Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 89 (2016) 804 – 811

Procedia Computer Science

Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016)

Robust Speaker Identification Incorporating High Frequency Features

Latha*

Amrita Vishwa Vidyapeetham, Amrita University, India

Abstract

Speaker identification system identifies the person by his/her speech sample. Speaker Identification (SI) system should posses a robust feature extraction unit and a good classifier. Mel frequency cepstral coefficient (MFCC) is very old feature extraction scheme, which has been regarded as standard set of feature vectors for speaker identification. The mel filter bank used in MFCC method, captures the speaker information more effectively in lower frequencies than higher frequencies. Hence high frequency region characteristics are lost. This problem is solved in the proposed method. The speech signal comprises both voiced and unvoiced segments. The voiced segment includes high energy, low frequency components and unvoiced segment includes low energy, high frequency components. In proposed method, the speech sample is divided into voiced and unvoiced segments. The voiced speech segment is filtered using mel filter bank to generate MFCC from lower frequencies of speech signal and unvoiced speech segment is filtered using inverted mel filter bank to generate IMFCC from higher frequencies of speech signal.

Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016

Keywords: IMFCC; Inverted Mel Filter Bank; MFCC; Short Time Energy; Voiced and Unvoiced Speech.

1. Introduction

Speaker identification (SI) is the important field of research from past many years. Speaker identification system identifies the person by his/her speech sample. There are two research topics in this field. They are feature extraction and feature matching. A basic speaker identification system is shown in Fig. 1. This system has two different stages. Registration or training is the first stage and the second stage is testing stage. In the training stage, each speaker wish to register has to provide samples of his/her speech to prepare a reference model of all registered speakers. In testing stage, the input speech signal of the speaker claiming the identity, is used to extract the feature and is matched with stored features to get the identification result.

The speaker identification is closed set[1], when it is known that all speakers of interest are included in the speaker model, and is open set when some unknown speaker is not the part of the speaker model. Most of the speaker identification systems are open set. The speaker identification is text dependent[2] if there is constraint on the utterance of the speaker, and it is text independent[3] if there is no restriction on the spoken word.

*Corresponding author. Tel.: +91 -9449107045.
E-mail address: s_latha@blr.amrita.edu

Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016
doi:10.1016/j.procs.2016.06.064

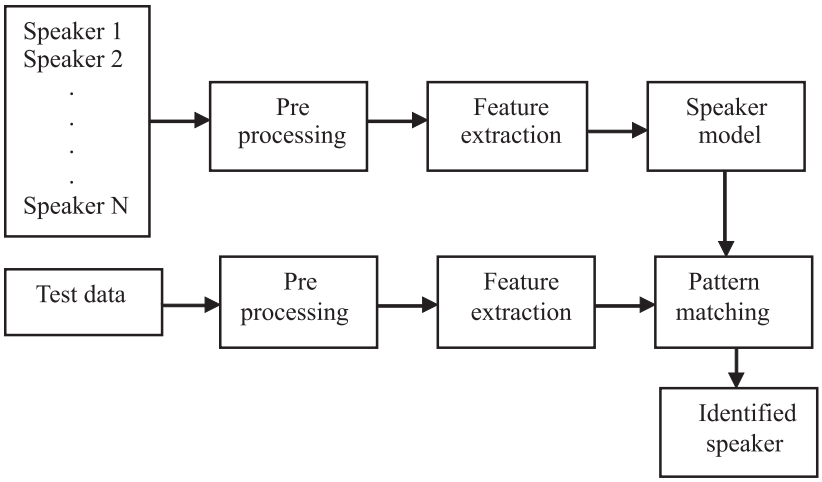

Fig. 1. Speaker Identification System.

1.1 Literature survey

The various feature extraction methods used by many authors include very old basic methods based on spectral averages[4], Pitch, Formats, Linear Predictive Coefficients (LPC)[5], Linear Predictive Cepstral Coefficients (LPCC)[6], Real Cepstral Coefficient (RCC)[6], Mel-Frequency Cepstral Coefficients (MFCC)[6], Perceptual Linear Predictive Coding (PLPC) and Linear Frequency Cepstral Coefficients (LFCC) etc. D.A.Reynolds made a detailed comparison of the feature extraction methods such as PLPC, MFCC, LPCC and LFCC[7] LFCC performance is very poor as it gives equal importance for all frequencies which resulted in increased redundancy. LPCC and PLPC performed better with increased filter orders where as MFCC performance is better with lower order filters also. Based on the literature, it is observed that MFCC outperformed all other methods.

Most frequently used feature matching techniques are Hidden Markov Model (HMM)[2], Gaussian Mixture Model (GMM)[3], Vector Quantization (VQ)[8] and Dynamic Time warping (DTW)[9]. HMM and DTW are the feature matching techniques mostly used for text dependent speaker identification. Different clustering techniques are compared[10] and is observed that the identification accuracy is improved with increased code book size.

1.2 Speaker specific features

In speaker identification the role of parametric representation of the speech signal which is effective in representing speaker specific characteristic is very important in the whole process of identifying the speaker. The identification accuracy is mainly influenced by speaker specific features. If features are more appropriate then the accuracy is high. But selection and extraction of speech features is not an easy process. For an SI system, speech features should occur frequently and naturally in speech signal, should be easy to extract and measure. The robust features are not affected by physical health conditions of the speaker and ambient noise. As we have seen from the literature, frequently used parameters and parametric representations of the speech signal are Pitch, Formats, Linear Predictive Coefficients (LPC), Mel-Frequency Cepstral Coefficients (MFCC), Real Cepstral Coefficient (RCC) and Linear Predictive Cepstral Coefficients (LPCC) etc.

The pitch (reciprocal of pitch period), represents periodicity of relaxation oscillations of the vocal folds of the speaker which intern depends on size and thickness of the vocal folds. Therefore combination of pitch along with other parametric speech features can make a set of robust feature of a speaker.

In Linear predictive coding (LPC) the vocal tract is modeled as an all pole filter (LTI system). The linear prediction method is most accurate method for estimating the parameters which characterize the vocal tract LTI system[11].

Since the vocal tract is unique in size and shape for a speaker, features extracted from LPC can make a set of speaker specific features.

The MFCC method of speech feature extraction is posed by Davis and Mermelstein[12]. MFC coefficients show the energy distribution of the signal in spectral domain Here the idea is to use mel scale which depicts characteristics of human auditory system. There are some features of speaker, which are present at higher frequencies of the speech spectrum and these features are given least importance in MFCC. The speaker information present in higher frequencies can be effectively extracted by the new feature vector IMFCC[13]. Here the idea is to invert the mel filter bank. Inverted Mel Frequency Cepstral Coefficients (IMFCC) are generated by using same processing steps used for MFCC but mel filter bank structure is replaced by inverted mel filter bank structure, which is complement to human hearing characteristics.

2. System Description

2.1 Mel frequency cepstral coefficients (MFCC)

Mel-frequency cepstrum (MFC) represents the power spectrum of the speech signal in cepstral domain. MFCC is based on human auditory system which can perceive frequencies only up to 1 kHz. MFCC applies Mel filter bank having a set of triangular band pass filters. These band pass filters are spaced linearly at lower frequencies, ie below 1 kHz and use logarithmic spacing above 1 kHz. Hence Mel filter bank covers low frequency regions more closely than high frequency regions. The overall process of the MFCC generation and frequency response of the Mel filter bank are shown in Fig. 2 and 3 respectively.

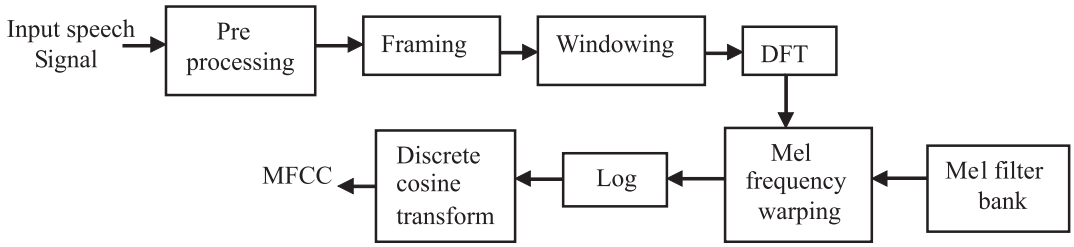

Fig. 2. Extraction of Mel Frequency Cepstral Coefficients.

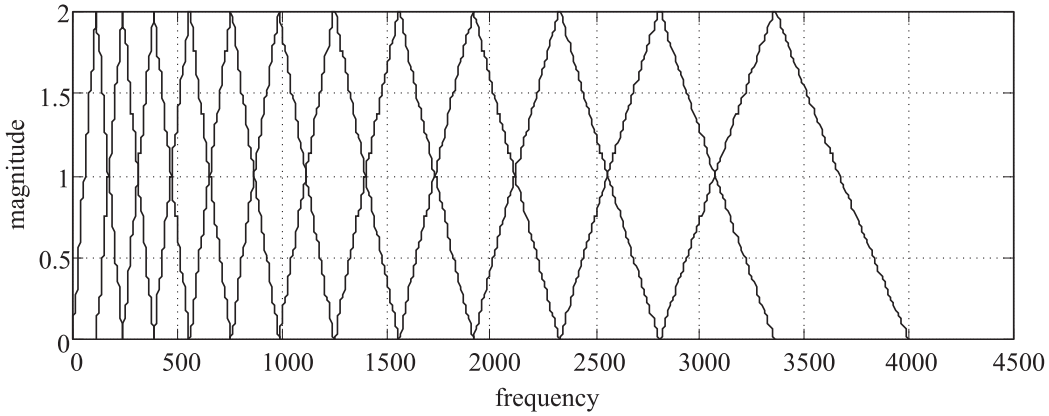

Fig. 3. Mel Frequency Scale Filter Bank.

In MFCC feature generation the mathematical relationship of Mel-scale frequency with linear frequency is given by the equation[12]

$$f_{\text{mel}} = 2595 \log\left(1 + \frac{f}{700}\right) \tag{1}$$

The linear scale frequency can be obtained back from mel scale frequency using

$$f = 700\left[\left(10^{\frac{f_{\text{mel}}}{2595}}\right) - 1\right] \tag{2}$$

2.2 Inverted mel frequency cepstral coefficients (IMFCC)

IMFCC has increased resolution in high frequency range compared to low frequency range. IMFCC[13] applies inverted mel filter bank having a set of triangular band pass filters, which are spaced logarithmically at low frequencies and linearly spaced at higher frequencies. Hence Inverted mel filter bank covers high frequency regions more closely than low frequency regions. The overall process of the IMFCC[13] and frequency response of the inverted mel filter bank are shown in Fig. 4 and 5 respectively.

The mathematical relationship between Inverted mel scale frequency and linear frequency is given by the equation

$$f_{\text{invertedmel}} = 2146.1 - 2595 \log\left(1 + \frac{(4000 - f)}{700}\right) \tag{3}$$

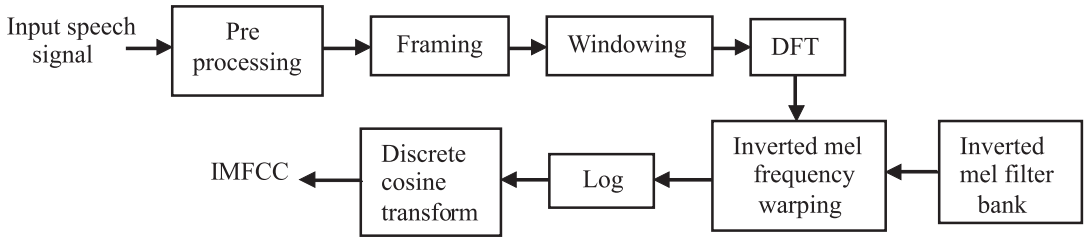

Fig. 4. Extraction of Inverted Mel Frequency Cepstral Coefficients.

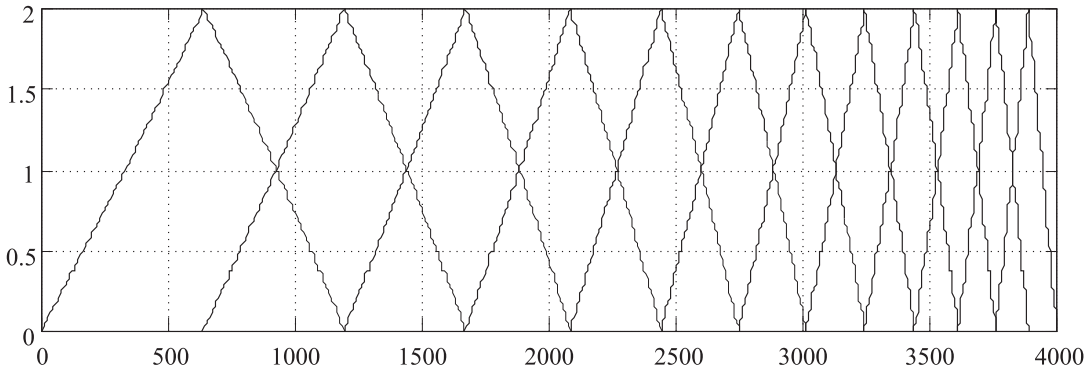

Fig. 5. Inverted Mel Frequency Scale Filter Bank.

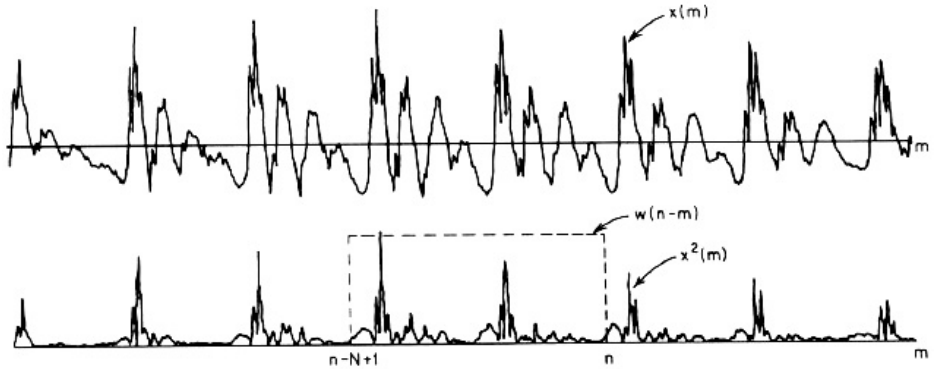

Fig. 6. Short Time Energy Computation. (After L. R. Rabiner and Schafer[13]).

The linear scale frequency can be obtained back from inverted mel scale frequency using

$$f = 4000 - 700\left(10^{\left(\frac{2146.1 - f_{\text{inverted}}}{2595}\right)} - 1\right) \tag{4}$$

2.3 Short time energy

Short time energy of a speech signal may be computed by the mathematical expression[11]

$$E_n = \sum_{l=-\infty}^{\infty} [x(l)\, w(n-l)]^2 \tag{5}$$

where $x(n)$ is speech signal under test and $w(n)$ is the window function used in the process of short time energy computation. Short time energy computation is illustrated in Fig. 6.

2.4 High frequency region of speech signal

Speech signal is concatenation of different sounds. It is found from the literature that all sounds contain equal amount of speaker specific information[14]. Temporal and spectral characteristics of each sound are different. It is found that Speaker specific information present in voiced and unvoiced phonemes have different frequency distributions. Based on the source of generation of the speech sound, the speech signal can be categorized in to voiced and unvoiced.

Voiced speech sound is generated by a periodic glottal source. In the production of voiced sound, air from the lungs is forced through the glottis with the tension of vocal cords and vocal cords begin to vibrate with relaxation oscillations. This produces quasi periodic pulses of air, which excites the vocal tract to produce voiced sound[11]. Generally the amplitude of the voiced sound is much higher than that of the unvoiced sound. Voiced part is of low frequency, high energy speech signal.

Unvoiced speech sound is generated by a noise like glottal source. Since source is random noise like, the resulting speech signal is also like random noise with low amplitude and high frequency. In the production of unvoiced sound, a narrow passage is created at some point in the vocal tract and the air current is forced through this passage at very high velocity to produce air turbulence. This noisy air current constitutes excitation for vocal tract to produce unvoiced sound. Unvoiced segment of the speech signal contains high frequency components which is the region of interest in the proposed algorithm.

The speech signal energy is a good representation which reflects the amplitude variations that occur in different sounds of the speech signal[11]. Thus short time energy of the speech signal serves as a metric for classifying the short segment of the speech signal into voiced and unvoiced[15]. As this can be observed from the Fig. 7, the value of the short time energy for the unvoiced segments is significantly smaller than that of voiced segments.

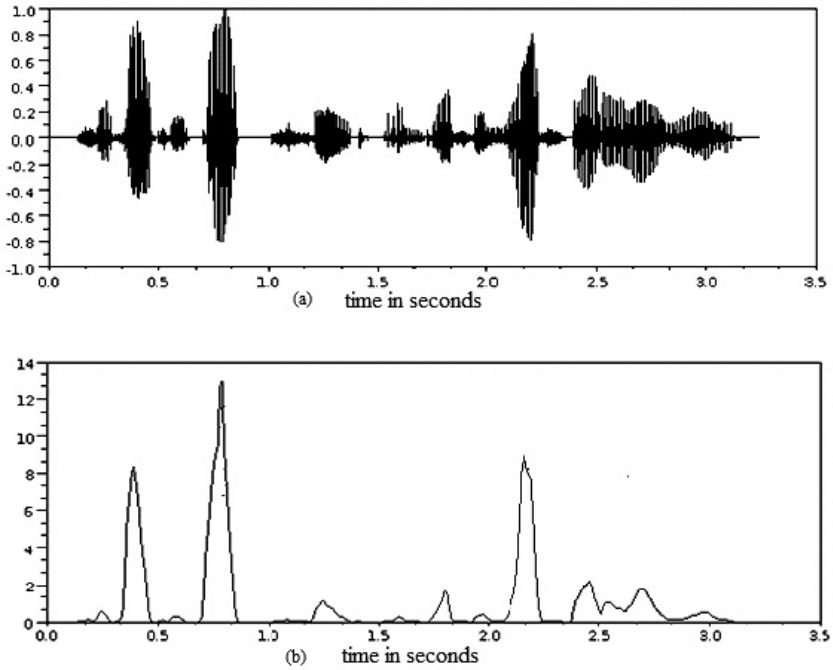

Fig. 7. (a) Speech Signal; (b) Short Time Energy.

3. Proposed Algorithm

The detailed proposed algorithm of feature extraction using MFCC and IMFCC is shown in Fig. 8.

In pre-processing the input speech signal amplitudes are normalized to a maximum of 1volt then the signal is passed through a filter which increases the signal energy at higher frequencies. After pre-processing the speech signal is divided into small frames. The speech signal is slowly varying signal with time. The temporal and spectral properties of the speech signal are reasonably stationary for a small period of time. Therefore short time analysis and processing is commonly adapted for speech signals. The speech sample is partitioned in to frames of N samples in the processes of frame blocking. There is a separation of M samples between two adjoining frames where $M < N$. The first N samples of the speech signal constitutes first frame and second frame begins after M samples of the first frame. This process continues until all speech samples are used up in the process of frame making. After frame blocking, each frame of the speech signal is multiplied with a causal window function which minimizes the signal discontinuities present at the beginning and end of each frame. The tapered window function also minimizes the spectral distortion. Typically Hamming window function is used for most of the speech signal applications which has the functional form shown in equation 7. The windowed frame of the speech signal has the form

$$y_n(m) = x_n(m)w(m) \quad 0 \leq m \leq N-1 \tag{6}$$

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad 0 \leq n \leq N-1 \tag{7}$$

Short time energy is computed for each windowed speech segment and compared with a threshold. If short time energy value is greater than the threshold, then we conclude that the speech segment is voiced and MFC coefficients are computed for the segment. If the energy obtained is lesser than the threshold, then the speech segment is unvoiced and IMFC coefficients are computed. MFCC and IMFCC will provide parametric representation of the speech signal which forms a set of robust feature vector of a speaker.

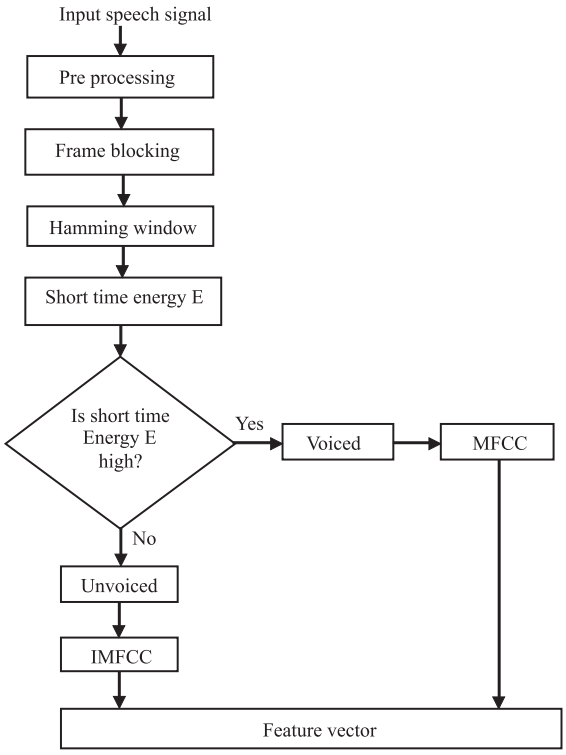

Fig. 8. Generation of Feature Vector Combining MFCC and IMFCC.

4. Implementation and Validation

The system is implemented on Mat lab 2010 platform. The data used for system validation is microphone recorded speech signals, sampled with 8000 samples per second. Recording is done for 5 seconds in normal room conditions. Totally 20 speech samples from both female and male speakers are recorded. Two speech samples with different utterances are taken from each speaker. One speech sample is used up in training the system and other is used for testing.

The input speech signal is segmented into frames of 256 samples with overlapping of 50% of the frame size and multiplied with Hamming window. For every frame 12 MFC or IMFC coefficients are computed. MFC and IMFC coefficients form speaker specific feature vector. These feature vectors are validated using vector quantization feature matching technique with 8 centroids. The accuracy of any speaker identification system is defined in terms of percentage of identification accuracy (PIA)[16].

$$\text{PIA} = \frac{\text{Number of utterances correctly identified}}{\text{Number of utterances under test}} \times 100 \tag{8}$$

5. Experimental Results

Here the metric used for performance analysis is percentage of identification accuracy (PIA). The results obtained are tabulated in Table 1. It is observed from table1 that, out of 20 speakers, 18 speakers are correctly identified in proposed method and 16 speakers are correctly identified in MFCC method. The PIA values are tabulated in Table 1. The results of proposed method cannot be compared with results of reference paper[16] because of difference in data base used for performance analysis. In reference paper[16] the PIA values are very high, database used are YOHO and

Table 1. Speaker Identification Results of 20 Speakers and PIA Values.

Feature Extraction Method	MFCC	MFCC and IMFCC
Number of speech samples tested	20	20
Number of speech samples identified correctly	16	18
Number of speech samples identified wrongly	4	2
PIA (%)	80	90

POLYCOST for performance analysis. For YOHO database the PIA values range from 89.55 to 98.9% for different feature extraction algorithm using MFCC and IMFCC. Whereas for POLYCOST, PIA values range from 82.17 to 95.44% for different feature extraction algorithm using fusion of MFCC and IMFCC.

6. Conclusions

A new feature extraction method using MFCC and IMFCC is proposed here. This method provides speaker information present in higher frequencies of speech signal with higher resolution which is being neglected in MFCC method. This makes the system more robust, and identification rate is improved. The performance of the system for both proposed method (90%) and MFCC (80%), is lesser compared to reference paper[16]. The speech samples for testing and training are recorded using ordinary lap top micro phone, in normal noisy room conditions. Secondly the number of speech samples used for system evaluation is very less which is only 20. But proposed method out performed traditional MFCC method of feature extraction. The drawback of the proposed method is processing time. The execution time required for proposed method is comparatively greater because the number of processing steps involved in feature vector computation is more. The proposed feature extraction method can generate better feature vectors but at the cost of processing time.

References

[1] J. P. Campell and Jr., Speaker Recognition: A Tutorial, *Proceeding of IEEE*, vol. 85, pp. 1437–1462, (1997).
[2] Osman Buyuk and Lavent M. Arslan, HMM-based Text-dependent Speaker Recognition with Handset-channel Recognition, *IEEE ICSPCA*, pp. 383–386, (2010).
[3] D. A. Reynolds and R. C. Rose, Robust Text-independent Speaker Identification using Gaussian Mixture Speaker Models, *IEEE Transaction on SAP*, vol. 3, no. 1, pp. 72–83, (1995).
[4] R. E. Wohiford, E. H. Jr. Wrench and B. P. Landell, A Comparison of Four Techniques for Automatic Speaker Recognition, *Proceedings of IEEE ICASSP*, vol. 5, pp. 908–911, (1980).
[5] B. Atal, Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification, *The Journal of the Acoustical Society America*, vol. 55, pp. 1304–1312, (1974).
[6] Sangeeta Biswas', Shamim Ahmadt and Md Khademul Islam Molladt, Speaker Identification using Cepstral Based Features and Discrete Hidden Markov Model, *Proceedings of IEEE ICICT*, pp. 303–306, (2007).
[7] D. A. Reynolds, Experimental Evaluation of Features for Robust Speaker Identification, *IEEE Transaction on SAP*, vol. 2, issue-4, pp. 639–643, (1994).
[8] Roma Bharti and Priyanka Bansal, Real Time Speaker Recognition System using MFCC and Vector Quantization Technique, *International Journal of Computer Applications*, vol. 117, no. 1, pp. 0975–8887, May (2015).
[9] Shahin and N. Botros, Speaker Identification using Dynamic Time Warping with Stress Compensation Technique, *Southeastcon'98, Proceedings of IEEE*, pp. 65–68, (1998).
[10] T. Kinnunen, T. Kilpeläinen and P. Fränti, Comparison of Clustering Algorithms in Speaker Identification, *Proceedings of IASTED, ICSPC*, pp. 222–227, (2000).
[11] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, India, Pearson Education, ISBN-13: 978-81-317-0513-1, (2012).
[12] S. Davis and P. Mermelstein, Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences, *IEEE Transactions on ASSP*, vol. 28, no. 4, pp. 357–366, August (1980).
[13] S. Chakroborty, A. Roy and G. Saha, Improved Closed Set Text Independent Speaker Identification by Combining MFCC with Evidence from Flipped Filter Banks, *IJSP*, vol. 4, no. 2, pp. 114–122, (2007).
[14] Kuruvachan K. George, K. Arunraj, K. T. Sreekumar, C. Santhosh Kumar and K. I. Ramachandran, Towards Improving the Performance of Text/Language Independent Speaker Recognition Systems, *IEEE EPSCICON*, pp. 38–41, (2014).
[15] R. G. Bachu, S. Kopparthi, B. Adapa and B. D. Barkana, Separation of Voiced and Unvoiced using Zero Crossing Rate and Energy of the Speech Signal, *American Society for Engineering Education*, vol. 2, pp. 114–122, (2008).
[16] Diksha Sharma and Israj Ali, A Modified MFCC Feature Extraction Technique for Robust Speaker Recognition, *Proceedings of ICACCI*, pp. 1052–1057, (2015).