Speaker Recognition using MFCC and Improved Weighted Vector Quantization Algorithm

C. Sunitha[#1], E. Chandra[*2]

[#] Dept. of Computer Applications and Software Systems
Sri Krishna College of Arts and Science, Coimbatore, India
phdsunithascholar@gmail.com
*Department of Computer Science
Bharathiar University, Coimbatore, India
crcspeech@gmail.com

Abstract—Speaker recognition is one of the most essential tasks in the signal processing which identifies a person from characteristics of voices . In this paper we accomplish speaker recognition using Mel-frequency Cepstral Coefficient (MFCC) with Weighted Vector Quantization algorithm. By using MFCC, the feature extraction process is carried out. It is one of the nonlinear cepstral coefficient functions. Then the pattern matching is accomplished by evaluating the similarity of the unknown speaker and the trained models from the database. For this process, weighted vector quantization is proposed that takes into account the correlations between the known models in the database. Experimentations express that the new methodologies provide higher accuracy and it can observe the correct speaker even from shorter speech samples more reliably.

Keyword—Feature extraction, MFCC, Weighted VQ, Mel-filter bank.

I. INTRODUCTION

Speaker recognition is the process of realizing the speaker from the database based on characteristics in the speech signal [1], [11]. Generally speaker recognition can be classified into two processes which are speaker identification and speaker verification. The main difference between these two categories is that, the speaker verification performs a binary decision to verify the speaker's identity whereas speaker identification performs multiple decisions and it consist the process of comparing the voice of the person speaking to a database or reference templates in an attempt to identify the speaker. Speaker identification can be further divided into two subcategories; one is text dependent and another one is text independent speaker identification [2].Text-dependent speaker identification varies from text-independent because the identification is performed on a voiced instance of a particular word, but in the second type, the speaker can say anything. The speaker recognition is being used for various popular applications which include automated dictation and command interfaces. Speaker recognition used the biometric technology [3] and [12].

The objective of feature extraction is to convert the speech signal waveform into some of the parametric representation for further analysis and processing. The speech signal is a slowly time varying signal. More number of possibilities is available to permanently represent the speech signal for the speaker recognition process such as Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC) and so on. MFCC is possibly the best known and most popular technique which is used for speaker extraction process since LPC is not applicable for real-time applications. The use of cepstrum permits for the resemblance between two cepstral feature vectors to be computed. The cepstrum derived from the MFCC features rather than LPC features produces best performance such as FAR (False Acceptance Ratio) and FRR (False Rejection Ratio) for a speaker recognition system.

MFCCs are based on the fluctuations of the human ear's critical bandwidths with the frequency. Filters separated linearly at low frequencies and logarithmically at high frequencies have been used to get the phonetically important characteristics of the speech signal which is considered in the MFCC technique.

The proposed work is focused on designing techniques by effectively extracting the information related to speaker and it is used to improve the speaker recognition system. This system is split into two models as mentioned below:

- Speech signal features are extracted by the Mel-Frequency Cepstral Coefficients (MFCC) as feature vectors.
- The pattern matching of the extracted signals are carried out by using the weighted vector quantization technique.

The rest of the paper is organized as follows: Section 2 gives the literature survey of the speaker recognition. Section 3 describes the proposed method for speaker recognition and the experimental results and discussions are explained in Section 4. Conclusion of the proposed work and scope for future work are given in Section 5.

II. LITERATURE SURVEY

Shivanker Dev Dhingra et al (2013) have proposed an approach of isolated speech recognition by using Mel-Scale Frequency Cepstral Coefficients (MFCC) and Dynamic Time Warping (DTW). Normally several features can be extracted from the speech signal of spoken words. MFCC algorithm is used for feature extraction and DTW is applied to deal with different speaking speeds in speech recognition. DTW is an algorithm which is used for measuring the similarity between two sequences which may vary in time or speed. The extracted features were stored in a .mat file by using MFCC algorithm. A distortion measure based on minimizing the Euclidean distance was applied when matching the unknown speech signal with the speech signal database. The experimental results were examined with the help of MATLAB and it is proved that the results are efficient. This process can be extended for n number of speakers [4].

Xinhui Zhou et at (2010) have compared the Linear Frequency Cepstral Coefficients and Mel Frequency Cepstral Coefficients for Speaker Recognition. This is prompted by insight from speech production that some speaker characteristics related with the structure of the vocal tract, in particular the vocal tract length, are speculated more in the high frequency region of speech. The performances between MFCC and LFCC in the NIST SRE 2010 extended core task. The produced results explained that, while they are complementary to each other, LFCC systematically outperforms MFCC mainly due to its better performance in the female trials by better capturing the spectral characteristics in the high frequency region [5]. The performance of LFCC is as robust as MFCC for the babble noise, but it is not similar while dealing with the white noise.

Geeta Nijhawanand Dr. M.K Soni (2014) have used MFCC and vector quantization for speaker recognition. The produced speech recognition rate is good by using the Voice Activity Detector (VAD), MFCC and LBG vector quantization algorithm. On an average, VAD approach produced 5% error rate reduction when compared to simply using a speech free segment from the beginning of the utterance for noise modeling. In this paper, the vector quantization distortion between the resultant codebook and MFCCs of an unknown speaker is used for the speaker recognition. MFCC is used for feature extraction since it mimics the human ear's response to the sound signals. The experimental results presented that the recognition percentage is about 95% and there is no false recognition rate [6]. The performance of the MFCC and VQ has been increased with increase in number of centroids. But these techniques have some limitations such as the decrement in efficiency when the database size is large.

Akanksha Singh Thakur and Namrata Sahayam (2013) have proposed the concept of MFCC and vector quantization for speaker recognition using Euclidean distance. They have explained the complexity of voice signal due to too much information contained in the signal. Hence the digital signal processes such as feature extraction and feature matching have been introduced to demonstrate the voice signal. The authors have been implemented an approach of speech verification by using a Euclidean distance. The speech recognition with MFCC has been enforced using software platform MATLAB R2010b [7].

Tomi Kinnunen and Pasi Fränti have proposed Speaker discriminative weighting method for VQ-based Speaker identification. They implemented and evaluated a weighted matching method for text independent speaker recognition. This method produced tremendous improvement over the previous method. The proposed method can also detect the correct speaker from much smaller speech samples. Furthermore, this method can be popularized to any other pattern recognition tasks because it is not designed for any particular features or distance metric [8].

III. PROPOSED METHODOLOGY

This section describes about the proposed work behind the speaker recognition. Block diagram for the proposed work is illustrated in Fig 1.

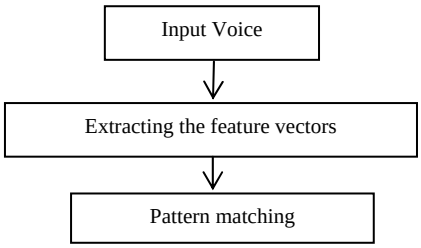

Fig.1. Speech recognition system

Initially input voice signal is received for the recognizing process. Then the features are extracted from the input signal and the extracted vectors are finally applied to the pattern matching process.

A. Mel-Frequency Cepstrum Coefficients

This section explains the extraction of the original signal into number of feature vectors for dimension reduction and probabilistic modeling. In the speech recognition, more methods are available to extract the features such as Mel frequency Cepstral coefficient (MFCC), linear prediction coefficients (LPC) and perceptual linear prediction coefficients (PLP) and so on. In the proposed methodology, features are extracted through MFCC and this method is one of the popular methods for extraction of speech signal. Sounds are represented in two ways, such as linear Cepstral and nonlinear Cepstral. MFCC is derived from nonlinear Cepstral representation of sound. Mel scale is used in the MFCC, and it is more responsible for human auditory system than linear Cepstral representation of sound [9]. Block diagram for MFCC is given in Fig. 2.

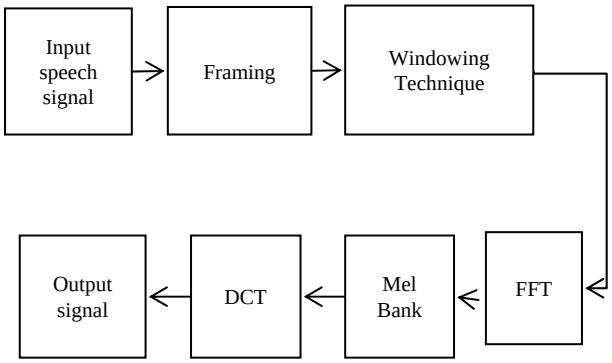

Fig.2. Block diagram for MFCC technique

In this process, initially the original signal from time domain to frequency domain is transformed by using Discrete Fourier Transform (DFT). The power spectrum is used for this conversion. Before DFT, hamming window is used for the reduction of frequency distortion due to segmentation. After this process, filter bank is used for wrapping the frequency from hertz scale to Mel scale. In conclusion Discrete Cosine Transformation (DCT) is used for the extraction of feature vectors on the logarithm of Mel scale power spectrum.

Step 1: In the first stage, original signal is multiplied by using Hamming window, and then the window speech frames are processed under DFT. This is obtained from Fourier transform.

$$x(k) = \sum_{n=0}^{N_D-1} x(n) e^{-j2\pi nk/N_D} \tag{1}$$

In the above equation, N_D defines the number of points in the DFT.

Step 2: Filter bank is created

$$e_s(i) = In\left[\sum_{k=0}^{N_f-1} |x(k)|^2 T_i(k)\right] \tag{2}$$

The above equation defines the energy spectrum $e_s(i)$, where number of filter is indicated by N_f and $i = 1,2,..,N_f$.

$$T_i(k) = \begin{cases} 0 & \text{for} \quad k < k_{b_{i-1}} \\ \frac{k - k_{b_{i-1}}}{k_{b_i} - k_{b_{i-1}}} & \text{for} \quad k_{b_{i-1}} \le k \le k_{b_i} \\ \frac{k_{b_{i+1}} - k}{k_{b_{i-1}} - k_{b_i}} & \text{for} \quad k_{b_i} \le k \le k_{b_{i+1}} \\ 0 & \text{for} \quad k > k_{b_{i+1}} \end{cases} \tag{3}$$

The above equation demonstrates the band pass filter $x(k)$ by triangular filter bank $T_i(k)$. Filter boundary points are indicated by $\{k_{b_i}\}_{i=0}^{N_f+1}$, where k denotes the index of the N_D point DFT.

Step 3: Mel-scale calculation using O'Shaughnessy [10] it is given by below equation

$$f_{mel} = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right) \tag{4}$$

In the above equation, f_{mel} denotes the sampling frequency.

$$k_{b_i} = \left(\frac{N_D}{F_s}\right) f_{me1}^{-1}\left[\begin{matrix} f_{mel}(f_{min}) \\ + \frac{i\{f_{mel}(f_{max}) - f_{mel}(f_{min})\}}{N_f+1} \end{matrix}\right] \tag{5}$$

In the above equation f_{min} and f_{max} denotes the low and high frequency boundary of the filter banks. Inverse transform f_{me1}^{-1} is given by below equation

$$f_{me1}^{-1}(f_{mel}) = 700\left[10^{f_{mel}/2595} - 1\right] \quad (6)$$

Step 4: MFCC coefficient is calculated, that the output of logarithmic filter bank is given to the DCT.

$$MFCC(n) = \sum_{i=0}^{N_f-1} e_s \cos\left(\frac{(\pi n(i-0.5))}{N_f}\right) 0 \le n \le N_f - 1 \quad (7)$$

Where n represents the number of MFCC coefficients.

B. *Improved Weighted Vector Quantization Technique*

Vector quantization is a function of framing the vectors from a large vector space to a finite number of regions in that space. Each region is called as cluster and can be represented by its center called as centroid [13,14]. The collection of all codewords is called a codebook [15 - 18].The codebook contains the resulting set of code vectors and it is stored in the speaker database. In this, each vector symbolizes a single acoustic unit typical for the particular speaker. Thus, a smaller set of sample vectors with similar distribution establishes the distribution of the feature vectors. The codebook should be reasonably high since the matching performance improves with the size of the codebook [19].

The matching of an unknown speaker is then accomplished by evaluating the resemblance between the feature vectors of the unknown speaker to the models of the known speakers in the database.

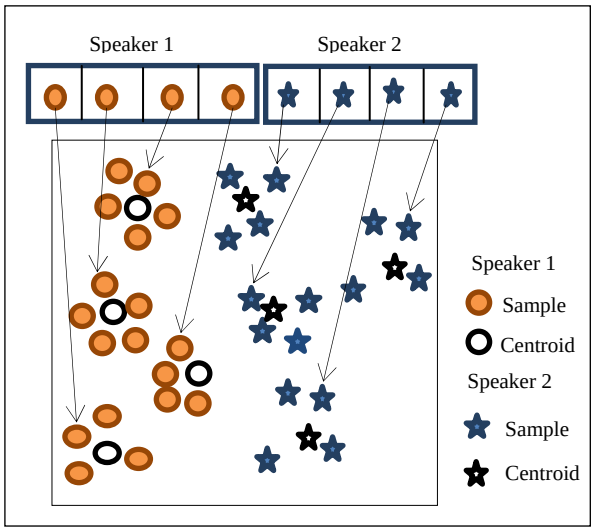

Fig.3. Speaker recognition process

The sequence of feature vectors extracted from the unknown speakers is denoted as follows.

$$X = \{x_1, x_2, \ldots, x_N\} \quad (8)$$

The aim of the speaker recognition is to find the best matching codebook C_{best} from the database of *N* codebooks. In the proposed system, codebook is generated using the LBG algorithm.

$$C = \{C_1, C_2, \ldots, C_T\} \quad (9)$$

Normally the matching is carried out by a dissimilarity measure that calculates the average distance of the matching as mentioned below.

$$d : X \times C \rightarrow \boldsymbol{R} \quad (10)$$

The best pattern matching codebook can then be determined by the codebook that minimizes the dissimilarity measure. Here similarity measure is used. In this way, we can define the weighting matching method more clearly. Thus, the best matching codebook is now determined as the codebook that maximizes the rate of the similarity measure of the mapping.

$$s : X \times C \rightarrow \boldsymbol{R} \quad (11)$$

This can also be written as,

$$C_{best} = arg_{1 \le i \le p} \max\{s(X, C_i)\} \quad (12)$$

Here the similarity measure is determined as the average of the inverse distance values:

$$S(X < C_i) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{d(x_t, c_{min}^{i,t})} \quad (13)$$

Where,

$c_{min}^{i,t}$- The nearest code vector to x_t in the codebook C_i

d – Given distance function in the feature space.

C. LBG Algorithm

Linde-Buzo-Gray (LBG) algorithm is a vector quantization algorithm which is similar to K-means clustering algorithm. It is used to derive a good codebook. It takes a set of input vectors as input and generates typical subset of vectors with the user specified K which is less than n as output according to the similarity measure. The steps of the LBG algorithm are given below.

a) Input training vectors

$$S = \{a_i \in R^d | i = 1,2,3, \dots, n\}$$

b) Create a codebook

$$C = \{b_j \in R^d | j = 1,2,3, \dots, K\}$$

c) Set $D_0 = 0$ and let $k = 0$.

d) Categorize the n training vectors into K clusters according to$a_i \in S_t \; if \; \|a_i - c_q\|_p \leq \|a_i - b_j\|_p$ for $j \neq q$.

e) Update cluster centers b_j, j = 1,2,3,…,K by $b_j = \frac{1}{|S_j|}\sum_{a_i \in S_j} a_i$.

f) Determine $k \leftarrow k + 1$ and calculate the distortion $\sum_{j=1}^{K}\sum_{a_i \in S_j} \|a_i - b_j\|_p$.

g) If $D_{k-1} - D_k / D_k > \in (a\ small\ number)$, repeat d~f.

h) Output the codebook

$$C = \{b_j \in R^d | j = 1,2,3, \dots, K\}$$

Best Matching Codebook is now determined as the Codebook that maximizes the rate of the similarity measure.

IV. RESULTS AND DISCUSSION

NTT database [20] and a large-scale Japanese Newspaper Article Sentences (JNAS) database [21] were used to evaluate proposed method. The proposed work is compared with the existing technique which is classified by DT-DWT with RVM and some other existing methods through performance metrics such as speaker identification accuracy.

Comparison is made for the algorithms which include MFCC with GMM, Inverted MFCC, Kullback-Leibler divergence, DT-CWT with RVM, MFCC with ELM and the proposed MFCC with Weighted VQ algorithm.

TABLE I. SPEAKER IDENTIFICATION RATE

Technique	Speaker identification rate (%)
MFCC with GMM	77.36
Inverted MFCC	77
Kullback-Leibler divergence	93
DT-CWT with RVM	95
MFCC with ELM	98.4
Proposed MFCC with Weighted VQ	98.75

The resultant windows obtained from the MFCC combined with the weighted vector quantization is shown in the following diagrams.

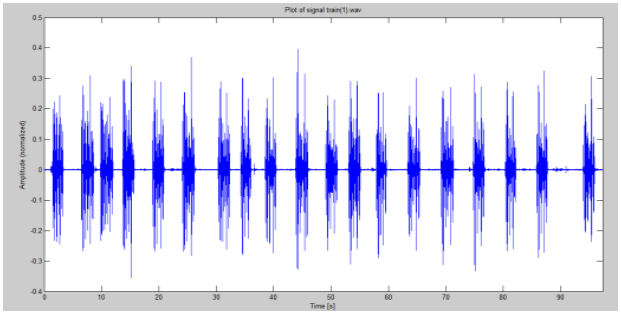

Fig.3. Input Signal

The input voice is shown in the Fig. 4 as the signal in terms of normalized amplitude and time. It implicates the variations in the voice signal in terms of amplitude variation.

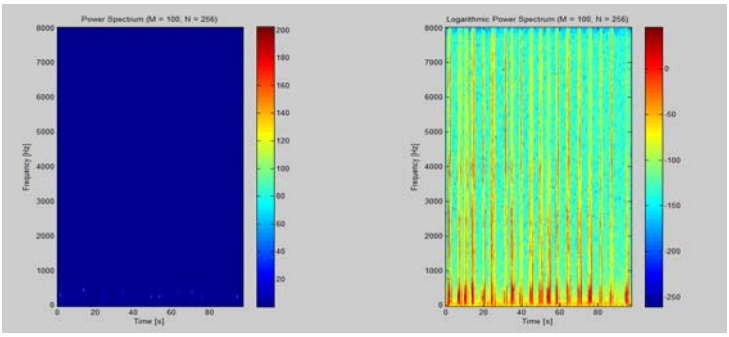

Fig.4. The linear and the logarithmic spectrum plot

The input signal is recognized and produced the linear and logarithmic spectrum plot of the given input signal in terms of frequency and time which is shown in the Fig. 5.

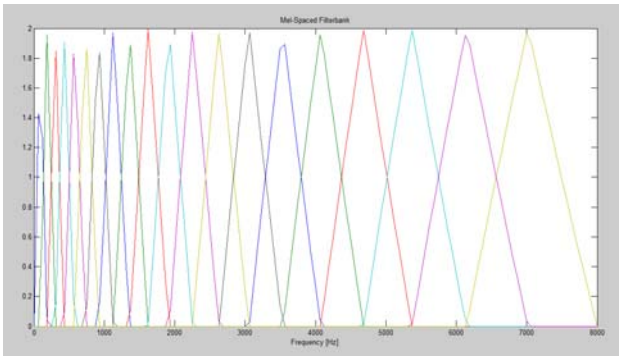

Fig.5. Mel-Spaced Filter Bank

The output of the MFCC technique is shown in Fig. 6 as the Mel-spaced filter bank in terms of frequency.

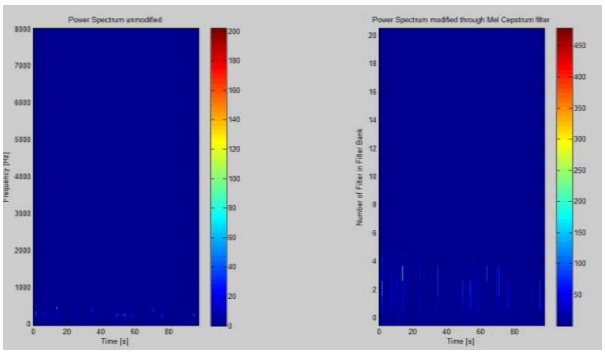

Fig.6. Unmodified power spectrum and the modified power spectrum through Mel Cepstrum filter

The unmodified power spectrum of the input signal is shown in the Fig. 7 in terms of frequency and time. The modified power spectrum through the Mel Cepstrum filter is given in Fig. 8 by considering the number of filters in the filter bank and time.

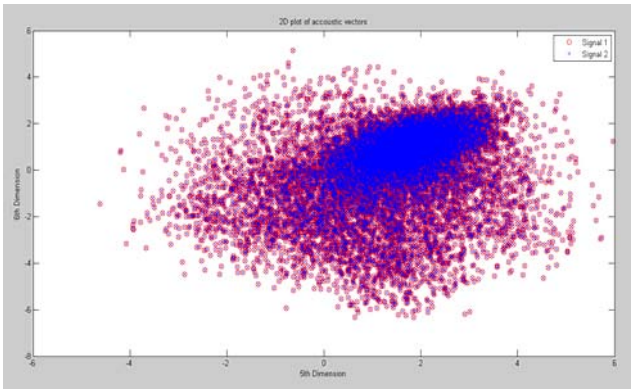

Fig.7. 2D plot of Acoustic vectors

The 2D plot of acoustic vectors is shown for the fifth and sixth dimension.

V. CONCLUSION

This paper provides the speaker verification using weighted vector quantization. In this paper speech signals are extracted by using MFCC technique where features are extracted using linearly spaced filters in Mel scale. Compared to other existing techniques, this method of MFCC provides the better feature extraction. The weighted vector quantization is suitable for 2D acoustic signal, leading to high material recognition accuracy than that of other system. The proposed method gives the better result which is observed from the experimental result.

REFERENCES

[1] Zilovic, M.S. Ramachandran, R.P. and Mammone, R.J "Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions". IEEE Transactions on Speech and Audio Processing, Volume 6, May 1998.
[2] 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, Dec. 2003.
[3] Claudio Becchetti and Lucio Prina Ricotti, "Speech Recognition", Chichester: John Wiley & Sons, 2004.
[4] Shivanker Dev Dhingra, Geeta Nijhawan and Poonam Pandit "Isolated speech recognition Using MFCC and DTW", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 2, Issue 8, August 2013.
[5] Xinhui Zhou, Daniel Garcia-Romero, Ramani Duraiswami, Carol Espy-Wilson and Shihab Shamma, "Linear versus Mel Frequency Cepstral Coefficients for Speaker Recognition", 2010.
[6] Geeta Nijhawan and Dr. M.K Soni, "Speaker Recognition Using MFCC and Vector Quantisation", International Journal on Recent Trends in Engineering and Technology, Vol. 11, No. 1, July 2014.
[7] Akanksha Singh Thakur1, Namrata Sahayam, "Speech Recognition Using Euclidean Distance", International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 3, March, 2013.
[8] Tomi Kinnunen and Pasi Fränti, "Speaker Discriminative Weighting Method for VQ-based Speaker identification", 2011.
[9] M. Bahoura, Pattern recognition methods applied to respiratory sounds classification into normal and wheeze classes, Comput. Biol. Med., 2009.
[10] O.C. Ai, M. Hariharan, S. Yaacob, L.S. Chee, Classification of speech dysfluencies with MFCC and LPCC features, Expert Syst. Appl., 2012.
[11] H. Beigi, Fundamentals of Speaker Recognition, Springer, New York, 2011.
[12] A.K. Jain, A. Ross, S. Prabhakar, An introduction to biometric recognition, IEEE Trans. Circuits Syst. Video Technol., 2004.
[13] Md. RashidulHasan, Mustafa Jamil, Md. GolamRabbani Md. SaifurRahman, "Speaker identification using mel frequency cepstral coefficients" ICECE 2004, 28-30 December 2004.
[14] SheerazMemon, Margaret Lech and Ling He "Using Information theoretic vector quantization for inverted MFCC based speaker verification" IEEE CCECE/CCGEI, Saskatoon, May 2005.
[15] He J., Liu L. and Palm G.,"A discriminative training algorithm for VQbased speaker identification", IEEE Transactions on Speech andAudio Processing,7(3): 353-356, 1999.
[16] Kinnunen T., Kilpeläinen T. and Fränti P. "Comparison of clustering algorithms in speaker identification", Proc. IASTED Int. Conf. Signal Processing and Communications, 2000.
[17] Kyung Y.J., Lee H.S.: "Bootstrap and aggregating VQ classifier for speaker recognition". Electronics Letters, 1999.
[18] Pham T., Wagner M., "Information based speaker identification", Proc. Int. Conf. Pattern Recognition (ICPR), 2000.
[19] Pham T., Wagner M., "Information based speaker identification", Proc. Int. Conf. Pattern Recognition (ICPR), 2000.
[20] T. Matusi and S. Furui, "Concatenated phoneme models for text-variable speaker recognition," in Proc. ICASSP'93, 1993, vol. II, pp. 391–394.
[21] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," J. Acoust. Soc. Jpn. (E), vol. 20, no. 3, pp. 199–206, 1999.

AUTHOR PROFILE

Mrs. C. Sunitha, MCA, M.Phil., is working as the Head, Dept. of Computer Applications and Software Systems at Sri Krishna College of Arts and Science, Coimbatore. She has more than 15 years of teaching and 10 years of administrative experience. She is guiding many M.Phil. Research scholars and has published papers in International journals. She has presented papers and organized various Seminars and Forums. She is pursuing her Ph.D. at Bharathiar University, Coimbatore and doing research work on Speech Recognition. She is a Life Member of Computer Society of India (CSI), Coimbatore Chapter and The Indian Science Congress Association (ISCA).

Dr.E. Chandra obtained her Ph.D. degree in the area of Speech recognition system from Alagappa University Karaikudi in 2007. She has totally 20 years of Rich experience in teaching including 6 months in the industry. At present she is working as Professor, Department of Computer Science, Bharathiar University, Coimbatore. She has published more than 40 research papers in National, International reputed journals and conferences in India and abroad. She is a Reviewer of international journals. She has guided more than 20 M.Phil. research scholars and 4 Ph.D. scholars. At present 8 Ph.D. scholars are working under her guidance. She wrote a book on “Fractal Image Compression Techniques: Genetic Algorithm and Efficient Domain Pool Algorithm for Fractal Image Compression” and “Distributed Data Mining”. She has completed a Minor Research Project on “Diverse Sub-band Adaptive Speech Enhancement for Hearing Aids” which is sanctioned by UGC. She has received best actively participating woman in CSI (India) for Coimbatore Chapter during the year 2012. She has delivered lectures to various Colleges in Tamil Nadu & Kerala. She is a Board of studies member at various colleges. Her research interest lies in the area of Neural Networks, Speech Recognition Systems, Fuzzy Logic and Machine Learning Techniques. She is a Life member of CSI, Society of Statistics and Computer Applications.