Mel Frequency Cepstral Coefficients (MFCC) Based Speaker Identification in Noisy Environment Using Wiener Filter

Paresh M. Chauhan
Dept. of Information Technology
Dharmsinh Desai University
Nadiad, India
pareshchauhanm4444@gmail.com

Nikita P. Desai
Dept. of Information Technology
Dharmsinh Desai University
Nadiad, India
npd_ddit@yahoo.com

***Abstract*—Speech processing is now an emerging technology of signal processing. Some research areas of speech processing are recognition of speech, speaker identification (SI), speech synthesis etc. Speaker identification is important research area of speech processing. SI means identifying the speaker based on his spoken speech. The main use of SI is to recognize the speech owner based on the speaking style of the speaker. SI is mainly used in forensic analysis, home control system, database access services etc. For SI two things are essential. One is feature extraction and another is feature matching. Feature extraction is extraction of small information from the available audio wave signal. That information can be used to represent the particular speaker. For SI, There are many feature extraction techniques like LPC (Linear Predictive Coefficients), MFCC (Mel Frequency Cepstral Coefficients), PLP (Perceptual Linear Predictive Coefficients) and many more are used. MFCC is one of them and it gives good (efficient) identification results. Factor affecting on SI is noise, sampling rate, number of frames etc., and among them noise is the most critical factor. We found that MFCC is not much effective in the noisy environment, especially when the noise condition mismatch. The identification rate becomes poor and poor when the noise level increases. To improve the performance of SI in a real world noisy environment, we propose a technique which is a variant of MFCC. Proposed MFCC includes wiener filter which is good for handling the noise in speech. In this paper, it is suggested that the wiener filter is effective in the frequency domain rather than the time domain based on our experiments. We got 88.57% average identification rate with NOIZEUS database by our proposed technique. In feature matching, the unknown speech is classified by using some classifier. We have used neural network for feature matching.**

Keywords—Speaker identification, Noise mismatch, Wiener filter, Real world noisy environment

I. INTRODUCTION

Speech processing is quite simple; it is man's basic principle means of communication and is, therefore, a convenient and effective way of communication with machines. Speech communication has to be efficient and robust. Basically, Speaker identification is of two types. One is text dependent and another is text independent. In text dependent, speaker identification is done on the closed set words or sentences. The utterances which are used for training of model are also used for testing. Second is text independent. In text independent, the utterances used for training and testing are not same. The user may speak any words he wants in text independent SI system. Let's see the modules of speaker identification. Fig.1 shows the training module.

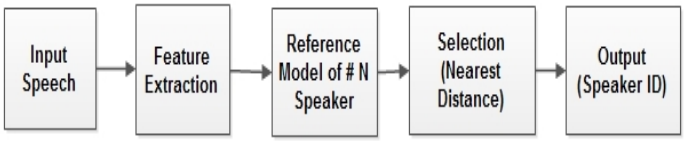

Fig. 1 Training module of SI

All identification systems have to pass from two different phases. The first one refers to the training phase while the second one refers to the testing phase. In the first training phase, each registered speaker has to provide speech samples, so that the system can build or train a reference model for that registered speaker. In case of speaker verification system, in addition, specific threshold (distance) is also computed from the training samples.

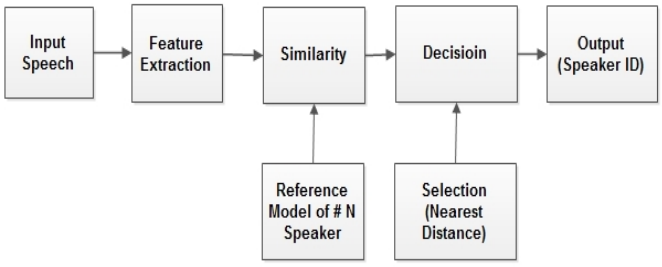

Fig. 2 Testing module of SI

During the testing phase (shown in Fig. 2), the input speech is matched with stored training reference model(s) and final decision is made.

Speaker identification is a difficult task and it is still an active hot research area. Speaker identification works based on the premise that a person's speech exhibits the unique characteristics. However, highly variant of input voice signal makes SI vary challenging. The principle source of variance is the speaker himself. Speech signals in training and testing period can be greatly different due to many facts such as

people's voice change with time, health conditions (e.g. the speaker has a fever), speaking rates, etc. There are also some other factors, except speaker variability, that present a challenge to speaker identification technology. Examples of these are acoustical noise, variations in recording environments and conditions (e.g. speaker uses different telephone handsets) etc. All these challenges make the system "Robust".

So what characterizes a "Robust System"? When people use the SI system in the real environment? The answer is quite simple. Human's ears can constantly adapt the characteristics of the environment such as the transmission channel, background noise, speaker of the speech etc., We always hope that our SI system do the similar things that the human's ear can do. We also expect that the recognition performance is effective. Unfortunately, at present, adaptation capacities to unknown conditions on machines are greatly poorer than that of ours. In fact, the performance of speaker identification systems trained with clean speech may degrade significantly in the real world because of the mismatch between the training and testing environments (noise mismatch condition). If the identification accuracy is very good under mismatch conditions of noise than that system is called "Robust". So, our aim is to make the speaker identification system robust in the real world environment. Our aim is to build a "Robust" SI system. The outlines of this research paper are as follows. In section II review of literature is shown; in section III we will see propose approach. In section IV we will see experimental results. And in section V we will see the conclusion.

II. REVIEW OF LITERATURE

Md. Rabiul Islam, Md. Fayzur Rahman [9] proposed their technique which is a PCA (Principle Component Analysis) based on genetic algorithm. They tested their technique on the NOIZEUS database and they got 85.73% identification rate.

Kevin R. Farrell et. al. [5] used neural tree network (NTN) which was examined for the text independent SI and did modifications in NTT. The NTN is a hierarchical classifier that combines the properties of decision trees and feed forward neural networks. The modified NTN uses discriminate learning to partition feature space as opposed to the more common clustering approaches, such as vector quantization. The modified NTN was evaluated for both closed and open set speaker identification experiments using the TIMIT database. The performance of the modified NTN is compared with vector quantization classifiers.

Juhani Saastamoinen, Zdenek Fiedler [4] evaluated the accuracy of the MFCC-based speaker recognition method. They analysed the recognition results using speech signals of everyday real life environments. They also studied the mismatch effects of sample length, speaking style, text-dependency, language, sample quality microphone and noise. The experiments were performed on self-collected corpus of 30 subjects. The most dominating factors are microphone, noise, degrading of the sample rate and quality.

Longbiao Wang, Kazue Minami [6] had stated the difference between conventional SI systems with their proposed system. They proposed the technique MFCC integrating with the phase information. They also describe the effectiveness of the phase information with MFCC that the error rate was reduced for clean database. They also found that the phase information with MFCC will reduced the error rate of identification of the speaker is reduced by 20% to 70%.

Satyanand Singh, Dr. E.G Rajan [11] presented the papers on various factors affecting in the accuracy of the speaker identification with MFCC and vector quantization (VQ). The investigation analysis was done on the speech signal from day to day life in the surrounding environments. They studied the mismatch that affecting of speaking style, mimicry, text-dependency, speaking language, voice sample length, utterance sample quality, the quality of microphone and surrounding noise. The corpuses of 10 people of 20 utterance subjects were collected which were indicate that any mismatch degrades the recognition accuracy. Noise is the most dominating factors among others factor is one of the conclusions of their research.

Martinez, J. et. al. [7] presented a fast and accurate automatic voice recognition algorithm. They had used MFCC to extract the features from voice and for classification vector quantization technique was used. MFCC is usually used in data compression, it allows building model for probability functions by the distribution of different vectors and the results they achieved were 100% of precision with a database of 10 speakers.

Sourabh Ravindran1, David V. Anderson [12] had published paper on the noise-robustness of mel-frequency cepstral coefficients (MFCCs) and gave a way to improve the performance in noisy conditions. They suggested for making MFCC more robust to noise for the improvements based on a more accurate model of the early auditory system while preserving their class discrimination ability. Speech versus non-speech classification and speech recognition are chosen to evaluate the performance gains afforded by the modifications.

One feature extraction technique proposed by Azzam Sleit, Sami Serhan [1] which had noteworthy improvement in the recognition rate of the SI system. They proposed two new techniques for SI based on utilizing the features generated from the MFCC. Those techniques are based on histograms for the features with some predefined interval lengths. The first technique builds histogram for all data in the feature vectors for each speaker while the second one builds a histogram for each feature column in the feature set of each speaker.

M. G. Sumithra, A. K. Devika [10] did the comparative work of various feature extraction techniques for the text independent SI. The techniques such as mel-frequency cepstral coefficients (MFCC), modified mel-frequency cepstral coefficients (MMFCC), revised perceptual liner prediction (RPLP), linear predictive coefficient cepstrum (LPCC) and bark frequency cepstral coefficients (BFCC) are implemented and based on the performance of computation time comparison was done. They had used vector quantization for feature matching. For testing, TIMIT database of 100 speakers was used.

III. PROPOSED METHOD

On an average MFCC gives the effective (efficient) identification accuracy in the clean environment. We propose a feature extraction technique with the modified architecture of MFCC. So, our proposed technique is a variant of MFCC which is for real world environment as well as for clean environment. The difference between MFCC and the proposed technique is the round shape block that is 'filtering'. Fig. 3 shows the proposed architecture with input and output of each step.

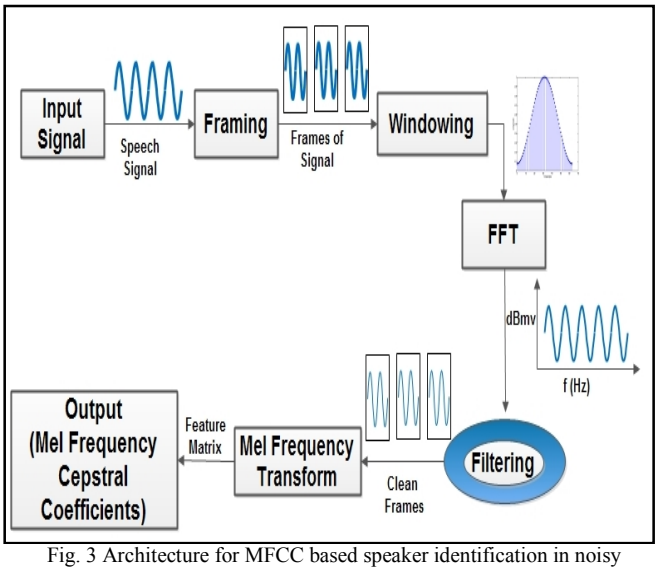

Fig. 3 Architecture for MFCC based speaker identification in noisy environment

Input signal: - An analog signal is continuous time varying signal i.e. in an analog audio signal, the instantaneous voltage of the signal varies continuously with the pressure of the sound waves. Fig.4 shows the example of our input speech signal.

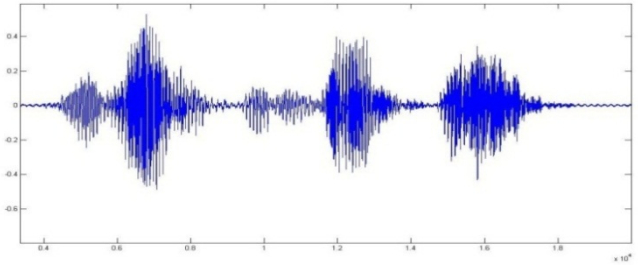

Fig. 4 Example of input speech signal

The input speech signal is stored as a .wav form. That .wav file is going as an input to the next step.

Framing: - An audio signal is constantly changing. So, to simplifying the things, we assume that on short time scales the audio signal doesn't change much (when we say it doesn't change, we mean that statistically i.e. statistically stationary, obviously the all samples are constantly changing on even short time scales). That's why we frame the signal into 20-40ms frames. Fig. 5 shows one frame of audio signal. Numbers of frames are generally 256 (in power of 2) because when FFT is calculated, it would be the simple if the numbers of frames are in power of 2.

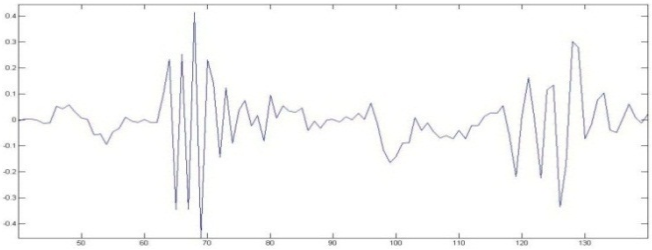

Fig. 5 One frame of speech signal

Windowing: - After framing, for minimization of the discontinuity of the signal, next step is windowing to each individual frame. The concept applied here is to minimize the spectral distortion by using the window function to the signal for zero at the beginning and end of the each frame. W (n) defined as the window function, where the range of n is between 0 to N-1. Here N is the length of frame. The result of windowing is the signal given by equation (1).

$$Y(n) = x(n)\, w(n),\ 0<n<N-1 \qquad (1)$$

We have considered the hamming window because the parameter side-lobe is good in that. Fig. 6 shows shape of hamming window function.

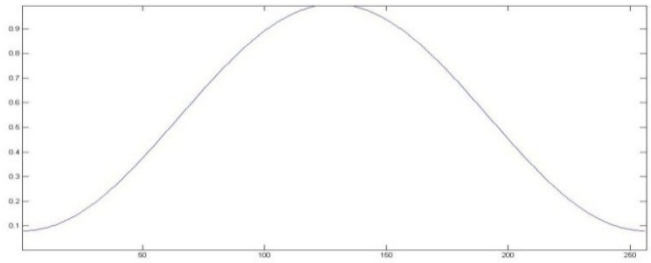

Fig. 6 Hamming window

Though there are other windows like triangle window function, rectangle window function but hamming window shows the gaussian characteristic that's why we have used the hamming window in this research work, which has the form:

$$W(n) = 0.54 - 0.46 \cos(2*PI*N / (N-1)),\ 0<n<N-1 \qquad (2)$$

Comparisons between different windows are given by Fredric J. Harris [2].

FFT (Fast Fourier Transform): - The next step is to take fast fourier transform of each frame. This transformation is a fast way of discrete fourier transform and it changes the domain from time to frequency. A Fourier transform converts time (or space) to frequency and vice versa; an FFT rapidly computes such transformations. As a result, fast fourier transforms are widely used for many applications in engineering, mathematics and science. A fast fourier transforms have been described as "the most important numerical algorithm[s] of our lifetime". In audio signal the frequency domain is effective than the time domain because frequency of particular person is

effective way to describe that person rather than the amplitude of the signal. Fig. 7 shows both the domain of analog signal.

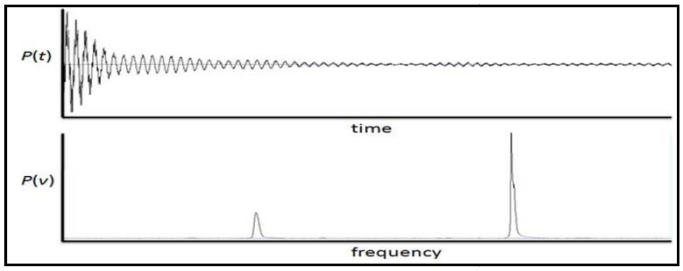

Fig. 7 Time domain and frequency domain

The equation of converting time domain to frequency domain is given by below:

$$S(f) = \int_{-\infty}^{\infty} s(t)e^{-i2\pi ft}dt \qquad (3)$$

Where *t* represents the time and *f* represents frequency.

Filtering: - The term Filtering means the process of removing or reducing the noise from the signal. The Fig.8 shows the graphical representation of noise filtering.

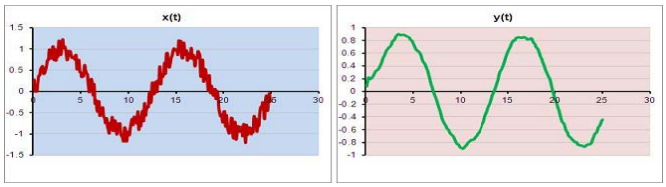

Fig. 8 Noise filtering

In our research we have consider NOIZEUS database having different variety of noise. In this step we have used wiener filter (removing noise by linear way). Numerous techniques were developed and among them wiener filter is the most fundamental approach and has been delineated in different forms and adopted in diversified applications [3]. More details about wiener filter for speech is available at reference [3].

Position of filter is chosen after FFT because we done the experiments related to the position that filtering is effective in the time domain (after framing) or in the frequency domain (after framing). We found that wiener filter is effective in frequency domain. Results we achieved are given in experiments and results section.

Mel frequency transformation: - Human ear perceives the frequencies non-linearly. Researches show that the scaling of perceiving frequency is linear up to 1 kHz and logarithmic above that. The main purpose of this is to motivate human hearing perception. This operation makes our features match more closely what humans actually hear. The mel-scale (Melody Scale) filter bank characterizes the human ear preciseness of the frequency. It is used as a band pass filtering for this stage of the identification. The whole signals for each frame is passed through the mel-scale band pass filter to mimic the human ear. Fig. 9 shows the mel-frequency scale. As mentioned above, psycho-physical studies have shown that speech signals do not follow a linear scale.

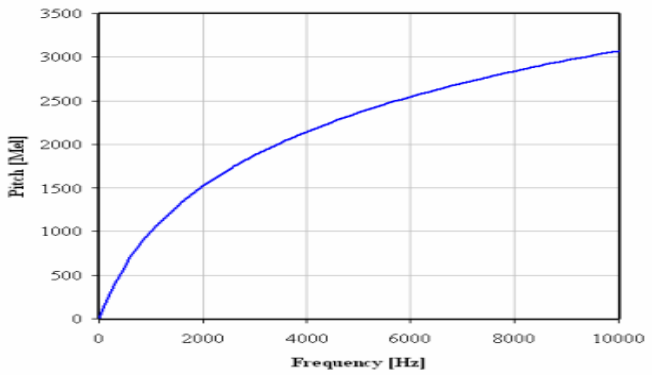

Fig. 9 Mel-frequency scale

Thus the actual frequency f, for each tone, is measured in Hz. An individual pitch is measured on a scale called the "mel-scale". Below 1000 Hz the mel-frequency scale is linear whereas above 1000 Hz is logarithmic spacing. 1 kHz tone is a pitch as reference point and the perceptual hearing threshold is above 40db, which is well-defined as 1000 Mels. Hence, the following estimated formula to calculate the Mels for a given frequency f in Hz:

$$\text{Mel}(f) = 2595 * \log_{10}(1 + f / 700) \qquad (4)$$

Mel scale filter bank is defined by a constant mel-frequency interval, having triangular band pass frequency response as well as spacing and bandwidth. The number of mel-cepstral coefficients, K, is typically chosen as 20. In the frequency domain, mel-warped filter bank is to view each of filters as a histogram bin, here bins have overlap. Fig. 10 shows the filter bank in mel-frequency scale.

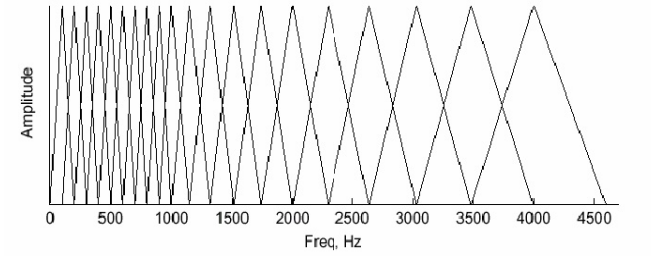

Fig. 10 Filter bank in mel-frequency scale

Classification

We have done the classification with Neural Network. One hidden layer in the neural network is required because our feature vectors have 40 attributes. We have kept the learning rate 0.3, momentum as 0.2 and training cycle as 100.

IV. EXPERIMENTS AND RESULTS

We have tested our proposed technique with the NOIZEUS database. There are 6 speakers. Each speaker speaks 5 sentences in seven different types of noise. The noises are of airport, bubble, car, exhibition, restaurant, station and street. The level of noise that is SNR (signal to noise ratio) are also types of 15dB, 10dB and 5dB. The results of our experiments are shown in the below table 1.

TABLE 1 IDENTIFICATION RATE (%) USING PROPOSED MFCC FOR NOIZEUS SPEECH CORPUS

SNR (Feature extraction technique)	Airport (%)	Bubble (%)	Car (%)	Exhibition (%)	Restaurant (%)	Station (%)	Street (%)	Average (%)
15dB (MFCC) [9]	88.33	90.00	90.67	89.00	90.00	88.33	90.00	89.47
15 dB (Existing MFCC)	100	100	100	80.00	100	96.67	93.52	95.74
15dB (proposed MFCC)	**100**	**100**	**100**	**96.67**	**100**	**100**	**100**	**99.52**
10dB (MFCC) [9]	85.67	87.67	86.00	87.33	85.33	86.67	87.67	86.62
10 dB (Existing MFCC)	86.67	96.67	93.33	83.33	100	96.67	73.33	90.00
10dB (proposed MFCC)	**93.33**	**100**	**90.00**	80.00	**100**	**100**	80.00	**91.90**
5dB (MFCC) [9]	83.00	83.67	79.67	78.33	83.33	83.00	83.33	82.04
15 dB (Existing MFCC)	86.67	66.67	56.67	73.33	83.33	66.67	53.33	69.52
5dB (proposed MFCC)	**86.67**	83.33	56.67	70.00	**86.67**	76.67	60.00	74.28

We have seven type of noise that is airport, bubble, car etc. The speeches having six types of different noise are used for training and the remaining speeches of the rest type of noise are used for testing. This set up is known as noise mismatch type of testing. The dark numbers show the accuracy we got with proposed technique which is good than the other two techniques which are mentioned in above table 1. In this experimental set up we got the average accuracy of 88.57% by our proposed MFCC which is 2.5% good than the MFCC with principal component analysis (PCA) with genetic algorithm [9] and 3.5% than the existing MFCC.

V. CONCLUSION

In this research work, we propose an approach which is a variant of a well-known technique MFCC for the noisy environment. MFCC gives the effective results in clean environment. But it drops the results in noisy environment. Using our proposed approach in noise mismatch condition with the classifier neural network, we achieved 88.57% accuracy to identify speaker in noisy environment. We analyzed that if the speech signal is more strong than noise signal means SNR ratio is greater than 0dB, then our propose technique gives good results for identification.

REFERENCES

[1] Azzam Sleit, Sami Serhan, "A Histogram Based Speaker Identification Technique", Computer Science Department - King Abdulla II School for Information Technology University of Jordan, IEEE 2008.

[2] Fredric J. Harris, Mexber, IEEE, "On then Use of Windows for Harmonic Analysis with the Discrete Fourier Transform*"in Proceeding of the IEEE,* January 1978.

[3] Jacob Benesty, Jingdong Chen, "Study of Wiener Filter for Noise Reduction" *in Springer Berlin Heidelberg*, 2005.

[4] Juhani Saastamoinen, "On Factors Affecting MFCC-Based Speaker Recognition Accuracy", Department of Computer Science, Finland.

[5] Kevin R. Farrellet. al, "Speaker Identification Using Neural Tree Networks", CAI€' Center, Rutgers University Piscataway, IEEE 1994.

[6] Longbiao Wang, Kazue Minami, "Speaker Identification By Combining Mfcc And Phase Information In Noisy Environments", Toyohashi University of Technology, Japan - 978-1-4244-4296-6/10, IEEE 2010.

[7] Martinez, J. Et. all "Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques" *in Electrical Communications and Computers (CONIELECOMP),* IEEE2012.

[8] McLoughlin I. Applied speech and audio processing (CUP, 2009)(ISBN 0521519543)-BOOK.

[9] Md. Rabiul Islam , Md. Fayzur Rahman, "Noise Robust Speaker Identification using PCA based Genetic Algorithm" *in International Journal of Computer Applications* (0975 – 8887) Volume 4– No.12, August 2010.

[10] M.G.Sumithra Et. all, "A Study on Feature Extraction Techniques for Text Independent Speaker Identification", Department of Electronics and Communication Engineering, Coimbatore, IEEE 2012.

[11] Satyanand Singh, Dr. E.G Rajan, "MFCC VQ based Speaker Recognition and Its Accuracy Affecting Factors" *in International Journal of Computer Applications*, may 2011.

[12] Sourabh Ravindran, David V.Anderson,"Improving the Noise-Robustness of Mel-Frequency Cepstral Coefficients forSpeech Processing" *in Yahoo Research*, Atlant.

[13] Mikhael, W.B et. al. "An improved speaker identification technique employing multiple representations of the linear prediction coefficients" *in IEEE International Symposium On Circuits And systems,* 2003.

[14] Zeinali, H., Sameti, H. "A fast Speaker Identification method using nearest neighbor distance" *in IEEE 11th International Conference on Signal Processing (ICSP),* 2012.

[15] Yuhuan Zhou, Jinming Wang, "Research on Adaptive Speaker Identification Based on GMM" *in International Forum on Computer Science-Technology and Applications,* 2009.

[16] Al-Dahri, S.S., Alotaibi, Y.A.,"A Word-Dependent Automatic Arabic Speaker Identification System" *in IEEE International Symposium on Signal Processing and Information Technology,* 2008.

[17] Reynolds, D.A. et. al., "Robust text-independent speaker identification using Gaussian mixture speaker models" *in IEEE Transactions on Speech and Audio Processing,* Jan 1995.

[18] Hossain, M., Ahmed, B., "A real time speaker identification using artificial neural network" *in 10th international conference on Computer and information technology,* 2007.

[19] Birnbaum, M., Brown, K.L., "Text-independent speaker identification" *in IEEE International Conference on Acoustics, Speech, and Signal Processing,* 1996.

[20] Sinith, M.S., Gowri Sankar, K., "A novel method for Text-Independent speaker identification using MFCC and GMM" *in International Conference on Audio Language and Image Processing (ICALIP),* 2010.

[21] Islam, M.R., Rahman, M.F., "Improvement of speech enhancement techniques for robust speaker identification in noise" *in 12th International Conference on Computers and Information Technology,* 2009.

[22] Sarangi, S.K., Saha, G., "A novel approach in feature level for robust text-independent speaker identification system" *in 4th International Conference on Intelligent Human Computer Interaction (IHCI),* 2012.