International Journal of Research in Advanced Engineering and Technology
ISSN: 2455-0876; Impact Factor: RJIF 5.44
www.engineeringresearchjournal.com
Volume 2; Issue 4; July 2016; Page No. 37-41

Noise robust text-independent speaker identification using GFCC

Pardeep Sangwan

Maharaja Surajmal Institute of Technology, GGSIPU, New Delhi, India

Abstract

Automated speaker recognition performs efficiently in matched conditions but performance degrades in noisy conditions. Recent research shows that a relatively new feature, Gammatone Frequency Cepstral Coefficients (GFCC), more noise robust as compared to generally used Mel-Frequency Cepstral Coefficients (MFCC). To understand the intrinsic robustness of GFCC as compared to MFCC, speaker identification experiments are done to analyze their (dis)similarities systematically. This study reveals that the nonlinear rectification accounts for the noise robustness differences primarily. The present research proposes a novel paradigm which utilizes the strong pattern matching capability of ANNs for identification of speakers. Here ten speech samples are collected from 40 different speakers. Gammatone Frequency Cepstral Coefficients (GFCCs) are extracted for all the speakers and these coefficients are used to train ANN and then test signals are validated and verified for ANN. The results of identification under noisy conditions are very encouraging.

Keywords: Speaker Recognition, MFCC, GFCC, ANN, Noise Robustness

1. Introduction

Speech recognition is the process of extracting the linguistic message underlying a spoken utterance, while Speaker Recognition is concerned with extracting the identity of the person speaking the utterance. There are many applications of the Speaker Recognition like: computer access control, telephone voice authentication for banking or long distance calling, intelligent answering machines with personalized caller greetings and automatic speaker labeling of the recorded meetings for speaker dependent audio indexing [1]. Depending upon the application, speaker recognition is divided in two different tasks: identification and verification [2]. Speaker identification means to identify a person from a group of persons by matching input voice sample with a group of known voices and best matching signal gives the identified signal. This is also called sometimes as closed-set speaker identification. In speaker verification from a voice sample, identity of the speaker is determined whether he/she is a person who he/she claims to be or not. This is also known as open-set problem, because it requires distinguishing a claimed speaker's voice known to the system from a potentially large group of voices from imposter speakers which are unknown to the system.

These applications are further distinguished by the constraints placed on the speech used to train and test the system and the environment in which the speech signal is recorded. There are two types of system wiz text dependent and text independent systems. In text dependent system, the speech used for training and testing of the system could only be the same word or phrase. In a text independent system, the speech used to train and test the system could be entirely unconstrained [2].

Speaker recognition has been a research topic for many years and various types of speaker models have been studied. Hidden Markov Models (HMM) have become the most popular statistical tool for this task. The best results have been obtained using continuous density HMM (CHMM) for modeling the speaker characteristics. For the text-independent task, where the temporal sequence modeling capability of the HMM is not required, one state CHMM, also called a Gaussian mixture model (GMM), has been widely used as a speaker model [3]. Previously, it has been shown that GMM can perform even better than CHMM with multi-states [4]. Soft computing techniques like ANN, Fuzzy Logic etc, are also very efficient for speaker recognition [5].

Automatic speaker recognition systems perform very well in certain conditions, e.g. without noise, room reverberation, or channel variations. However, such conditions can hardly be met in practice. Real acoustic environments present various challenges to speaker recognition systems. Robustness of speaker recognition systems must be addressed for practical applications.

One challenge is channel/session variation. Recent NIST speaker recognition evaluations (SRE) have mainly focused on addressing the problem of channel variations in speaker verification. State-of-art systems use techniques such as joint factor analysis and i-vector based probabilistic linear discriminant analysis [6, 7]. Another challenge is additive noise. In our daily listening environments, speech often occurs simultaneously with noise sound. To improve noise robustness, Ming *et al*. propose to train speaker models in multiple noisy conditions to alleviate the mismatch with noisy test conditions [8]. Alternatively, one can employ speech enhancement algorithms to clean up noisy speech prior to speaker recognition.

The human ability to perform speaker recognition in noisy conditions has motivated studies of robust speaker recognition from the perspective of computational auditory scene analysis. In one such study [9], it is shown that a new speaker feature, Gammatone Frequency Cepstral Coefficients (GFCC) exhibits superior noise robustness to commonly used Mel-Frequency Cepstral Coefficients (MFCC) in speaker identification tasks [10]. As for the reason, the front-end of GFCC, the Gammatone filter bank, must be more noise-robust than that of MFCC, the Mel-filter bank. In particular, the frequency scales employed in the two filter banks were

believed to be the key difference although no convincing evidence was presented to support this hypothesis.
In this paper, we have used an ANN with back propagation algorithm for training of the neural network and weight adaptation with GFCC so that unknown speaker can be identified in noisy conditions.

2. Cepstral Coefficients

a) Mel-Frequency Cepstral Coefficients

Mel-frequency wrapping: Human perception of frequency contents of sounds for speech signal does not follow a linear scale. Thus for each tone with an 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 spacing below 1000 Hz and a logarithmic spacing above 1000Hz.As a reference point, the pitch of a 1 KHz tone, 40dB above the perceptual hearing threshold, is defined as 1000 Mels [8]. Therefore the following approximate formula is used to compute the Mels for a given frequency f in Hz.

$$f_{Mel} = 2595 * \log_{10}\left(1 + \frac{f}{700}\right) \quad (1)$$

Subjective spectrum is simulated by the use of a filter bank, one filter for each desired Mel-frequency component. That filter bank has a triangular band pass frequency response and the spacing as well as the bandwidth is determined by a constant Mel-frequency interval [11]. The Mel scale filter bank is a series of triangular band pass filters that have been designed to simulate the band pass filtering believed to occur in the auditory system. This corresponds to series of band pass filters with constant bandwidth and spacing on a Mel frequency scale [12].
Cepstrum: In this final step, the log Mel spectrum is converted back to time. The result is called the Mel Frequency Cepstral Coefficients (MFCC).The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the Mel spectrum coefficients (and so their logarithm) are real numbers, these can be converted in to the time domain. In this final step log Mel spectrum is converted back to time. The result is called the Mel Frequency Cepstral Coefficients (MFCC).The Discrete Cosine Transform (DCT) is done for transforming the Mel coefficients back to time domain [13].

$$C_n = \sum_{k=1}^{K} (\log S_k) \cos\left\{n\left(k - \frac{1}{2}\right) * \frac{\pi}{k}\right\} \quad (2)$$

n = 1, 2,... k

Whereas S_k, k = 1, 2, … K are the outputs of last step.
Complete process for the calculation of MFCC is shown in the figure 1.1. The speech signal is frame blocked with each frame having length of 30 ms with an overlap length of 20 ms. Hamming window is used for the analysis. The spectrum is calculated with the help of Discrete Fourier Transform (DFT) for each windowed frame. The Mel spectrum is obtained bypassing the spectrum through Mel-filter bank. Here 40 filters are used. Finally on the output of the Mel filter bank cepstral analysis is done using only 13 coefficients out of 40, then Discrete Cosine Transform (DCT) of the Mel spectrum is applied to obtain the set of feature vectors known as MFCC.

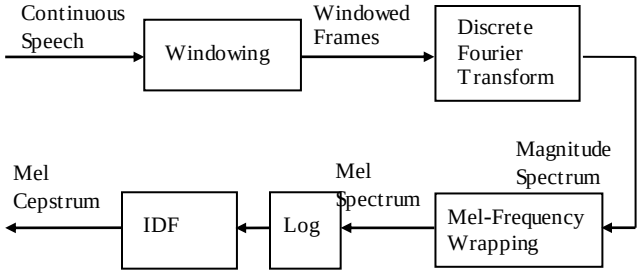


Fig 1.1: Calculation of MFCC coefficients

b) Gammatone Frequency Cepstral Coefficients

To obtain GFCC, first of all auditory filtering is done by decomposing an input signal into the time-frequency (T-F) domain using a bank of gammatone filters. Gammatone filters are derived from psychophysical and physiological observations of the auditory periphery and this filter bank is a standard model of cochlear filtering [14]. A bank of 64 filters is used whose centre frequencies range from 50 Hz to 4000 Hz or 8000 Hz depending on the sampling frequency of speech data. Since the filter output retains the original sampling frequency, fully rectified 64-channel filter responses are decimated to 100 Hz along the time dimension. This yields a corresponding frame rate of 10 ms, which is used in many short-time speech feature extraction methods. The magnitudes of the decimated outputs are then loudness-compressed by a cubic root operation,

$$G_m[i] = ||g|_{decimate}[i,m]|^{1/3} \quad (3)$$

where, $i = 0 \ldots N-1, m = 0 \ldots M-1$.
Here, N=64 refers to the number of frequency (filter) channels and M is the number of time frames obtained after decimation. The resulting responses $G_m[i]$ form a matrix, representing the T-F decomposition of the input. This T-F representation is a variant of cochleagram. Note that, unlike the linear frequency resolution of a spectrogram, a cochleagram provides a finer frequency resolution at low frequencies than at high frequencies.
A time slice of the above matrix is called Gammatone Feature (GF), and used to denote its i^{th} channel. Time index is dropped for simplicity. Here, a GF vector comprises 64 frequency components. Note that the dimension of a GF vector is larger than that of MFCC vectors used in a typical speaker recognition system. Additionally, because of the frequency overlap among neighbouring filter channels, GF components are correlated with each other. In order to reduce dimensionality and de-correlate the components, DCT is applied to a GF. The resulting coefficients are called Gammatone Frequency Cepstral Coefficients (GFCC) [10, 15, 16]. Specifically, cepstral coefficients (C_j), are obtained from a GF as follows:

$$C_j = \sqrt{\frac{2}{N}} \sum_{i=0}^{N-1} G[i] \cos\left(\frac{j\pi}{2N}(2i+1)\right) \quad (4)$$

Where j = 0…..N-1.

In fact, the newly derived coefficients are not cepstral coefficients because a cepstral analysis requires a log operation between the first and the second frequency analysis for the deconvolution purpose [17]. Here, these are called cepstral coefficients because of the functional similarities between the above transformation and that of a typical cepstral analysis in the derivation of MFCC.

3. Artificial neural network

Artificial neural networks are massively connected networks of computational "neurons", and represent parallel-distributed processing structures. The inspiration for ANN has come from the biological architecture of the neurons in the human brain. A key characteristic of neural networks is their ability to approximate arbitrary nonlinear functions. Since machine intelligence involves a special class of highly nonlinear decision making, neural network would be effective there. Through the use of neural networks, an intelligent system would be able to learn and perform high-level cognitive task [18]. For example, an intelligent system would only need to be presented a goal; it could achieve its objective through continuous interaction with its environment and evaluation of the responses by means of the neural networks. A neural network consists of a set of nodes, usually organized into layers, and connected through weight elements called synapse. At each node, the weighted inputs are summed (aggregated), thresholded, and subjected to an activation function in order to generate the output of that node. These operations are shown in the figure 1.2.

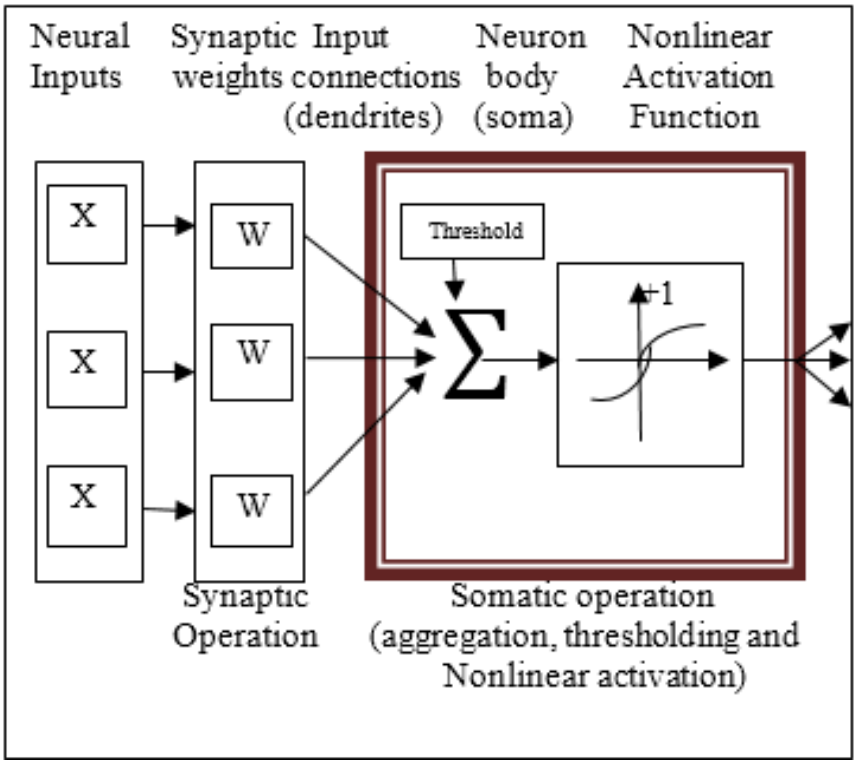
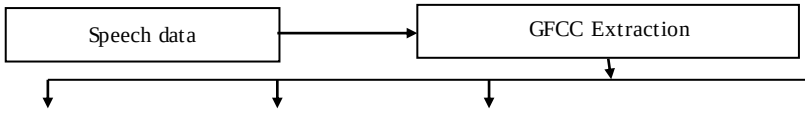


Fig 1.2: The operation at a node of a neural network [19]

4. Proposed Paradigm

Ten speech samples are collected from 40 different speakers out of which 30 speakers are male and 10 speakers are female. GFC Coefficients are extracted for all the speakers.



(Speaker1-0).1	(Speaker1-0).0	(Speaker3-0).0			(SpeakerN-0).0
(Speaker1-1).1	(Speaker1-1).0	(Speaker3-1).0			(SpeakerN-1).0
(Speaker1-2).1	(Speaker1-2).0	(Speaker3-2).0			(SpeakerN-2).0
(Speaker1-3).1	(Speaker1-3).0	(Speaker3-3).0			(SpeakerN-3).0
(Speaker1-4).1	(Speaker1-4).0	(Speaker3-4).0	-	-	(SpeakerN-4).0
(Speaker1-5).1	(Speaker1-5).0	(Speaker3-5).0			(SpeakerN-5).0
(Speaker1-6).1	(Speaker1-6).0	(Speaker3-6).0			(SpeakerN-6).0
(Speaker1-7).1	(Speaker1-7).0	(Speaker3-7).0			(SpeakerN-7).0
(Speaker1-8).1	(Speaker1-8).0	(Speaker3-8).0			(SpeakerN-8).0
(Speaker1-9).1	(Speaker1-9).0	(Speaker3-9).0			(SpeakerN-9).0
(Speaker2-0).0	(Speaker2-0).1	(Speaker2-0).0			(Speaker2-0).0
(Speaker2-1).0	(Speaker2-1).1	(Speaker2-1).0	-	-	(Speaker2-1).0
(Speaker2-2).0	(Speaker2-2).1	(Speaker2-2).0			(Speaker2-2).0
(Speaker2-3).0	(Speaker2-3).1	(Speaker2-3).0			(Speaker2-3).0

(Speaker2-4).0	(Speaker2-4).1	(Speaker2-4).0			(Speaker2-4).0
(Speaker2-5).0	(Speaker2-5).1	(Speaker2-5).0			(Speaker2-5).0
(Speaker2-6).0	(Speaker2-6).1	(Speaker2-6).0			(Speaker2-6).0
(Speaker2-7).0	(Speaker2-7).1	(Speaker2-7).0			(Speaker2-7).0
(Speaker2-8).0	(Speaker2-8).1	(Speaker2-8).0			(Speaker2-8).0
(Speaker2-9).0	(Speaker2-9).1	(Speaker2-9).0			(Speaker2-9).0
(Speaker3-0).0	(Speaker3-0).0	(Speaker3-0).1			(Speaker3-0).0
(Speaker3-1).0	(Speaker3-1).0	(Speaker3-1).1			(Speaker3-1).0
(Speaker3-2).0	(Speaker3-2).0	(Speaker3-2).1			(Speaker3-2).0
(Speaker3-3).0	(Speaker3-3).0	(Speaker3-3).1			(Speaker3-3).0
(Speaker3-4).0	(Speaker3-4).0	(Speaker3-4).1	-	-	(Speaker3-4).0
(Speaker3-5).0	(Speaker3-5).0	(Speaker3-5).1			(Speaker3-5).0
(Speaker3-6).0	(Speaker3-6).0	(Speaker3-6).1			(Speaker3-6).0
(Speaker3-7).0	(Speaker3-7).0	(Speaker3-7).1			(Speaker3-7).0
(Speaker3-8).0	(Speaker3-8).0	(Speaker3-8).1			(Speaker3-8).0
(Speaker3-9).0	(Speaker3-9).0	(Speaker3-9).1			(Speaker3-9).0
|	|	|	-	-	|
|	|	|			|
|	|	|			|
(SpeakerN-0).0	(SpeakerN-0).0	(SpeakerN-0).0			(SpeakerN-0).1
(SpeakerN-1).0	(SpeakerN-1).0	(SpeakerN-1).0			(SpeakerN-1).1
(SpeakerN-2).0	(SpeakerN-2).0	(SpeakerN-2).0			(SpeakerN-2).1
(SpeakerN-3).0	(SpeakerN-3).0	(SpeakerN-3).0			(SpeakerN-3).1
(SpeakerN-4).0	(SpeakerN-4).0	(SpeakerN-4).0	-	-	(SpeakerN-4).1
(SpeakerN-5).0	(SpeakerN-5).0	(SpeakerN-5).0			(SpeakerN-5).1
(SpeakerN-6).0	(SpeakerN-6).0	(SpeakerN-6).0			(SpeakerN-6).1
(SpeakerN-7).0	(SpeakerN-7).0	(SpeakerN-7).0			(SpeakerN-7).1
(SpeakerN-8).0	(SpeakerN-8).0	(SpeakerN-8).0			(SpeakerN-8).1
(SpeakerN-9).0	(SpeakerN-9).0	(SpeakerN-9).0			(SpeakerN-9).1

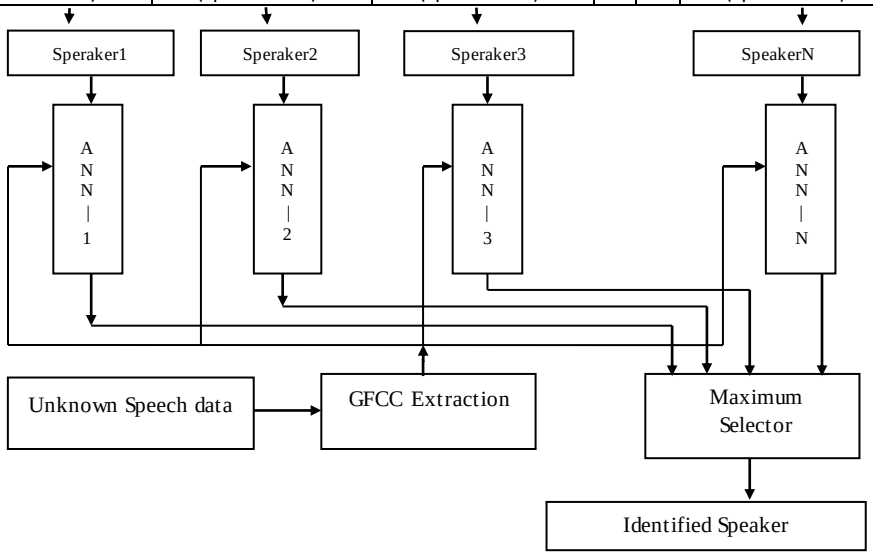


Fig 1.3: ANN based Speaker Identification System [19]

Then, the ANN is trained for a particular speaker by taking output parameter of that particular speaker as '1' and of all other speakers as '0' and repeating this procedure for all the speakers one by one to get feature matrix of same. The block diagram for the ANN based speaker identification is as shown in figure 1.3. The steps for the proposed paradigm are as follows:

Step 1: Extraction of the GFC Coefficients of all the speech signals.

Step 2: Training the ANNs for each data vector.

Step 3: Getting the test speech signal from speaker to be identified.

Step 4: Extraction of the GFC Coefficients of the test signal.

Step 5: Input these GFC Coefficients to the each trained ANN's.

Step 6: Collecting the outputs of each ANN and the ANN which gives the maximum output corresponds to the identified speaker.

5. Experimental Results

Speech samples have been collected and to check the robustness of the model, data is collected in a noisy environment so that realistic results could be obtained. The extracted MFCC and GFCC are used to train and test an artificial neural network having 2 hidden layers and 10 neurons, which has given the best result in a previous study [19]. The results show that ANN with GFCC have better success rate than ANN with MFCC.

Table 1: Validation Results

S. No.	No. of Neuron	No. of Hidden Layers	No. of Test Sample	With MFCC		With GFCC	
				Accurately Identified	Success Rate	Accurately Identified	Success Rate
1	5	1	50	31	62%	37	74%
2	10	2	50	37	74%	41	82%
3	30	3	50	35	70%	38	76%

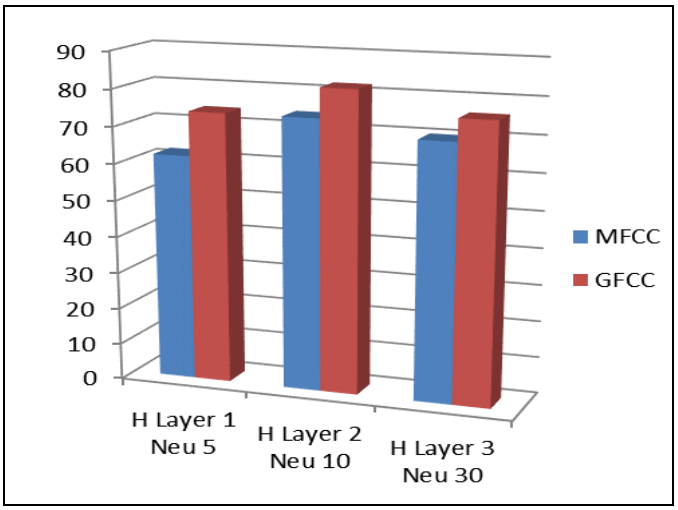


Fig 1.3: Percentage Success Rate

6. Conclusion

Speaker recognition is an emerging and very important technique in this new era of human-machine interaction. It has two main tasks: speaker identification and speaker verification. Various methods have been proposed on the basis of statistical techniques in the area of speaker identification. Ten speech samples are collected from 40 different speakers. MFCC coefficients are extracted for all the speakers and these coefficients are used to train ANN and Fuzzy Logic and then test signals are validated and verified using MATLAB. The results are promising using ANN in comparison to fuzzy logic.

7. References

1. Barbu T. Comparing Various Voice Recognition Techniques, *Proceedings of the 5th Conference on Speech Technology and Human-Computer Dialogue*, 2009.
2. Reynolds DA. An Overview of Automatic Speaker Recognition Technology, *IEEE*, 2002.
3. Reynolds DA, Rose RC. Robust Text Independent Speaker Identification using Gaussian Mixture Speaker Models, *IEEE Transaction on Speech and Audio Processing*, Jan, 1995.
4. Kinnunen T, Karpov E, Franti P. Real Time Speaker Identification and Verification, *IEEE Transaction on Audio, Speech and Language Processing*, Jan, 2006.
5. Anup Kumar Paul, Dipankar Das, Md. Mustafa Kamal. Bangla Speech Rrecognition System using LPC and ANN, *Seventh International Conference on Advances in Pattern Recognition*, 2009.
6. Dehak N, Kenny P, Dehak R, Dumouchel P, Ouellet P. Front-end Factor Analysis for Speaker Verification", *IEEE Trans. On Audio, Speech, and Language Processing*, May 2010; 19(4):788-798.
7. Burget L, Plchot O, Cumani S, Glembek O, Matejka P, Brummer N. Discriminatively Trained Probabilistic Linear Discriminant Analysis for Speaker Verification, In *Proc. ICASSP*, 2011, 4832-4835.
8. Ming J, Hazen TJ, Glass JR, Reynolds DA. Robust Speaker Recognition In Noisy Conditions, *IEEE Trans. Audio, Speech, and Language Processing*, 2007; 15(5):1711-1723.
9. Zhao X, Shao Y, Wang DL. CASA-Based Robust Speaker Identification, *IEEE Trans. Audio, Speech and Language Processing*, 2012; 20(5):1608-1616,
10. Shao Y, Srinivasan S, Wang DL. Incorporating Auditory Feature Uncertainties In Robust Speaker Identification, in *Proc. ICASSP*, 2007, 277-280.
11. Chen J, Paliwal KK, Mizumachi M, Nakamura S. Robust MFCC Derived from Differentiated Power Spectrum, *Eurospeech*, 2001.
12. Sheeraz Memon, Margaret Lech, Ling He. Using Information Theoretic Vector Quantization for Inverted MFCC based Speaker Verification, *2nd International Conference on Computer, Control and Communication*, 2009.
13. Md. Sahidullah, Goutam Saha. On the use of Distributed DCT in Speaker Identification, *Annual IEEE India Conference (INDICON)*, 2009.
14. Patterson RD, Holdsworth J, Allerhand M. Auditory Models as Pre-Processors for Speech Recognition, in *The Auditory Processing of Speech: From Sounds to Words*, 1992, 67-83.
15. Xiaojia Zhao, DeLiang Wang. Analyzing Noise Robustness of MFCC and GFCC Features in Speaker Identification, in *Proc. ICASSP*, 2013, 7204-7208.
16. Shao Y, Wang DL. Robust speaker identification using auditory features and computational auditory scene analysis, in *Proc. ICASSP*, 2008, 1589-1592.
17. Oppenheim AV, Schafer RW, Buck JR. *Discrete-Time Signal Processing*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 1999.
18. Karrey FO, Clarence De Silva. *Soft-Computing and Intelligent System Design*, Pearson Education.
19. Sangwan P, Sheoran D. Text-Independent Speaker Identification System using Soft-computing Techniques, *UACEE-IJCSIA*, 2013; 3(3):92-95,
20. www.voxforge.com