Proc. of the Second Intl. Conf. on Advances in Electronics, Electrical and Computer Engineering -- EEC 2013

ISBN: 978-981-07-6935-2

Text Independent Speaker Identification Using Soft-Computing Techniques

[Pardeep Sangwan, Dinesh Sheoran]

***Abstract*—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. In the past various models have been proposed for the identification of speakers with the help of statistical techniques like Hidden Markov Model, Gaussian Mixture Model. As the Artificial Neural Networks (ANNs) are the universal classifiers. 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 Mel Frequency Cepstral Coefficients (MFCC) are extracted for all the speakers and these coefficients are used to train ANN and then test signals are validated and verified for ANN as well as for Fuzzy Logic. The results of identification are very encouraging.**

Keywords— Speaker Identification, Artificial Neural Network, Fuzzy Logic, MFCC, Hidden Markov Model, Discrete Fourier Transform, Discrete Cosine Transform

I. Introduction

Modern technology is advancing in the direction of better man-machine interaction. Initial steps for human-machine communications led to the development of the keyboard, the mouse, the trackball, the touch-screen, and the joystick. However none of these communication devices provides the ease of use of speech, which has been the most natural form of communication between humans for centuries. Speech recognition by machine refers to the capability of a machine to convert human speech to a textual form, providing a transcription or interpretation of everything the human speaks while the machine is listening.

Speech recognition is the classification of spoken words by a machine. The words are transformed into a format that a machine can understand then matched in some way against a template or dictionary of previously identified sounds. 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 weather 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]. In this paper, we have used an ANN with back propagation algorithm for training of the neural network and weight adaptation so that any unknown speaker can be identified.

II. Soft Computing techniques

A. *Artificial Neural Networks*

Artificial neural network 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

Pardeep Sangwan, Assistant Professor

Maharaja Surajmal Institute of Technology, GGSIPU, New Delhi
India

Dinesh Sheoran, Assistant Professor

Maharaja Surajmal Institute of Technology, GGSIPU, New Delhi
India

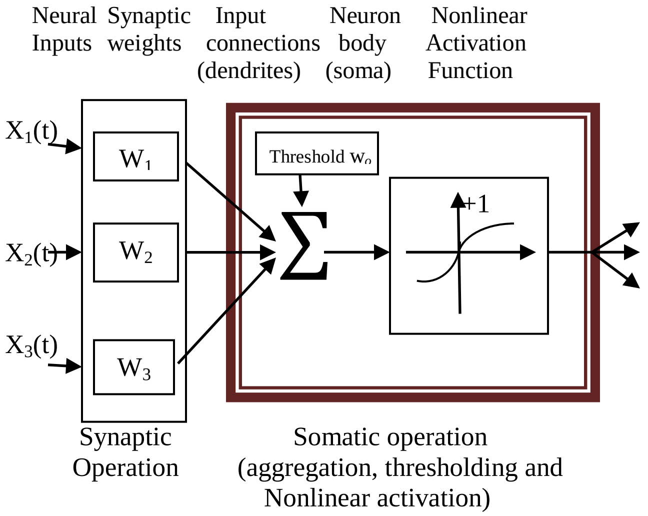

Neural Inputs
Synaptic weights
Input connections (dendrites)
Neuron body (soma)
Nonlinear Activation Function
X1(t)
X2(t)
X3(t)
W1
W2
W3
Threshold w0
Σ
+1
Synaptic Operation
Somatic operation (aggregation, thresholding and Nonlinear activation)

Fig 1.1: The operation at a node of a neural network [6]

approximate arbitrary nonlinear functions. Since machine intelligence involves a special class of highly nonlinear decision making, neural network would be effective there. Furthermore, the process of approximation of a nonlinear function (i.e. system identification) by interacting with a system and employing data on its behavior may be interpreted as “Learning”. Through the use of neural networks, an intelligent system would be able to learn and perform high-level cognitive task [6]. 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 1.1.

B. *Fuzzy Logic*

Fuzzy Logic techniques have been used in image-understanding applications such as detection of edges, feature extraction, classification, and clustering. Fuzzy logic poses the ability to mimic the human mind to effectively employ modes of reasoning that are approximate rather than exact. In traditional hard computing, decisions or actions are based on precision, certainty, and vigor. Precision and certainty carry a cost. In soft computing, tolerance and impression are explored in decision making. The exploration of the tolerance for imprecision and uncertainty underlies the remarkable human ability to understand distorted speech, decipher sloppy handwriting, comprehend nuances of natural language, summarize text, and recognize and classify images. With FL, we can specify mapping rules in terms of words rather than numbers. Computing with the words explores imprecision and tolerance. Another basic concept in FL is the fuzzy if–then rule. Although rule-based systems have a long history of use in artificial intelligence, what is missing in such systems is machinery for dealing with fuzzy consequents or fuzzy antecedents. In most applications, an FL solution is a translation of a human solution. Thirdly, FL can model nonlinear functions of arbitrary complexity to a desired degree of accuracy. FL is a convenient way to map an input space to an output space. FL is one of the tools used to model a multi-input, multi-output system.

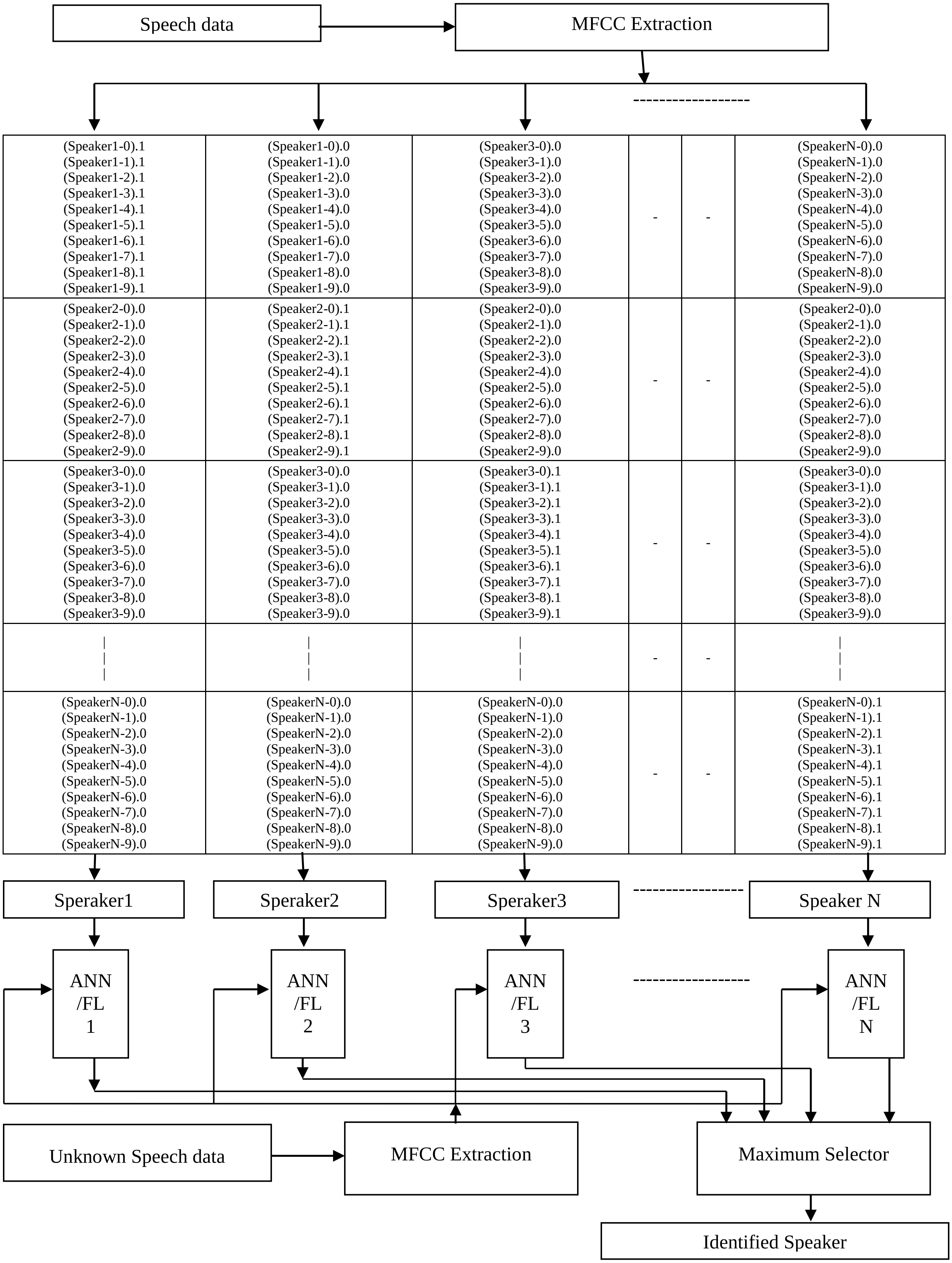
Zadeh introduced the term fuzzy logic in his seminal work “Fuzzy sets,” which described the mathematics of fuzzy set theory (1965). Plato laid the foundation for what would become fuzzy logic, indicating that there was a third region beyond True and False. It was Lukasiewicz who first proposed a systematic alternative to the bi-valued logic of Aristotle. The third value Lukasiewicz proposed can be best translated as “possible,” and he assigned it a numeric value between True and False. Later he explored four-valued logic and five-valued logic, and then he declared that, in principle, there was nothing to prevent the derivation of infinite-valued logic. FL provides the opportunity for modeling conditions that are inherently imprecisely defined. Fuzzy techniques in the form of approximate reasoning provide decision support and expert systems with powerful reasoning capabilities. The permissiveness of fuzziness in the human thought process suggests that much of the logic behind thought processing is not traditional two-valued logic or even multi valued logic, but logic with fuzzy truths, fuzzy connectiveness, and Fuzzy rules of inference. A fuzzy set is an extension of a crisp set. Crisp sets allow only full membership or no membership at all, whereas fuzzy sets allow partial membership. In a crisp set, membership or non-membership of element x in set A is described by a characteristic function $\mu_A(x)$, where $\mu_A(x) = 1$ if $x \in A$ and $\mu_A(x) = 0$ if $x \notin A$. Fuzzy set theory extends this concept by defining partial membership. A fuzzy set A on a universe of discourse U is characterized by a membership function $\mu_A(x)$ that takes values in the interval [0, 1]. Fuzzy sets represent common sense linguistic labels like slow, fast, small, large, heavy, low, medium, high, tall, etc. A membership function is essentially a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1.

III. Proposed Paradigm

Ten speech samples are collected from 40 different speakers out of which 30 speakers are male and 10 speakers are female. MFCC coefficients are extracted for all the speakers. 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 size and then same is done with the Fuzzy Logic Technique. The block diagram for the ANN based speaker identification is as shown in figure 1.2. The steps for the proposed paradigm are as follows:

Step-1: Extraction of the MFCC coefficients of all the speech signals.

Step 2: Training the ANNs for each data vector.



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

Fig 1.2: ANN/FL based Speaker Identification System

Step 3: Getting the test speech signal from speaker to be identified.
Step 4: Extraction of the MFCC coefficients of the test signal.
Step 5: Input these MFCC 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.
Step 7: The whole procedure is repeated with Fuzzy Logic.

IV. Experimental Results

Speech samples have been collected from [12]. To check the robustness of the model data is collected in a noisy environment so that realistic results could be obtained.

Table 1: Validation results

S. No.		No. of Neurons	No. of Hidden Layers	No. of Test Samples	Success Rate	%age Success Rate
1	ANN	5	1	50	34	68
2		10	2	50	39	78
3		30	3	50	37	74
4	**Fuzzy Logic**			50	35	70

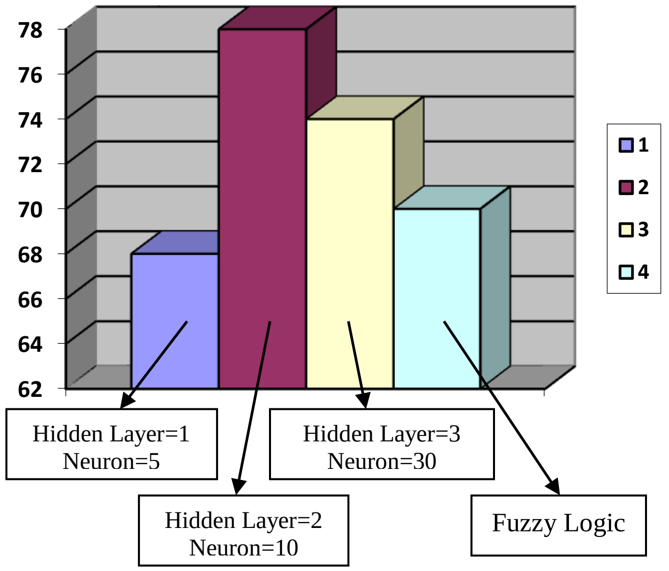


Fig 1.3: Percentage Success Rate

V. 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.

References

[1] T. Barbu, "Comparing Various Voice Recognition Techniques,"Proceedings of the 5th Conference on Speech Technology and Human-Computer Dialogue, 2009.

[2] D. A. Reynolds, "An Overview of Automatic Speaker Recognition Technology", IEEE, 2002.

[3] D. A. Reynolds, R.C. Rose, "Robust Text Independent Speaker Identification using Gaussian Mixture Speaker Models", IEEE Transaction on Speech and Audio Processing, Jan, 1995.

[4] T. Kinnunen, E. Karpov, P. Franti, "Real Time Speaker Identification and Verification", IEEE Transaction on Audio, Speech and Language Processing, Jan, 2006.

[5] Anup Kumar Paul, Dipankar Das and Md. Mustafa Kamal, "BanglaSpeechRrecognition System using LPC and ANN", Seventh InternationalConference on Advances in Pattern Recognition, 2009.

[6] F. O. Karrey, Clarence DeSilva, "Soft-Computing and Intelligent System Design", Pearson Education.

[7] Wang Chen, Miao Zhenjiang and Meng Xiao, "Differential MFCC and Vector Quantization used for Real-Time Speaker Recognition System",Congress on Image and Signal Processing, 2008.

[8] Gong Wei-Guo, Yang Li-Ping and Chen Di, "Pitch Synchronous Based Feature Extraction for Noise-Robust Speaker Verification", Congress onImage and Signal Processing(CISP '08), vol. 5, 2008.

[9] J. Chen , K. K. Paliwal, M. Mizumachi and S. Nakamura, "RobustMFCC Derived from Differentiated Power Spectrum ", Eurospeech 2001.

[10] SheerazMemon, Margaret Lech and Ling He, "Using Information Theoretic Vector Quantization for Inverted MFCC based Speaker Verification", 2nd International Conference on Computer, Control andCommunication, 2009.

[11] Md. Sahidullah and GoutamSaha, "On the use of Distributed DCT inSpeaker Identification", Annual IEEE India Conference(INDICON),2009.

[12] WWW.VOXFORGE.COM