker identification process. The entire coding is done in MATLAB. Fig.8 shows the spectrogram for hindi word ‘ek’.

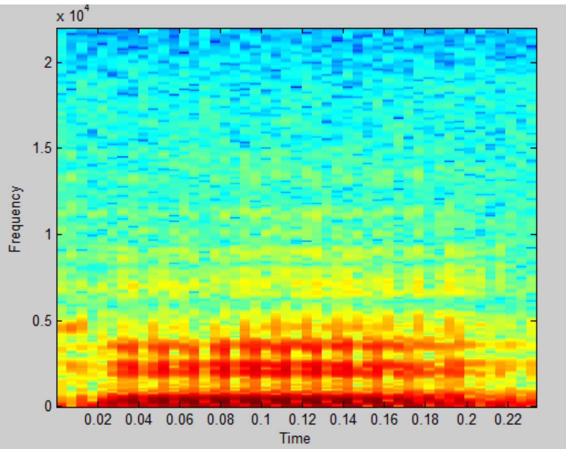


Fig 8: Spectrogram

The experimental results are shown in Table 1 from which it can be seen that the diagonal element has the minimum VQ distance value in their respective row. It indicates that S1 matches with S1, S2 matches with S2 and so on. The designed system identifies the speaker on the basis of the smallest Euclidean distance compared to the codebooks in the database.

Table 1: Experimental result for speaker recognition system

	S1	S2	S3	S4	S5
S1	5.5332	5.9783	5.9480	5.8662	9.4432
S2	5.8565	5.5785	6.7809	5.8752	8.9524
S3	6.3978	6.4021	5.1524	5.9761	9.8761
S4	6.5434	6.4367	6.2678	5.1567	9.8672
S5	5.2051	5.4356	4.9872	4.8675	4.8608

Table 2 shows the effect of changing the number of centroids on the identification rate of the system. It can be concluded that increasing the number of centroids definitely improves the identification rate but it comes at the expense of increasing computation time.

Table 2: System identification rate with number of centroids

Number of centroids	Identification rate (%)
8	75
16	78
64	80
256	80

V. Conclusions

The recognition rate obtained in this work using VAD, MFCC and LBG-VQ is good. Experiments show that by using VAD, we get a 5% error rate reduction compared to simply using noisy speech.

The VQ distortion between the resultant codebook and MFCCs of an unknown speaker is used for the speaker recognition. MFCCs are used because they mimic the human ear’s response to the sound signals. Experimental results presented shows that the % accuracy of recognition is around 80% and there is no false recognition which shows the robust performance of the proposed design approach. Analysis of the % accuracy for recognition with codebook size shows that the performance of the proposed system increases with increase in number of centroids. However VQ has certain limitations and its efficiency deteriorates when the database size is large and hence HMM techniques or Neural Network technique can be used to improve the performance and to increase the accuracy.

References

[1] Ch.Srinivasa Kumar, Dr. P. Mallikarjuna Rao, 2011, "Design of an Automatic Speaker Recognition System using MFCC, Vector Quantization and LBG Algorithm’’, International Journal on Computer Science and Engineering,Vol. 3 No. 8 ,pp:2942-2954.

[2] Amruta Anantrao Malode,Shashikant Sahare,2012 , "Advanced Speaker Recognition”, International Journal of Advances in Engineering & Technology ,Vol. 4, Issue 1, pp. 443-455.

[3] A.Srinivasan, "Speaker Identification and verification using Vector Quantization and Mel frequency Cepstral Coefficients”,Research Journal of Applied Sciences,Engineering and Technology 4(I):33-40,2012.

[4] Vibha Tiwari, "MFCC and its applications in speaker recognition”,International Journal on Emerging Technologies1(I):19-22(2010).

[5] Md. Rashidul Hasan,Mustafa Jamil,Md. Golam Rabbani Md Saifur Rahman, "Speaker Identification using Mel Frequency Cepstral coefficients”,3rd International Conference on Electrical & Computer Engineering,ICECE 2004,28-30 December 2004,Dhaka ,Bangladesh.

[6] Fu Zhonghua; Zhao Rongchun; “An overview of modeling technology of speaker recognition”, IEEE Proceedings of the International Conference on Neural Networks and Signal Processing Volume 2, Page(s):887 – 891, Dec. 2003.

[7] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-34, pp. 52-9, Feb. 1986.

[8] Sasaoki Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acoust., Speech, Signal Process., vol. 29(2), pp. 254-72, Apr. 1981.

[9] D.A. Reynolds, "Experimental evaluation of features for robust speaker identification," IEEE Trans. Speech Audio Process., vol. 2(4), pp. 639-43, Oct. 1994.

[10] Geeta Nijhawan, Dr. M.K Soni , “A New Design Approach for Speaker Recognition Using MFCC and VAD "I.J. Image, Graphics and Signal Processing, 2013, 9, 43-49.

[11] C.S. Gupta, "Significance of source features for speaker recognition," Master's thesis, Indian Institute of Technology Madras, Dept. of Computer Science and Engg., Chennai, India, 2003.

[12] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D.A. Reynolds, and B. Xiang, "Using prosodic and conversational features for high-performance speaker recognition," in Int. Conf. Acoust., Speech, Signal Process., vol. IV, Hong Kong, Apr. 2003, pp. 784-7.

[13] Nagaraja B.G., H.S. Jayanna” Kannada Language Parameters for Speaker Identification with The Constraint of Limited Data," I.J. Image, Graphics and Signal Processing, 2013, 9, 14-20.

[14] Y. Linde, A. Buzo, and R.M. Gray, "An algorithm for vector quantizer design," IEEE Trans. Communications, vol. COM-28(1), pp. 84-96, Jan. 1980.

[15] R. Gray, "Vector quantization," IEEE Acoust., Speech, Signal Process. Mag., vol. 1, pp. 4-29, Apr. 1984.

[16] F.K. Soong, A.E. Rosenberg, L.R. Rabiner, and B.H. Juang, "A Vector quantization approach to speaker recognition," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., vol. 10, Detroit, Michingon, Apr. 1985, pp. 387-90.

[17] T. Matsui, and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and Discrete/continuous HMMs," IEEE Trans. Speech Audio Process., vol. 2(3), pp. 456-9, July 1994.

[18] DSP Mini-Project: An Automatic Speaker ecognition System,http://www.ifp.uiuc.edu/~minhdo/teaching/speaker_recognition.

[19] Voice Recognition using DSP:http://azhar /paperpresentation.blogspot.in/2010/04/voice recognition-using-dsp.html.

[20] Vedant Dhandhania, Jens Kofod Hansen, Shefali Jayanth Kandi, and Arvind Ramesh,"A Robust Speaker Independent Speech Recognizer for Isolated Hindi Digits "International Journal of Computer and Communication Engineering, Vol. 1, No. 4, November 201.

Geeta Nijhawan did her M.Tech in Electronics and Communication Engineering in 2006 and B.E (Electronics) in 1995 from Government Engineering College, Raipur (Now NIT Raipur). She has a rich experience of 15 years in academics. She has authored three books on Electronics & Communication. Her core area of interest is Signal Processing. Presently she is doing her research work in the area of speech processing.

Dr. M. K Soni did his B.Sc (Engg.) in 1972 and M.Sc (Engg.) in 1975 from REC Kurukshetra (Now NIT Kurukshetra) and thereafter completed his Ph.D from REC Kurukshetra (in collaboration with IIT Delhi) in 1988. He has a total 39 years of rich experience into Academics. His area of interest is microprocessor based control systems and digital system design. He has more than 100 research papers in the International and National Journals to his credit. Presently he is Executive Director & Dean, Faculty of Engineering and Technology, Manav Rachna International University.

How to cite this paper: Geeta Nijhawan, M.K Soni,"Real Time Speaker Recognition System for Hindi Words", IJIEEB,vol.6, no.2, pp.35-40, 2014. DOI: 10.5815/ijieeb.2014.02.04