Journal of Computer Science 7 (4): 459-465, 2011
ISSN 1549-3636

A New Speaker Recognition System with Combined Feature Extraction Techniques

[1]M.G. Sumithra, [2]K. Thanuskodi and [3]A. Helen Jenifer Archana
[1]Department of Electronics and Communication Engineering,
Bannari Amman Inst. of Technology, Sathyamangalam,
[2]Department of Electrical and Electronics,
Engineering Akshaya College of Engineering and Technology, Coimbatore
[3]Department of Electronics and Communication Engineering,
Bannari Amman Institute of Technology, Sathyamangalam

Abstract: Problem statement: This study introduces a new method for speaker verification system by fusing two different feature extraction methods to improve the recognition accuracy and security. **Approach:** The proposed system uses Mel frequency cepstral coefficients for speaker identification and Modified MFCC for verification. For speaker modeling vector quantization is used. **Results:** The proposed system was investigated the effect of the different length segmental feature as well as speaker modeling for speaker recognition. The performance was evaluated against 1000 speakers for 10 different languages with duration of 10 sec for training the system and for testing 5 sec. duration samples were used. **Conclusion/Recommendations:** Experimental results of the proposed system showed that higher recognition accuracy of 93% is achieved by increasing the number of filter banks used for feature extraction method, more competitive with existing system using vector quantization with lesser computational complexity. The system efficiency may further be improved using other speaker modeling techniques like GMM, HMM.

Key words: Feature extraction, speaker modeling, vector quantization, false acceptance, false rejection

INTRODUCTION

Speaker recognition is the task of recognizing people from their voices. Strictly speaking there is a difference between speaker recognition (recognizing who is speaking) and speech recognition (recognizing what is being said). Speaker recognition system is categorized into speaker verification (to authenticate a claimed speaker identity from a voice signal based on speaker-specific characteristics reflected in spoken words) and speaker identification (to find the identity of a talker, in a known population of talkers, using the speech input).Speaker identification is the task of determining an unknown speaker's identity. In a sense speaker verification is a 1:1 match where one speaker's voice is matched to one template (and possibly a general world template) whereas speaker identification is a 1: N match where the voice is matched to N templates.

A speaker verification system can be text dependent or text-independent. Examples of former case are user specific pass-phase or a system prompted phrase (sometimes used as a liveness test). The prior knowledge and constraint of the text can greatly boost performance of a verification system. In a text-independent application, there is no prior knowledge by the system of the text to be spoken, such as when using extemporaneous speech. The general approach to Automatic speaker verification consists of five steps: digital speech data acquisition, feature extraction, pattern matching, making an accept/reject decision and enrolment to generate speaker reference models.

A speaker verification system is composed of two distinct phases, a training phase and a test phase. Each of them can be seen as a succession of independent modules. The first and foremost module is the feature extraction module conveying speaker information extracted from the speech. This is the pedestal module, where the entire system performance relies. The next module is speaker modeling module, represent that speaker’s voice and acoustic features. The selection of modeling is primarily dependent on the type of speech to be used, desired performance, the ease of training and updating and storage and computation

Corresponding Author: Sumithra M.G., Department of Electronics and Communication Engineering,
Bannari Amman Inst. of Technology, Sathyamangalam, Tamil Nadu, India

considerations. The final module is for making decision based on the training and testing phase.The system, in turn, outputs a binary decision: Either accept or reject the authenticity for the claimed speaker. Success in speaker verification depends on extracting and modeling the speaker dependent characteristics of the speech signal which can effectively distinguish one talker from another.

Literature review: The most commonly used acoustic vectors are Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC) and Perceptual Linear Prediction Cepstral (PLPC) Coefficients and zero crossing coefficients (Yegnanarayana *et al*., 2005; Vogt *et al*., 2005). All these features are based on the spectral information derived from a short time windowed segment of speech. They differ mainly in the detail of the power spectrum representation. A new modification of Mel-Frequency Cepstral Coefficient (MFCC) feature has been proposed for extraction of speech features for Speaker verification (SV) application (Saha and Yadhunandan, 2000).This is compared with original MFCC based feature extraction method and also on one of the recent modification. The study uses multi-dimensional F-ratio as performance measure in Speaker Recognition (SR) applications to compare discriminative ability of different multi parameter methods.An MFCC like feature based on the Bark scale is shown to yield similar performance in speech recognition experiments as MFCC (Aronowitz *et al*., 2005).The BFCC features perform well for text dependent speaker verification systems. Revised perceptual linear prediction was proposed by Kumar *et al*. (2010), Ming *et al*. (2007) for the purpose of identifying the spoken language; Revised Perceptual Linear Prediction Coefficients (RPLP) was obtained from combination of MFCC and PLP.

The objective of modeling technique is to generate speaker models using speaker-specific feature vectors. Such models will have enhanced speaker-specific information at reduced data rate. This is achieved by exploiting the working principles of the modeling techniques. Earlier studies on speaker recognition used direct template matching between training and testing data. In the direct template matching, training and testing feature vectors are directly compared using similarity measure. For the similarity measure, any of the techniques like spectral or Euclidean distance or Mahalanobis distance is used (Liu *et al*., 2006).

The disadvantage of template matching is that it is time consuming, as the number of feature vectors increases. For this reason, it is common to reduce the number of training feature vectors by some modeling technique like clustering. The cluster centres are known as code vectors and the set of code vectors is known as codebook. The most well-known codebook generation algorithm is the K-means algorithm (Mporas *et al*., 2007; Ming *et al*., 2007). In 1985, Soong *et al*. used the LBG algorithm for generating speaker-based vector quantization (VQ) codebooks for speaker recognition. In order to model the statistical variations, the hidden Markov model (HMM) for text-dependent speaker recognition was studied. The system performances in neural network based networks were also studied (Clarkson *et al*., 2006). In HMM, time-dependent parameters are observation symbols. Observation symbols are created by VQ codebook labels. Continuous probability measures are created using Gaussian mixtures models (GMMs) (Krause and Gazit, 2006). The main assumption of HMM is that the current state depends on the previous state.

In 1995, Reynolds proposed Gaussian mixture modeling (GMM) classifier for speaker recognition task (Krause and Gazit, 2006; Clarkson *et al*., 2006). This is the most widely used probabilistic technique in speaker recognition. The GMM needs sufficient data to model the speaker and hence good performance. In the GMM modeling technique, the distribution of feature vectors is modelled by the parameters mean, covariance and weight.GMM outperformed the other modeling techniques. The disadvantage of GMM is that it requires sufficient data to model the speaker well (Aronowitz *et al*., 2005).

MATERIALS AND METHODS

For designing an efficient speaker verification system,the system has to identify the speaker from the trained database if he/she is an enrolled speaker or not.After identifying the identity of the speaker, the system verifies the speaker. Identifying a single user among N users in the databse requires a much efficient features. Then only the speaker model can further enhance the speaker specific information. MFCC features helps to cluster the speaker efficiently that improves the identification rate, because the inter speaker variabilty is high. But the intra speaker variabilty is also high, yields poor verification rate. At the same time the intra speaker variability is low in MMFCC features due to the normalized magnitude spread of the extracted coefficients from the speech signal.

Therefore to improve the speaker verification system performance for identifying the speaker's entry the proposed method uses LBG speaker model,which uses MFCC features for code book generation.

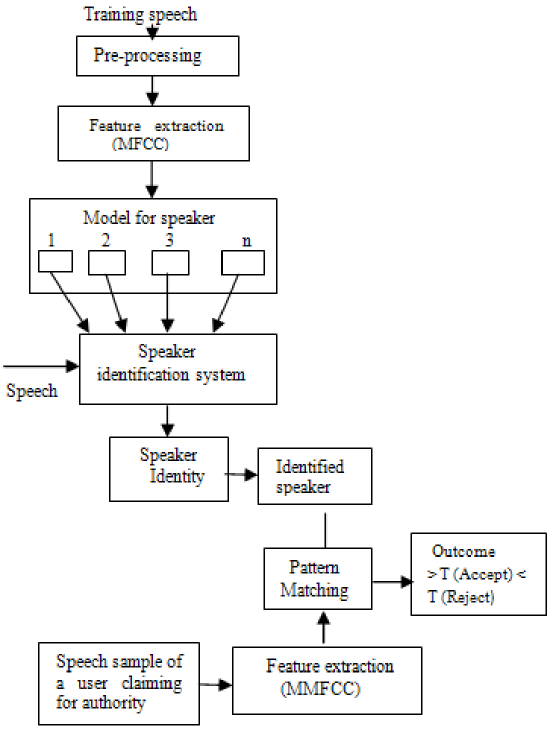
Training speech
Pre-processing
Feature extraction (MFCC)
Model for speaker
1
2
3
n
Speech
Speaker identification system
Speaker Identity
Identified speaker
Pattern Matching
Outcome >T (Accept) < T (Reject)
Speech sample of a user claiming for authority
Feature extraction (MMFCC)

Fig. 1: Proposed speaker recognition system block diagram

For verifying the speaker's authority, the system uses MMFCC features. By combining these two qualities of MFCC and MMFCC the system efficiency increases.The proposed method effectively clusters the speaker's identity due to the high inter and low intra speaker variabilty among the extracted features.This makes the clustering process easy and efficient even with low amount of training data. The proposed system performed well,even in multiple language entry database. The false rejection rate is zero for proposed mentod .Thus the proposed system is suitable for highly secured environment. Figure1 represents detailed the block diagram of the proposed system.

The terms used in the proposed scheme are as follows.

Pre-processing: The pre-processing stage convert the analog speech signal into digital samples and then segment the continuous speech signal into shorter length frames .After segmenting windowing techniques are used.

Framing: The speech signal is divided into short fixed length frames. The continuous speech signal is divided into frames where each frame consists of N samples and successive frames are overlapping with each other by M samples (Jayanna and Prasanna, 2009).

Windowing: After frame segmentation, windowing is carried out to minimize the spectral distortion by using the window to taper the signal on both ends thus reducing the side effects caused by signal discontinuity at the beginning and at the end due to framing. We have used Hamming window which is multiplied with each frame:

$$w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{N-1})$$

where N is the number of samples in each frame.

Feature extraction technique: In this proposed method feature extraction based on MFCC is used for speaker identification and MMFCC is used for Speaker verification. Both techniques were discussed below.

Mel-frequency cepstral coefficients: Mel-frequency cepstral coefficients (MFCC) are one of the most popular methods for extracting features from the speech signal. MFCC's are shown to be less susceptible to the variation of the speaker's voice and surrounding environment. It is based on the known variations of human ears. Group neighbouring frequency bins into overlapping triangular bands with equal bandwidth according to theme scale. Critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech.

A Mel is a unit of measure of perceived pitch or frequency of a tone. The characteristics is expressed on the mel frequency scale, which is a linear frequency spacing below 1KHz and a logarithmic spacing above 1KHz. The following function transforms real (linear frequency) to Mel frequency:

$$\text{Mel}(f) = 2595\log(1 + \frac{f}{700})$$

Modified Mel-frequency cepstral coefficients: A modified Mel-frequency cepstral coefficient (MMFCC) is the improvised version of conventional MFCC. The weightining function was introduced which is unique for each frame of an utterance for each speaker (Clarkson *et al*., 2006; Aronowitz *et al*., 2005). From the 20 filter bank outputs, we have to calculate the average. Next, the city block distance of each filter bank output from the average and then sum the log of the distance for all filters were calculated. It is called 'Sweep' that is unique for each frame (window) since it is calculated from filter bank output of that frame. The sweep represents the total variation in magnitude of the filter outputs for each frame and gives a measure of the magnitude spread of the coefficients, equivalent to variance in Euclidean distance measure:

$$\text{avg} = \frac{1}{20}\sum_{i=1}^{20} S_k[i]$$

where $S_k[i]$ be the filter bank outputs, where i = 1, 2, ...20 and:

$$M_k[i] = \left|S_k[i] - \text{avg}\right|$$

Therefore:

$$\text{Sweep} = \sum_{i=1}^{20} \log M_k[i]$$

So finally weighting function is defined as:

$$W[i] = \log\left[\frac{S_k[i]}{\text{Sweep}}\right]$$

The modification in above through the weighting function gives the Modified MFCC coefficients as:

$$C_n = \left(\log\left(S_k[i].W[i].\cos\left(n\left(i - \frac{1}{2}\right)\frac{\pi}{2}\right)\right)\right)$$

MMFCC uses compensation based on the magnitude of spread, through a frame based weighting function to preserve the speaker dependent information in different frames. The variation of intensity/loudness at different segments of a spoken word may influence the magnitude of the coefficients affecting cluster formation in parameter space for a speaker. MMFCC is a frame based technique to reduce these effects through normalization of coefficients in each frame by its total spread, so that coefficients of all the frames are brought to same level of spread. This also minimizes effect of change in background noise level in SR applications where the speaker while speaking is moving from one environment to another. MMFCC features shows enhanced discriminative ability for the coefficients that is important in Speaker Verification applications.

Speaker modeling technique:

Vector quantization: Vector Quantization is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a codeword. The collection of all code words is called a codebook. The training material is used to estimate the code book. Here a speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors. The codebook is a set of cells in a multidimensional space. Each cell defines a small part of the total space and contains a point centered within the cell, the centroid (Goto *et al*., 2008). The cepstral coefficients derived from each frame are regarded as a vector (a point) in the space and thereby belonging to one of the cells. The vector always belongs to the cell containing the closest located centroid. Hence, a vector quantizer Q of dimension k and size N is a mapping from a vector in the k-dimensional space into one of N centroids in the space.

K-means clustering: This is an algorithm to classify or to group data vectors based on attributes/features into K groups (or clusters). The K-means algorithm (Jayanna and Prasanna, 2009; Memon *et al*., 2009) was developed for the Vector Quantization codebook generation. It represents each cluster by the mean of the cluster centroid vector. The grouping of data is done by minimizing the sum of squares of distances between the data vectors and the corresponding cluster's centroids (Jayanna and Prasanna, 2009).

Linde-Buzo-Gray (LBG) clustering: The LBG (Memon *et al*., 2009; Alsulaiman *et al*., 2010) algorithm is a finite sequence of steps in which, at every step, a new quantizer, with an average distortion less or equal to the previous one, is produced. We can distinguish two phases, the initialization of the codebook and its optimization. The codebook optimization starts from an initial codebook and after some iteration, generates a final codebook with a distortion corresponding to a local minimum.

Feature matching: Initially the speech signal of the unknown speaker is acquired and the Mel frequency wrapping is done as explained earlier. This yields the feature vectors. These feature vectors of the unknown speaker are combined together to form the feature matrix. The feature matrix thus formed is compared with the vector quantized code book matrices present in the stored data base. This comparison is performed using Euclidean Distance(ED) calculation.

The performance of a speaker verification system is measured in terms of false acceptance rate (FA %) and false rejection rate (FR %). False acceptance error consists of accepting identity claim from an imposter (Goto *et al*., 2008). False rejection error happens when a valid identity claim is rejected. It is represented as:

$$T_E = F_A + F_R$$

where:

F_A = False acceptance
F_R = False rejection
T_E = Total error of verification system

RESULTS AND DISCUSSION

For the enrolment of the user a data record has to be maintained in the database with different text information. This database contains 10 different language entries. The amount of speech given for training and testing the speaker verification system is 10 and 5 sec respectively, to analyze the speech signal hamming window is used. The proposed speaker recognition system was analysed by varying the number of filter banks for extracting the features, length of the frames with different percentages of overlaps.

Figure 2-3 shows the acoustic vectors distribution of a user for different samples using MFCC and MMFCC. The feature vector space for MMFCC has lower variability among the different samples. But measure up to MMFCC, MFCC features has higher variability among the different speakers which is used in identification, shown in Fig. 4a and b.

The efficiency of the speaker modeling also relies on the extracted features for generating the code book. Most of the speech information is positioned in lower frequency than the higher frequencies. The higher frequency components may or may not contain the information related to speaker. Only 13 lower filter bank coefficients were taken into account for the code book generation. Usually the first filter bank component will be omitted for mapping the speaker model .This improves the speaker verification system by improving the high inter variability among the speaker models.

The proposed system efficiency were analysed by using 20 filter banks for extracting features. From the Fig. 5 we can understand that while increasing the length of the frame there were loses in the features and this makes creating speaker specific model a difficult one. At the same time increasing shift over the frame increases the redundancy. Compared to K-means LBG performs better for large segment frames with 50% shift over the frames. The shorter length features with 60% shift performs well irrespective of speaker modeling with the efficiency of 89%.

By increasing the number of filter banks we can extract the detailed features from the speech signal. Compared to K-means LBG performs better while using 24 filter banks for feature extraction.

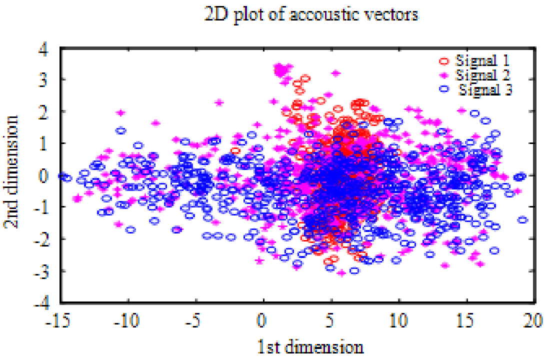

Fig. 2: 2D acoustic vectors distribution of a user for different samples using MFCC

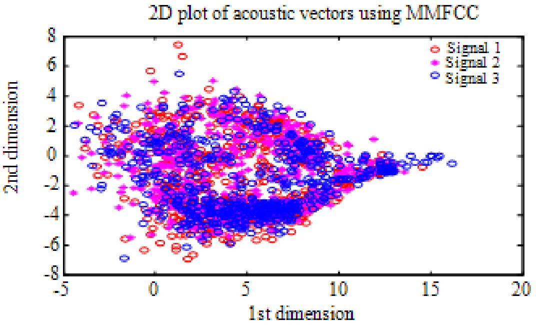

Fig. 3: 2D acoustic vectors distribution of a user for different samples using MMFCC

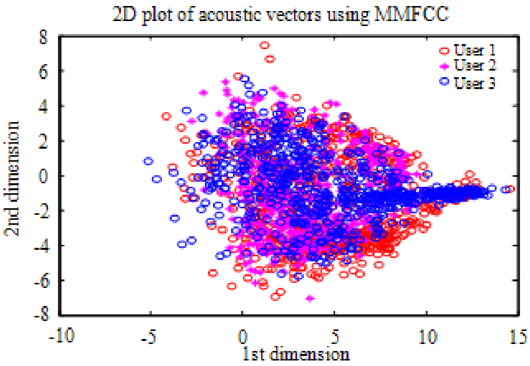

Fig. 4a: 2D acoustic vectors distribution of different users using MMFCC

The LBG algorithm models each speaker efficiently with enhanced speaker specific information at reduced data rate. As a result the system efficiently distinguishes the speaker. From Fig. 4a with 60% overlap, the 256 and 512 sample frame performed better compared to 1024 and 2048 frames.

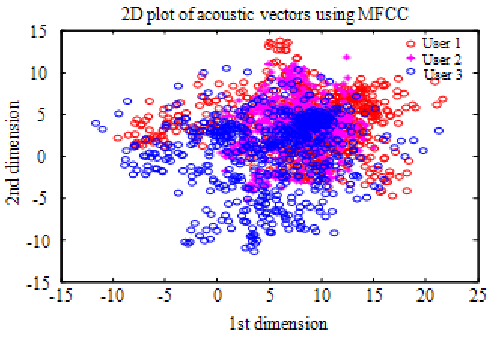

Fig. 4b: 2D acoustic vectors distribution of different users using MFCC

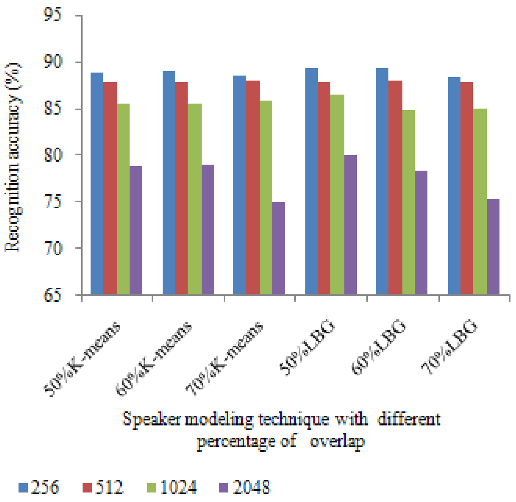

Fig. 5: Proposed speaker recognition system accuracy using 20 filter bank for different segmental length features

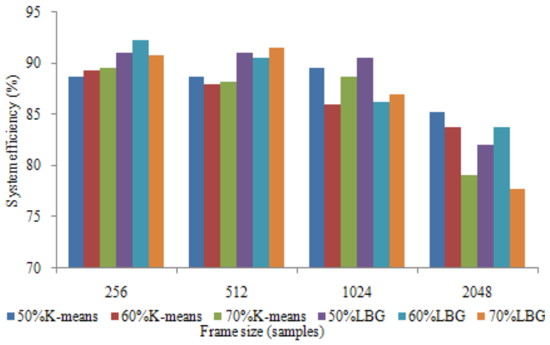

Fig. 6: Proposed speaker recognition system accuracy using 24 filter bank for different segmental length features

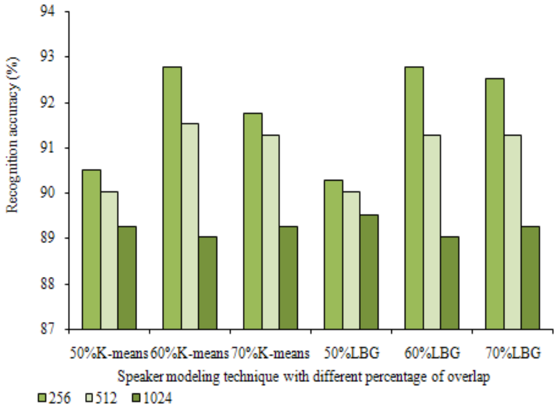

Fig. 7: Proposed speaker recognition system accuracy using 40 filter bank for different segmental length features

Even though the 256 and 512 size frames performed equally well in extracting features the system performance rely on speaker modeling module. LBG algorithm performs better than K-means for 256 length frames. At the same time K-means performed comparatively well than LBG for large frame segments.

Further the proposed speaker recognition system was evaluated using 40 filter banks for extracting the features, From the 40 filter banks lower 13 filter bank outputs (40-13) were taken for generating the speaker model. The speaker recognition system performs well under 60% shift over the segment both in LBG and K-means modeling shown in Fig. 7. But by considering the overall system performance LBG algorithm performs better than K-means even in large segmental frames.

Proposed and existing method: The database contains 10 different languages and there may be chances in existences of phonotical similarities among the languages. The efficiency of the system relies on how well the modeling technique distinguishes the extracted feature during codebook generation. From Fig. 8 for Arabic, Korean, Polish, Portuguese and English language speakers were recognized well than the existing speaker recognition system (MFCC were used for both identification and verification). The system efficiency is high for the languages having minimum number of users compared to other languages in database. While increasing the population of the database obviously the system efficiency degrades.

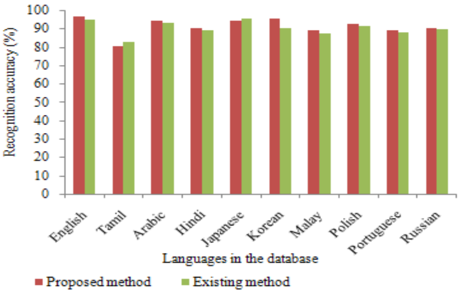

Fig. 8: Comparison of recognition accuracy for Proposed and existing method

CONCLUSION

A new speaker recognition scheme is proposed and the proposed system uses MFCC features for identification and MMFCC features for verification and LBG algorithm for speaker modeling. The proposed scheme is evaluated against the database containing 1000 speakers. The proposed system is suitable for highly secured environments, because of zero false rejection rate. Even with this high population the system performed well since it has produced comparatively good performance than the existing algorithms. The proposed system efficiency may further be improved by using other speaker modeling techniques like HMM, GMM. The proposed system can be extended to multilingual text independent speaker recognition system.

REFERENCES

Aronowitz, H., D. Irony and D. Burshtein, 2005 Modeling Intra-Speaker Variability for Speaker Recognition. Tel-Aviv University, Israel. http://eprints.pascal-network.org/archive/00001688/01/isis12.pdf

Alsulaiman, M., Y. Alotaibi, M. Ghulam, M.A. Bencherif and A. Mahmoud, 2010. Arabic speaker recognition: Babylon levantine subset case study. J. Comput. Sci., 6: 381-385. DOI: 10.3844/jcssp.2010.381.385

Clarkson, T.G., C.C. Christodoulou, Y. Guan, D. Gorse and D.A. Romano-Critchley *et al*., 2006. Speaker identification for security systems using reinforcement-trained pRAM neural network architectures. IEEE Trans. Syst. Man Cybernet., 31: 65-76. DOI: 10.1109/5326.923269

Goto, Y., T. Akatsu, M. Katoh, T. Kosaka and M. Kohda, 2008. An Investigation on Speaker Vector-Based Speaker Identification Under Noisy Conditions. Proceedings of the International Conference on Audio, Language and Image Processing, July 7-9, IEEE Xplore, Shanghai, pp: 1430-1435. DOI: 10.1109/ICALIP.2008.4590119

Jayanna, H.S. and S.R.M. Prasanna, 2009. Analysis, feature extraction, modeling and testing techniques for speaker recognition. IETE Technical Rev., 26: 181-190. DOI: 10.4103/0256-4602.50702

Kumar, P., A.N. Astik Biswas and M. Chandra, 2010. Spoken Language identification using hybrid feature extraction methods. J. Telecommun., 1: 11-5.

Krause, N. and R. Gazit, 2006. SVM-based Speaker Classification in the GMM Models Space, Proceedings of the IEEE Odyssey Speaker and Language Recognition Workshop, June 28-30, IEEE Xplore, San Juan, pp: 1-5. DOI: 10.1109/ODYSSEY.2006.248138

Liu, M., B. Dai, Y. Xie and Z. Yao, 2006. Improved GMM-UBM/SVM for Speaker Verification. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 14-19, IEEE Xplore, Toulouse, pp: 1-1. DOI: 10.1109/ICASSP.2006.1660173

Mporas, I., T. Ganchev, M. Siafarikas and N. Fakotakis, 2007. Comparison of speech features on the speech recognition task. J. Comput. Sci., 3: 608-616. DOI: 10.3844/jcssp.2007.608.616

Ming, J., T. Hazen, J. Glass and D. Reynolds, 2007. Robust speaker recognition in noisy conditions. IEEE Trans. Audio Speech Language Proc., 15: 1711-1723. DOI: 10.1109/TASL.2007.899278

Memon, S., M. Lech and N. Maddage, 2009. Speaker verification based on different vector quantization techniques with gaussian mixture models. Proceedings of the 3rd International Conference on Network and System Security, Oct. 19-21, Gold Coast, Queensland, Australia, pp: 403-408.

Saha, G. and U.S. Yadhunandan, 2000. Modified Mel-Frequency Cepstral Coefficient. Prince of Songkhla University. http://fivedots.coe.psu.ac.th/~montri/Research/Publications/icep2003_modified.pdf

Vogt, R., J.B. Baker and S. Sridharan, 2005. Modelling session variability in text independent speaker verification. Proceedings of the 9th European Conference on Speech Communication and Technology, Sept. 4-8, Lisbon, Portugal. http://eprints.qut.edu.au/15490/

Yegnanarayana, B., S.R.M. Prasanna, J.M. Zachariah and C.S. Gupta, 2005. Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system. IEEE Trans. Speech Audio Proc., 13: 575-82. DOI: 10.1109/TSA.2005.848892