Volume 5, Issue 1, January 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Feature Extraction and Classification for Automatic Speaker Recognition System – A Review

Kirandeep Kaur*, Neelu Jain
E&C Department
PEC University of Technology, Chandigarh, India

Abstract*— Automatic speaker recognition (ASR) has found immense applications in the industries like banking, security, forensics etc. for its advantages such as easy implementation, more secure, more user friendly. To have a good recognition rate is a pre-requisite for any ASR system which can be achieved by making an optimal choice among the available techniques for ASR. In this paper, different techniques for the system have been discussed such as MFCC, LPCC, LPC, Wavelet decomposition for feature extraction and VQ, GMM, SVM, DTW, HMM for feature classification. All these techniques are also compared with each other to find out best suitable candidate among them. On the basis of the comparison done, MFCC has upper edge over other techniques for feature extraction as it is more consistent with human hearing. GMM comes out to be the best among classification models due to its good classification accuracy and less memory usage.

Keywords*— Automatic Speaker Recognition, Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Cepstral Coefficients, Gaussian Mixture Model (GMM), Vector Quantization (VQ), Dynamic Time Warping (DTW), Hidden Markov Model (HMM), Wavelet decomposition

I. INTRODUCTION

Automatic speaker recognition systems are categorized as two types of systems – speaker identification and speaker verification [1, 2]. In speaker verification systems, a person claims its identity and is verified against that particular person's database but on the other hand in speaker identification systems, a person is identified by comparing his attributes with each template stored in the database.

Apart from this classification, the system could be text dependent or text independent. Text dependent system involves same text being spoken both in training phase as well as in testing phase whereas there is no restriction on the text being spoken in text independent systems. Recognition rate comes out to be better in text dependent system as compared to text independent systems [2] because there is better calibration of the text used in both the phases and there is more control over the user and environment. But on the other side, we can record longer samples of voice in text independent systems which also helps in improving system performance due to large feature space. The main steps involved in any general speaker recognition system are shown in figure 1:

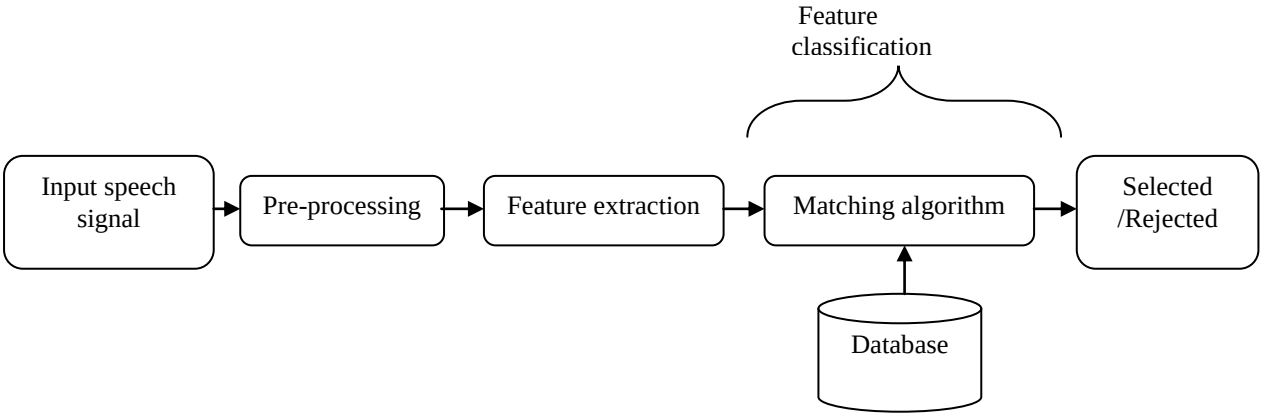

Figure 1: Process flow in ASR

The preprocessing is the first step to make data ready for feature extraction. Feature extraction reduces the dimensionality of the data and makes easy to process the speech data. Speech signal also consists of non relevant information. Therefore, important features of speech signal are extracted and are used for further processing. The widely used techniques for feature extraction are linear predictive cepstral coefficients (LPCC), perceptual linear predictive coefficients (PLPC), cepstrum analysis, linear frequency cepstral coefficients (LFCC), mel frequency cepstral coefficients (MFCC). After the features are extracted, classifiers are used to model the feature data for different users.

Whenever a person speaks its features are compared with the database of speaker models and selected/rejected depending upon match/mismatch of the models. The classifiers are based on three models : Stochastic model (e.g. Gaussian Mixture Model (GMM), Hidden Markov Model (HMM)), Deterministic model (e.g. Support Vector Machines (SVM)), Template model (Dynamic Time Warping (DTW), Neural Networks (NN), Vector Quantization (VQ)). The classifier algorithms can also be supervised or unsupervised. In supervised learning, labels assigned to data are known before whereas in unsupervised learning, labels are not known.

Rest of the paper is structured as follows. Some key concepts and literature survey is given in section 2. Section 3 explains the techniques involved in feature extraction and their comparison. Section 4 discusses and compares different classifier models. Section 5 concludes the paper.

II. LITERATURE SURVEY

The preprocessing involves frame segmentation, detection of active frames and windowing technique. If speech signal is enhanced during preprocessing then it increases the recognition rate of the system [3]. For speech enhancement, spectral subtraction and adaptive noise cancellation methods are used for single channel and multichannel speech signals respectively. Various modifications have been proposed in these conventional methods such as adaptive noise in wavelet domain, spectral subtraction with oversubtraction model, SS with adaptive averaging [3]. Empirical mode decomposition is also used for speech enhancement [4] in which the signal is decomposed into IMFs (Intrinsic mode functions) and then selected IMFs depending upon threshold value are used for reconstruction of the speech signal.

Feature extraction techniques evolved starting from the very basic techniques based on long and short term spectral averages [5], predictive coefficients (LPCC, PLPC) to widely used filterbank coefficients (MFCC, LFCC). D.A. Reynolds (1994) has compared MFCC, LFCC, LPCC, PLPC techniques with each other [6]. MFCC proves to be better than all other techniques for lower filter orders. LPCC, PLPC gives better performance with increasing filter order but performance degrades in linear coefficients (LFCC) because it gives equal detail to entire band of the signal hence highlights the superfluous information also [6].

LFCC outperforms MFCC in case of female trials [7] due to short vocal tract length or higher formant frequencies in females and MFCC gives more detail to lower frequencies only whereas LFCC gives equal detail to all the frequencies and captures the high formant frequencies of females. T. Barbu (2007) has used DDFMCC (Delta Delta Mel Frequency Cepstral Coefficients), the modified MFCCs which are second order derivatives of mel cepstral coefficients and this proposed system has produced 85% recognition rate [8]. The problem of short length data (<15s) has been solved by adding some noise to the speech signals. This increased the feature sample size and hence improved performance [9].

Performance of MFCC based system has been improved by using modified window function in MFCC technique [10]. This system represents power spectrum of the original spectrum as well as its derivative and also includes the phase information. Khaled Daqrouq et al. (2008) have designed a speaker identification system based on continuous wavelet transform method [18].

Wavelet decomposition has been used in fusion with MFCC in which MFCC features are extracted in frequency sub-bands obtained by wavelet decomposition [17]. This proposed system has shown increase in recognition rate of 1.18% over MFCC features. B.G. Nagaraja et al. (2013) have combined MFCC and LPCC features together to improve the recognition rate in multilingual speaker identification systems because both techniques extract different types of information [11].

R. Schwartz et al. (1982) has compared probability density estimation with Mahalanobis distance method [12] and has shown that former method performed substantially better than the latter. B. Wildermoth et al. (2003) have implemented GMM based classifier on MFCC based system and evaluated on three readily available databases i.e. TIMIT, YOHO and ANDOSL. Benefits and limitations of these databases have been discussed [13].

Loh Mun Yee et al. (2013) have compared different classifiers – DTW, GMM and SVM on MFCC feature vectors in which SVM produced worst results among all the three due to its restriction to work with fixed length vectors [14]. GMM-UBM performed better than DTW and HMM due to the alleviation of the problem of computational complexity and storage space [20].

III. FEATURE EXTRACTION

The widely used techniques for feature extraction can be classified into two categories based upon their underlying concept. One is filterbank coefficients (MFCC, LFCC) and the other one is predictive coefficients modeled by all pole model (LPC, LPCC). Wavelet decomposition can also be used in combination with these techniques more specifically with MFCC.

A. Linear Predictive Coefficients

In this technique, the speech production model is represented by these coefficients. The vocal tract is modeled by the all-pole filter [6]. The linear predictive coefficients are the coefficients of this all pole filter. The speech signal is represented as a linear combination of past outputs and present and past inputs.

$$s(n) = -\sum_{k=1}^{p} a(k)\, s(n-k) + Gu(n)$$

Where S_n is speech signal and U_n is an excitation signal. Here, $a_1, a_2, \ldots, a_k$ are predictive coefficients which are evaluated by autoregression method or covariance method. LPC carries a disadvantage that it is not consistent with

human hearing because it gives detail to all the frequencies equally due to which high frequency coefficients are also included which normally add noise [6].

B. Linear Predictive Cepstral Coefficients

This technique is just an extension to the above mentioned LPC technique. When linear predictive coefficients are represented in cepstrum domain then the obtained coefficients are linear predictive cepstral coefficients. Cepstrum is obtained by taking inverse DFT of logarithm of the magnitude of the DFT of the speech signal. They are more robust and reliable than LPC [6].

C. Mel-Frequency Cepstral Coefficients

Mel frequency cepstrum is a representation of a short term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a non linear mel scale of frequency. Mel frequency cepstral coefficients are the coefficients that collectively make up the MFCC. The flow process in MFCC calculation is as shown in figure 2 and the frequency response of mel scaled filterbank is shown in figure 3.

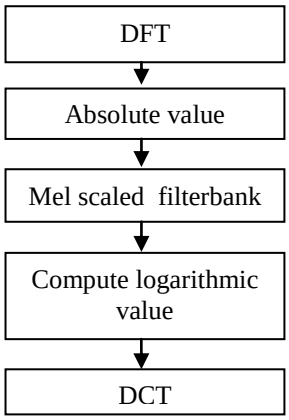

Figure 2: Flowchart for MFCC technique

For each frame of the speech signal, DFT is calculated as:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi kn/N} \quad ; N-1 \leq k \leq 0$$

After that absolute value of the powers obtained by DFT is determined and mapped by mel-scaled filterbank. Then calculating logarithmic value of these mapped power spectrums, cepstral coefficients are obtained by taking DCT as:

$$X(k) = \sum_{n=0}^{N-1} x(n) \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right] \quad ; N-1 \leq k \leq 0$$

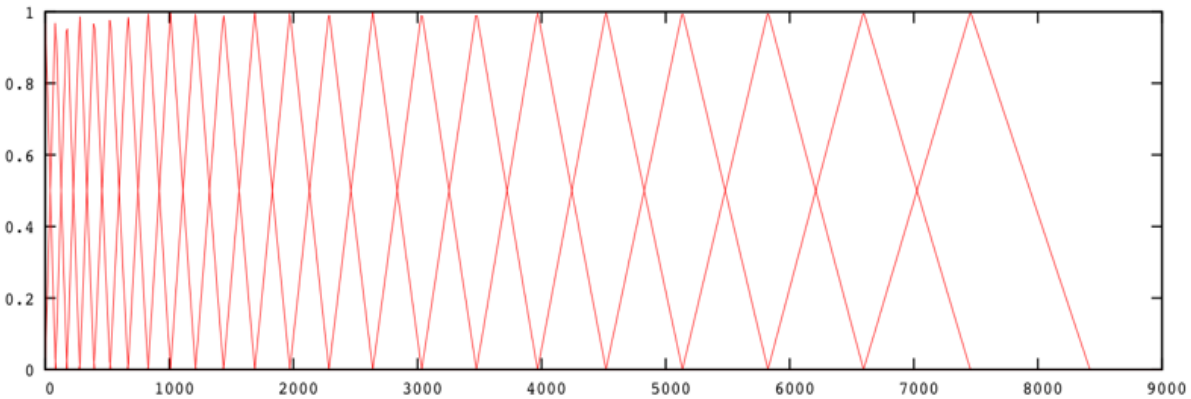

Figure 3: Frequency response of mel scaled filterbank

With this filterbank less detail is given to higher frequencies and more detail is given to lower frequencies because speech characteristics are supposed to be present more at lower frequencies. The relationship between mel scale and hertz scale is represented as:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

MFCC technique is considered more consistent with human hearing as compared to LPCC, LFCC because of mel scale representation [15].

D. Wavelet decomposition

In this method the speech signal is decomposed into different frequency sub-bands. Decomposition levels can be different. At each level, the signal is partitioned into low frequency (approximate coefficients) and high frequency bands

(detail coefficients). Approximate coefficients at each level can be used for further decomposition. The frequency resolution keeps on increasing with increasing number of decomposition levels. The advantage of this technique is that only speech carrying frequency bands i.e. low frequencies can be used for further processing and can be used in fusion with MFCC by extracting features from selected frequency sub –bands [15].

TABLE I COMPARISON OF DIFFERENT FEATURE EXTRACTION TECHNIQUES

S.No.	Technique	Principle	Merits and De-merits
1.	LPC	Modeled by all pole model	Based on basic principle of sound production, performance degradation in presence of noise [15].
2.	Cepstral coefficients	FFT based	Not much consistent with human hearing due to representation by linearly spaced filters [16].
3.	LPCC	Modeled by all pole model	Gives smoother spectral envelope and stable representation as compared to LPC [15], drawback due to linearly spaced frequency bands [16].
4.	MFCC	Filterbank coefficients	More information about lower frequencies than higher frequencies due to mel spaced filter banks hence behaves more like a human ear as compared to other techniques [15,17], based on STFT which has fixed time-frequency resolution [17].
5.	Wavelet decomposition	Decomposition to sub-bands	Characteristics of time-frequency localization and the multi-resolution analysis are suitable for non-stationary speech signal [17], Continuous wavelet transform is preferable over discrete wavelet transform as it gives more data per scale [18].

IV. FEATURE CLASSIFICATION

Using feature extraction, a spoken utterance can be represented with feature vectors. The speech signal of a person will have similar still differently arranged feature vectors. To recognize these feature vectors voice modeling is done using classifier algorithms by which a template of features is generated for a particular registered user and that is used as a reference in recognition process. It means every registered user will have a reference model in database and if a new user comes then it will be declared as unregistered one. The different classifiers are discussed and compared here:

A. Vector quantization

A large set of feature vectors are divided into groups having approximately the same number of points closest to them. Each group is represented by its centroid point. VQ can be defined as a mapping function that maps k-dimensional vector space to a finite set CB = {C_1, C_2, C_3, ……, C_N}. The set CB is called codebook consisting of N number of code vectors and each code vector C_i = {c_{i1}, c_{i2}, c_{i3},……, c_{ik}} is of dimension k. The method most commonly used to generate codebook is the Linde-Buzo-Gray (LBG) algorithm. Feature vectors are extracted from input speech signal and the Euclidean distance between input speech signal and each code vector is calculated. The input vector belongs to the cluster of the code vector that yields the minimum distance [21].

B. Gaussian mixture model

The GMM is a density estimator. The distribution of the feature vector x is modeled clearly using a mixture of M Gaussians. Expectation maximization algorithm is used to estimate mean, covariance parameters. During recognition, a sequence of features are extracted from the input signal. Then the distance of the given sequence from the model is obtained by computing the log likelihood of given sequence. The model that provides the highest likelihood score is verified as the identity of the speaker [21].

C. Support Vector Machines

Support Vector Machine is a supervised learning algorithm. It needs training of the tool before classification procedure gets started. This is the best tool for binary classification of the data. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the input.The hyperplane is constructed defined by set of weights W, data points X and a bias or offset b, such that:

$$W.X + b = 0$$

where W.X denotes the dot product of the data and the normal vector to the hyperplane. The parameter b determines the offset of the hyperplane from the origin along the normal vector. Figure 2 shows the partition of the input data into two classes.

Points lying on the hyperplane satisfy the equation (1):

$$W.X + b = 0 \quad (1)$$

Points lying on one side of this hyperplane are denoted by class C1 as positive examples satisfying:

$$W.X + b > 0 \text{ , } d(i) = +1$$

Points lying on the other side of this hyperplane are denoted by class C2 as negative examples satisfying:

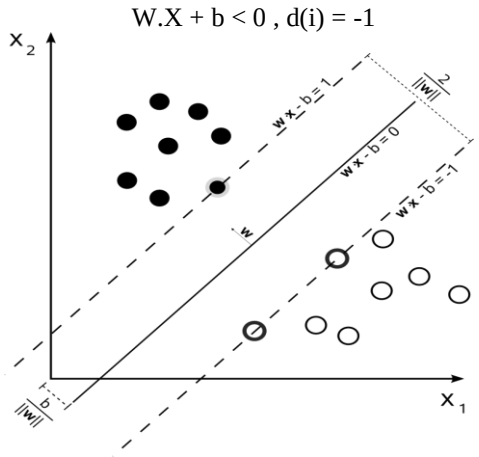

Figure 4: Hyperplane dividing input data into two classes

SVMs do not perform good for speech signals due to its restriction to work with fixed length vectors [19].

D. Dynamic Time Warping

This is used specifically to deal with variance in speaking rate and variable length of input vectors because this algorithm calculates the similarity between two sequences which may vary in time or speed. To normalize the timing differences between test utterance and the reference template, time warping is done non-linearly in time dimension. After time normalization, a time normalized distance is calculated between the patterns. The speaker with minimum time normalized distance is identified as authentic speaker [21]

E. Hidden Markov Model

HMM is a stochastic model. The elements of HMM are number of states, number of distinct observation symbols, the probability of going from one state to another, the observation symbol probability distribution and probability of being in a particular state initially. It assumes that the observation at some time is generated by some process whose state is hidden from the observer. It also follows markov property that the current state S_t does not depend upon any state prior to time t-1 and depends upon only on S_{t-1} state [22]. The speech training features are represented by probability measures which train the HMM speaker model. For each speaker model, the metric is determined as the probability of the observation sequence. The HMM speaker model which yields the highest probability is selected [21].

TABLE II COMPARISON OF DIFFERENT CLASSIFIERS

S.No.	Classifier	Type of algorithm	Merits and De-merits
1.	DTW	Unsupervised	Problem in dealing with cross-channel issue, requires less storage space [20], beneficial for variable length input features [21].
2.	HMM	Unsupervised	Computationally more complex and needs more storage space, needs more training data to deal with inter-session issue [20].
3.	GMM	Unsupervised	Needs less training and test data, compromise between DTW and HMM [20].
4.	SVM	Supervised	Beneficial in case of binary classification, poor performance in speaker recognition due to its restriction to work with fixed length vectors [14].
5.	NN	Unsupervised	Requires storage of huge amount of training data, extensive computation to find the closest neighbour [21].
6.	VQ	Unsupervised	Memory requirement is feasible for real-time applications, Computationally less complex [21].

V. CONCLUSIONS

This paper has reviewed the research done in the area of automatic speaker recognition. Different techniques for feature extraction and classification have been discussed. Each technique has got its advantages and limitations as shown in the comparison tables. Some techniques are preferred over others such as MFCC for feature extraction and GMM for classification. MFCC is more consistent with human hearing due to mel scale representation. MFCC can be integrated with other techniques such as wavelet decomposition and LPCC to improve the performance of the system as by integrating features more information is added to the input training data. Otherwise choice can be made depending upon

certain parameters such as number of system users, storage space, classification time etc. VQ is preferred for real time systems because of its less memory requirement. To align the input features in time dimension, DTW is used. For large number of users, GMM performs better as it requires less amount of data to train the classifier hence memory usage also decreases for the system.

REFERENCES

[1] J.P. Campbell, "*Speaker Recognition: A Tutorial*", Proceedings of the IEEE, vol.85, issue-9, pp. 1437-1462, September, 1997.

[2] G.R. Doddington, "*Speaker Recognition – Identifying People by their Voices*", Proceedings of the IEEE, vol. 73, issue-11, pp. 1651-1664, November, 1985.

[3] P. Bactor, A. Garg, "*Different Techniques for the Enhancement of the Intelligibility of a Speech Signal*", Inernational Journal of Engineering Research and Development, vol. 2, issue-2, pp. 57-64, July, 2012.

[4] L. Zao, R.Coelho, P. Flandrin, "*Speech Enhancement with EMD and Hurst-Based Mode Selection*", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, issue-5, pp. 899-911, May, 2014.

[5] R. E. Wohiford, E. H. Wrench, Jr., and B. P. Landell, "*A Comparison of Four Techniques for Automatic Speaker Recognition*", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol.5, pp. 908-911, April, 1980.

[6] D.A. Reynolds, "*Experimental Evaluation of Features for Robust Speaker Identification*", IEEE Trans. on Speech and Audio Processing, vol. 2, issue-4, pp. 639-643, October, 1994.

[7] X. Zhou, D. G. Romero, R. Duraiswami, C.E. Wilson, S. Shamma, "*Linear versus Mel Frequency Cepstral Coefficients for Speaker Recognition*", IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Waikoloa, HI, pp. 559 – 564, 11-15 December, 2011.

[8] T. Barbu, "*A Supervised Text-independent Speaker Recognition Approach*", Proceedings of World Academy of Science, Engineering and Technology, International Journal of Computer, Information, Systems and Control Engineering, vol.1, issue-9, pp. 2678-2682, January, 2007.

[9] P. Krishnamoorthy, H.S. Jayanna , S.R.M. Prasanna, "*Speaker Recognition under Limited Data Condition by Noise Addition*", Expert Systems with Applications (Elsevier), vol.38, issue-10, pp.13487–13490, September, 2011.

[10] Md Sahidullah, G.Saha, "*A Novel Windowing Technique for Efficient Computation of MFCC for Speaker Recognition*", IEEE Signal Processing Letters, vol.20, issue-2, pp. 149-152, February, 2013.

[11] B.G. Nagaraja, H.S. Jayanna, "*Combination of Features for Multilingual Speaker Identification with the Constraint of Limited Data*", International Journal of Computer Applications, vol.70, issue-6, pp. 1-6, May, 2013.

[12] R. Schwartz, S. Roucos, and M. Berouti, "*The Application of Probability Density Estimation of Text-Independent Speaker Identification*," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol.7, pp. 1649-1652, May, 1982.

[13] B. Wildermoth, K.K. Paliwal, "*GMM Based Speaker Recognition on Readily Available Databases*", Proceedings of the Microelectronic Engineering Research Conference, Brisbane, Australia, November, 2003.

[14] http://comp.utm.my/pars/files/2013/04/Comparative-Study-of-Speaker-Recognition-Methods-DTW-GMM-and-SVM.pdf .

[15] S. Malik, F. A. Afsar, "*Wavelet Transform based Automatic Speaker Recognition*", IEEE 13th International Multitopic Conference, INMIC, Islamabad, pp. 1-4, 14-15 December, 2009.

[16] S. Tripathi, S. Bhatnagar. "*Speaker Recognition*", IEEE Third International Conference on Computer and Communication Technology (ICCCT), Allahabad, pp. 283-287, 23-25 November, 2012.

[17] P. Kumar, M. Chandra, "*Hybrid of Wavelet and MFCC Features for Speaker Verification*", IEEE World Congress on Information and Communication Technologies (WICT), Mumbai, pp. 1150-1154, 11-14 December, 2011.

[18] K. Daqrouq, W. A. Sawalmeh, A. R. Qawasmi, I. N. Isbeih, "*Speaker Identification Wavelet Transform based Method*", IEEE 5th International Multi-Conference on Systems, Signals and Devices (SSD), Amman, pp. 1-5, 20-22 July, 2008.

[19] C. J. C. Burges, "*A Tutorial on Support Vector Machines for Pattern Recognition*", Data Mining and Knowledge Discovery (Springer), vol. 2, issue-2, pp. 121-167, June, 1998.

[20] WW. Chen, Q. Hong, X. Li, "*GMM-UBM for Text-Dependent Speaker Recognition*", IEEE International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, pp. 432-435, 16-18 July, 2012.

[21] R. P. Ramachandran, K.R. Farrell, R. Ramachandran, R. J. Mammone, "*Speaker Recognition—General Classifier Approaches and Data Fusion Methods*", Pattern Recognition in Information Systems, vol. 35, issue-12, pp. 2801-2821, December, 2002.

[22] Ghahramani Z., "*An Introduction to Hidden Markov Models and Bayesian Networks*", International Journal of Pattern Recognition and Artificial Intelligence, vol. 5, issue-1, pp. 9-42, 2001.

